ФЕДЕРАЛЬНОЕ АГЕНТСТВО НАУЧНЫХ ОРГАНИЗАЦИЙ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. № 11 Москва 2015

ОБЩИЙ РАЗДЕЛ

УДК 004.056 : 004.89

А.А. Грушо, М.И. Забежайло, А.А. Зацаринный, В.О. Писковский, С.В. Борохов

О возможностях приложений интеллектуального анализа данных в задачах обеспечения информационной безопасности облачных сред*

Представлен обзор некоторых областей применения методов и моделей интеллектуального анализа данных (ИАД) в прикладных системах обеспечения информационной безопасности. Особое внимание уделено новому активно развивающемуся направлению – облачным вычислительным средам. Обсуждаются как имеющиеся, так и перспективные возможности использования моделей и методов искусственного интеллекта при решении задач информационной безопасности.

Ключевые слова: *интеллектуальный анализ данных, облачные вычисления, информационная безопасность, математические модели и методы*

ВВЕДЕНИЕ

Информационная безопасность (ИБ) – одно из наиболее динамично развивающихся направлений в области компьютерных наук и информационных технологий (ИТ). Растущий интерес к ней демонстрируют не только исследовательское и технологическое сообщества, но и государственные структуры самого высокого уровня. Как и в ряде других стран, в Российской Федерации разработана и воплощается в жизнь национальная Доктрина информационной безопасности (см., например [1, 2] и др.).

Сегодня, наряду с такими уже традиционно ассоциируемыми с проблематикой информационной

* Работа выполнена при поддержке РФФИ (проект № 15-29-07981 офи-м).

безопасности (понимаемой как область обеспечения гарантированного управления заданными рисками и противодействия идентифицированным угрозам) задачами как, например, развитие эффективных средств криптографической защиты или систем обнаружения и предотвращения вторжений (IDS\IPS[1]), управление доступом к информационным ресурсам и другие все более широкий интерес и разработчиков, и пользователей соответствующих математических моделей, методов, алгоритмов и прикладных программно-аппаратных решений привлекает круг проблем, связываемых с расширением представлений об анализируемых угрозах и рисках – проблематикой так называемого обеспечения непрерывности бизнеса. Здесь базовые для области обеспечения ИБ понятия конфиденциальности, целостности и доступности, требования к которым в каждом конкретном случае задаются соответствующей политикой безопасности (представленной в той или иной, например, – декларативной либо процедурной – форме), могут быть интегрированы с рядом других (призванных определить нормальный режим функционирования соответствующего программно-технического комплекса) требований процедурного характера в систему проблемно-ориентированных Соглашений об Уровне Сервиса (так называемых Service Level Agreements (SLA), выполнимость которых обычно отслеживается системами мониторинга и оптимизации ИТ-ресурсов и сервисов). Наблюдаемые отклонения от подобным образом формализованной НОРМЫ поведения защищаемой ИТ-системы анализируются на предмет классификации таких отклонений на:

- не являющиеся результатом преднамеренных (целенаправленных) внешних воздействий сбои (будем называть их случайными *несущественными* техническими сбоями) и
- представляющие собой результат целенаправленных внешних вредоносных воздействий (компьютерных атак) существенные[2] отклонения от НОРМЫ.

Задачами системы обеспечения информационной безопасности (СОИБ) в таких ситуациях становятся в первую очередь:

- идентификация фактов компьютерных атак,
- причинный анализ структуры таких целенаправленных вредоносных влияний на нормальный режим работы защищаемого информационно-технологического комплекса

и, конечно же,

- выбор и оптимизация стратегий противодействия выявляемым компьютерным атакам (вместе с реализацией соответствующих мер противодействия их последствиям).

С точки зрения особенностей математических моделей и алгоритмов, требуемых для успешной идентификации и причинного анализа компьютерных атак, по-видимому, следует обратить особое внимание на необходимость:

- обучаться на прецедентах (описаниях ранее зафиксированных и изученных фактов как успешных – достигших своих целей, так и не успешных атак);
- оперировать надежным[3] образом, в том числе, как очень большими[4], так и малыми (статистически не значимыми) выборками накапливаемых эмпирических данных;
- вести исчерпывающий анализ причин успешности конкретных компьютерных атак с целью формирования эффективных средств противодействия подобным целенаправленным вредоносным воздействиям в будущем;
- оперировать как числовыми (такими, например, как частоты встречаемости тех или иных событий и т.п.), так и нечисловыми (например, наиболее значимые факторы влияния и отношения между ними) характеристиками описаний анализируемых прецедентов

и др.

Особой проблемой здесь может оказаться выбор адекватного языка описаний накапливаемых прецедентов. С одной стороны, это должен быть инструмент с достаточно полными дескриптивными возможностями, так как с его помощью предстоит внести в организуемый причинный анализ все необходимые «характерные особенности» изучаемых прецедентов компьютерных атак. Однако, с другой стороны, хотелось бы, чтобы подобный дескриптивный инструментарий имел минимальную сложность, способную обеспечить эффективный поиск соответствующих причин (а вместе с этим и поиск надежных мер противодействия их влиянию).

В совокупности перечисленные выше особенности обсуждаемой предметной области (задачи «диагностики» и противодействия опасным отклонениям от нормального режима функционирования сложных ИТ-систем) позволяют говорить о ней как об одной из сфер приложений интеллектуального анализа данных (ИАД), понимаемого как компьютерный анализ данных с помощью интеллектуальных компьютерных систем, позволяющих моделировать рассуждения и схемы принятия решений, которые характерны для экспертов в рассматриваемой области исследований и разработок.

НЕКОТОРЫЕ ОСОБЕННОСТИ ЗАДАЧ ОБЕСПЕЧЕНИЯ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ ОБЛАЧНЫХ СРЕД

Облачная вычислительная среда как программно-аппаратный комплекс имеет ряд существенных особенностей, внимание к которым критически важно при организации безопасного режима функционирования такой среды.

Прежде всего – это трехуровневая архитектура подобной среды (ее ИТ-ландшафта), где разделены:

- системно-техническая программно-аппаратная среда (физический уровень оборудования и системного программного обеспечения);

[1] Intrusion Detection\Protection Systems.

[2] Т.е. выводящие за рамки предусмотренных SLA допустимых отклонений от НОРМЫ.

[3] Обеспечивающим гарантии получения необходимого результата.

[4] Например, при выделении «шаблонов типового поведения» участников информационного обмена в крупных компьютерных сетях и др.

- управление виртуальной облачной средой – распределением нагрузки на физическое оборудование и обеспечением формирования текущей конфигурации (актуальной «топологии») облачной среды, где исполняемые приложения «не задумываются», на каком именно оборудовании каждое из них выполняется (middleware – поддерживающее виртуальную облачную среду на физическом оборудовании);
- исполняемые приложения.

Не менее важно иметь в виду, что виртуальная облачная среда, вообще говоря, одновременно оперирует тремя видами облачных функций\ресурсов:

- виртуализуемыми вычислениями,
- виртуализуемым хранением данных и
- виртуализуемыми сетевыми взаимодействиями.

Именно в связи с этими обстоятельствами приходится иметь дело с некоторыми существенными дополнительными особенностями при проектировании и построении СОИБ для облачных вычислительных сред. Так, в частности, отдельного внимания требует уточнение понятия периметра зоны ответственности СОИБ облачной среды. Заметим, что при наличии специального middleware, «отвечающего» за распределение прикладной нагрузки по устройствам «физического» уровня облачной среды, вопрос о конфигурации соответствующего периметра имеет в значительной мере иную сложность, чем, например, в обычных (необлачных) ИТ-архитектурах. Не менее существенные дополнения в условиях работы на трехуровневом ИТ-ландшафте облачной вычислительной среды приходится вносить и в процедуры формулирования и мониторинга исполнения политики безопасности, где выполнение ряда хорошо известных типов требований (например, обязательное разнесение некоторых информационных потоков по различным комплексам физического оборудования и т.п.) оказывается нетривиальной математической задачей как на стадии планирования и организации вычислений, так и на стадии контроля исполнения требований соответствующей политики безопасности[5].

В свою очередь, если придерживаться расширенного толкования информационной безопасности как комплекса мер, призванного обеспечить «непрерывность бизнеса» (т.е. непрерывность штатного режима функционирования защищаемого программно-технического комплекса), то в соответствующих Соглашениях об Уровне Сервиса (SLA) необходимо определить «границы» допустимых отклонений от нормального (штатного) режима функционирования облачной среды на всех трех уровнях ее ИТ-ландшафта. При этом режим НОРМЫ может быть определен, в том числе, и с помощью фиксации[6] заданных:

- уровней доступности заданных информационных ресурсов облачной среды, в том числе – для участников информационного обмена в облачной среде;
- уровней предельно допустимых задержек (времени ожидания) доступа к заданным типам информационных ресурсов в облачной среде, в том числе – для участников информационного обмена в облачной среде.

ВОЗМОЖНОСТИ ОРГАНИЗАЦИИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В СОИБ ОБЛАЧНЫХ ВЫЧИСЛИТЕЛЬНЫХ СРЕД: ОБЩИЕ ХАРАКТЕРИСТИКИ ПОДХОДА

Требующая своего решения в рамках мониторинга событий безопасности (см., например, [5] и др.) проблема классифицировании (экспертизы и оценки) каждой ситуации выхода за рамки SLA (и, соответственно, определения НОРМЫ), как уже было показано, может быть сведена к «опознанию» и дифференциации ситуаций двух типов:

- случайного технического сбоя (флюктуации),
- результата целенаправленного вредоносного воздействия.

Первым шагом в этом процессе является выбор языка описания данных, используемого для фиксации накапливаемых сведений о наблюдаемых событиях. При этом вопрос об адекватных выразительных (описательных) возможностях такого языка, по-видимому, имеет принципиально важное значение: использование компьютерных средств анализа данных, полученных при фиксации событий, предполагает возможности представить его средствами (пусть и в неявном виде) все существенные факторы, взаимодействие которых и привело к возникновению каждого фиксируемого события. (Действительно, только в этом случае можно рассчитывать на приемлемость результатов подобного компьютерного анализа). Однако в каждом конкретном случае представляется естественным выбирать язык адекватной (например, минимальной достаточной) сложности, так как требования к скорости анализа данных и реальные объемы этих данных при решении прикладных задач могут быть весьма чувствительны к объемам возникающих здесь вычислений.

При организации обсуждаемой классификации (экспертизы и оценки) результатов мониторинга событий безопасности представляется естественным потребовать, чтобы собственно процесс вычислений (компьютерного анализа данных) и процедуры оценки его результатов (оценки достаточности оснований для принятия результатов выполненной экспертизы) были разделены и выполнялись как самостоятельные процедуры соответствующего регламента.

Место для ИАД-экспертизы в таком процессе определяется по крайней мере двумя обстоятельствами. Во-первых: соответствующая компьютерная система анализа данных может использоваться как «усилитель интеллектуальных возможностей»[7], выполняя те же, что и эксперт операции (что позволяет ему понимать и однозначным образом интерпретировать порождаемые ею результаты), однако делать это существенно быстрее и в существенно больших объемах (чем использующий ее результаты эксперт).

[5] См., например, работы [3, 4] и др., представляющие вариант математической техники и программных решений для задач этого типа.

[6] В виде соответствующих SLA (Service Level Agreement).

[7] Термин, введенный У.Р.Эшби – см. [6].

Во-вторых, накапливаемые в процессе мониторинга событий безопасности данные могут быть использованы при экспертизе и оценке вновь фиксируемых событий, если единообразным образом описанные события, уже зафиксированные СОИБ в процессе мониторинга, использовать как соответствующую обучающую выборку. (При этом факты результативных, достигших своих целей, и нерезультативных компьютерных атак могут быть использованы как множества примеров и контрпримеров для организации машинного обучения на прецедентах).

Так, общая схема ИАД-экспертизы может быть описана следующими тремя шагами:

1. Анализ накапливаемых данных.
2. Формирование эмпирических зависимостей («*решающих правил*»).
3. Экспертиза новых инцидентов (проводимая с помощью порожденных из обучающей выборки эмпирических зависимостей).

При этом, еще в середине 80-х годов на стадии формирования ориентированных на СОИБ-приложения экспертных систем продукционного характера (см. подробнее в следующем Разделе) было осознано, что:

■ шаги1-2 могут оказаться весьма ресурсоемкими (например, здесь в дополнение к собственно процедурам порождения эмпирических зависимостей – «решающих правил» – необходимо уметь поддерживать полноту и непротиворечивость их актуального на каждый момент времени набора, причем, не только пополняемого, но и, возможно, изменяемого в процессе работы системы), а

■ шаг 3 с ростом объемов анализируемых данных требует все более эффективной (прежде всего - в части быстродействия) организации процедур экспертизы.

Таким образом, стало понятно, что шаг 3 придется исполнять в условиях все более жестких ограничений режима реального времени, а вот шаги 1-2 есть возможность реализовать в «фоновом» по отношению к собственно экспертизе режиме (что сегодня и можно наблюдать в архитектуре и организации функционирования промышленных СОИБ-систем). Более того, при таком разделении при исполнении шагов 1-2, вообще говоря, уместно, если это потребуется, и подключение экспертов, что в современных промышленных СОИБ-решениях практически исключено в случае экспертизы новых инцидентов из-за актуальных ограничений режима реального времени.

При организации процесса порождения таких зависимостей придется разделить два самостоятельных (организованных по разным принципам и требующих использования соответствующих инструментов анализа – математических моделей, методов и алгоритмов) типа «сред» мониторинга. Первую из них характеризуют большие выборки часто повторяющихся событий, позволяющие выявлять «устойчивые шаблоны поведения». Для второй, наоборот, существенны возможности оперировать малыми (статистически не значимыми) выборками прецедентов, где роль каждого отдельного события может быть критически важна.

Далее в обсуждении мы будем использовать следующие, на наш взгляд, удобные понятия и обозначения.

События (инциденты), которые могут быть квалифицированы как

♦ случайные флюктуации (непреднамеренные технические сбои) или же, наоборот, как

♦ результат целенаправленного вредоносного воздействия.

Прецеденты:

♦ события, классифицированные как «успешные» (достигшие целей) целенаправленные вредоносные воздействия;

♦ события, по своим описаниям «подобные» целенаправленным вредоносным воздействиям, однако классифицированные как «не имевшие успеха» (не достигшие целей) целенаправленные вредоносные воздействия.

В процессе классификации событий безопасности мы будем различать следующие три случая (результаты предпринимаемой экспертизы):

(i) «*да*» (прецедент, классифицированный как «успешное» целенаправленное вредоносное воздействие);

(ii) «*нет*» (прецедент, классифицированный как «не имевшее успеха» целенаправленное вредоносное воздействие);

(iii) «*не знаем*» (нет или же не хватает данных для аргументированной классификации прецедента как «успешного» или же «не имевшего успеха» целенаправленного вредоносного воздействия).

Основной целью нашего дальнейшего обсуждения будет обзор наиболее существенных свойств (функциональных возможностей) математических моделей, методов и алгоритмов:

- анализа первичных событий (формирования обучающей выборки),
- формирования «правил» (процедур, методов, правил принятия решения и т.п.) классификации событий безопасности,

и, наконец,

- исполнения требований политики безопасности, обусловленных выполненной на базе порожденных «правил» классификацией.

При этом в анализе первичных событий представляется важным (учитывая обсужденные выше требования к языку описания данных) особое внимание уделить:

■ доступным возможностям фиксации «всех имеющих отношение к делу» деталей,

■ выбору (формированию) системы языков описания прецедентов, допускающих детализацию и усложнение,

а также

■ анализу[8] «сходства» и «различий» в имеющихся описаниях прецедентов.

В процессе исполнения классификации вторичных событий на базе «правил» следует обратить внимание на корректность процедур «экстраполяции», выраженных «правилами» эмпирических зави-

[8] См. выше определение негативных («не имевших успеха») прецедентов

симостей, на описания вторичных событий, чтобы обеспечить аргументированную «достаточность оснований» для принятия результатов реализуемой «экстраполяции».

НЕСКОЛЬКО ЗАМЕЧАНИЙ ПО ИСТОРИИ ВОПРОСА И ВЫБОРУ ИСТОЧНИКОВ ИНФОРМАЦИИ ДЛЯ ПРЕДПРИНЯТОГО ОБЗОРА

Историю развития приложений комплекса моделей, методов и алгоритмов, который впоследствии (во второй половине 1990-х гг.) будет назван интеллектуальным анализом данных, в задачах обеспечения информационной безопасности, по-видимому, следует вести с начала 1980-х, гг., когда появился цикл работ Дороти Деннинг (Dorothy E. Denning) и ее коллег о моделях и программных системах обнаружения вторжений в охраняемые компьютерные сети (см., например, [7] и др.). Появление таких процедурных подходов и инструментальных средств противодействия несанкционированным проникновениям в компьютерные сети стало естественным ответом на явно обозначившиеся технологические вызовы, обусловленные развитием сетевых технологий, а вместе с ними – и возможностей осуществления внешних вредоносных воздействий на те или иные компьютерные системы. Развиваемая Д. Деннинг и ее коллегами технология базировалась на двух группах функций: «диагностике» известных типов вторжений (реализуемой средствами продукционной экспертной системы) и «идентификации» (установлении фактов) вторжений, использующей статистические методы анализа данных и формируемые в процессе мониторинга объекта защиты профили сетевого поведения пользователей и компонентов охраняемой сети.

Следующим (вполне естественным) шагом развития стало дополнение процедурного «инструментария» таких систем достаточно популярными на тот момент (вторая половина 80-х – начало 90-х гг. XX в.) проблемно-ориентированными статистическими моделями и технологиями нейронных сетей. Например, в работах Терезы Лант (Teresa F. Lunt) и ее коллег из SRI International (см., например, [8, 9] и др.) представлена технология (и построенная на ее основе экспертная система следующего поколения), интегрирующая решение класса IDES [7] и нейронную сеть, обеспечивая при этом режим реального времени для выполняемого компьютерного анализа данных.

Следующим (не менее ожидаемым) шагом развития рассматриваемых технологий стала интеграция в них средств систематического анализа аномалий в поведении объекта защиты и порождения на основе результатов такого анализа новых правил обнаружения аномалий. Так, в системе обнаружения аномалий W&S [10], разработанной Ваккэйро (H.S. Vaccaro) и Лиепиньшем (G.E. Liepins) в Лос-Аламосской Национальной лаборатории (США), результатом статистического анализа данных становились правила выявления аномалий, которые далее использовались для поиска новых аномалий в данных.

Следующим шагом вперед стала интеграция в уже используемый в системах обеспечения информационной безопасности набор моделей и методов интеллектуального анализа данных также и математических моделей и методов индуктивного обучения (машинного обучения на прецедентах) – см., например, [11] и другие.

Проводимые в последующие десятилетия исследования и разработки в обсуждаемой нами области можно в самом общем виде разделить на два направления:

1) собственно разработка математических моделей, методов и алгоритмов, а также экспериментальных компьютерных систем интеллектуального анализа данных для их последующего использования в прикладных решениях

и на

2) создание и развитие промышленных СОИБ-решений.

Коммерческие результаты второго из этих направлений весьма обширно представлены в открытой печати различными материалами обзорного или же обще-информационного характера (см., например, обзоры [12, 13] или же аналитические сайты типа [14] др.), в которых помимо рекламно-маркетинговых данных о функциональных возможностях верхнего уровня для соответствующих решений\продуктов присутствуют лишь важнейшие эксплуатационные характеристики (скорость обработки данных и т.п.).

В то же время, информация первого типа практически отсутствует в открытых источниках (исключая материалы специальных научных конференций типа USENIX Security Symposium – см., например, [15], или IEEE Symposium on Security and Privacy – см., например, [16][9] и др.). В крупных корпорациях (ведущих самостоятельные исследования и разработки по обсуждаемой тематике – таких как IBM, McAffe, Cisco и т.п.) подробные материалы этого типа, как правило, остаются внутренними техническими отчетами (Technical Reports, White Papers, …), которые недоступны за пределами исследовательских служб этих компаний.

Однако, если воспользоваться хорошо известными в аналитическом сообществе приемами обобщения косвенной информации, то и в рассматриваемой ситуации (опираясь на общедоступные данные об исследованиях и разработках IBM, SAS Institute, Intel Security, HP, Cisco и др.) можно получить ряд полезных заключений о функциональных возможностях, а также некоторых эксплуатационных характеристиках подходов, математических моделей и построенных на их основе прикладных систем интеллектуального анализа данных, используемых в обсуждаемой нами области.

Поясним эти соображения несколькими примерами. Проблематика AML&FP - так называемых Anti-Money Laundering (противодействия отмыванию де-

[9] Материалы которого можно посмотреть, например, в *IEEEXplore* (через ссылку http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6954656).

нег – легализации незаконных доходов) и Fraud Protection\Detection\Prevention (противодействия мошенничеству в финансовой сфере) – см. например, [17, 18] и др. – с точки зрения используемых здесь математических моделей и методов представляет собою весьма близкую к обсуждаемой нами проблематике обеспечения информационной безопасности предметную область. Здесь так же, как и в области ИБ, актуальны (характерные для интеллектуального анализа данных) и задачи реконструкции по постоянно накапливаемым эмпирическим данным устойчивых «шаблонов» поведения объектов мониторинга, и задачи выявления аномалий в «типовом» поведении таких объектов, и, наконец, задачи идентификации вновь наблюдаемых событий как аномальных отклонений от заранее описанных (в том числе – на основании уже сформированных «шаблонов» «типового» поведения) представлений о НОРМЕ функционирования объекта мониторинга и защиты. Одним из бесспорных лидеров в области исследований и разработки промышленных программных систем для этой области финансовой безопасности является американская компания SAS Institute [19]. Ее проблемно-ориентированными решениями в области AML&FP пользуются крупнейшие банки мира. В плане исследований и разработок она известна своими партнерскими связями с ведущими университетами по всему миру, а об активности ее собственных исследовательских подразделений можно судить, например, по объемам патентуемых этой компанией ИТ-решений [20]. Таким образом, можно утверждать, что математические модели, методы и алгоритмы интеллектуального анализа данных, используемые компанией SAS Institute в ее промышленных AML&FP-решениях, отражают лучшие практики и достижения соответствующего сообщества исследователей и разработчиков. (В следующем Разделе мы поговорим о них более подробно).

Покупка корпорацией Intel одного из многолетних лидеров в секторе разработки программных продуктов и решений в области ИБ – компании McAffe (сегодня она выступает под брендом Intel Security [21]) – обозначила новое направление перспективного развития промышленных решений в рассматриваемой нами области. Оно ориентировано на глубокую интеграцию программных и аппаратных компонентов, результатом которой является встраивание функций безопасности в аппаратную архитектуру вычислительных систем всех уровней – от микросхем до облачных решений. Такая интеграция ориентирована в том числе и на взаимодействие с внешними партнерами – разработчиками конечных решений, которым обеспечивается доступ как к перспективным аппаратным компонентам (chip-set'ам), так и к проблемно-ориентированным средствам поддержки разработки (tool-kit'ам). Некоторые возможности именно такого типа, открывающиеся в обсуждаемой нами проблематике безопасности облачных вычислений, мы обсудим далее в одном из следующих разделов настоящей статьи, опираясь на информацию об активном интересе Intel в том числе и к быстро развивающимся облачным решениям [22].

В области интеллектуального анализа так называемых Big Data комплекс технологий IBM Watson [23, 24 и др.] сегодня – один из безусловных лидеров как исследовательского, так и технологического сообществ. Развиваемые в этом комплексе возможности каузального анализа больших объемов неструктурированной информации (текстов на естественном языке, порождаемых сложной диагностической аппаратурой[10] графических образов и др.), а также возможности машинного обучения на контекстах (специально подобранных массивах информации) позволяют использовать этот инструментарий анализа компьютерного данных для построения каузальных сетей очень крупных размеров и восстановления (скрытых в фактических данных – !) эмпирических зависимостей между большими множествами взаимодействующих факторов. Технологии такого типа имеют очевидные перспективы в анализе больших объемов данных технологического мониторинга физических и виртуальных (в том числе – облачных) вычислительных сред, ориентированном на устранение сбоев в нормальном режиме функционирования ИТ-инфраструктуры, идентификацию и противодействие компьютерным атакам (см., например, специализированные решения в области мониторинга ИТ-инфраструктуры и поддержки непрерывности бизнеса, предлагаемые компаниями BMC Software [25, 26], HP [27], Splunk [28, 29] и др.). Однако, не меньший интерес могут представлять возможности применения таких технологий ИАД и в области компьютерного анализа причинно-следственных связей, скрытых в больших объемах текстовой информации (см., например, проект Big Mechanism агентства DARPA [30] и др.).

ФОРМИРОВАНИЕ «*ПРАВИЛ*» *КЛАССИФИКАЦИИ ПРЕЦЕДЕНТОВ*: (ОБЗОР ПРОЦЕДУРНЫХ ТЕХНИК)

Среди используемых в рассматриваемой нами области формальных процедур формирования и классификации событий безопасности наиболее полного внимания, по-видимому, заслуживают следующие два класса математических моделей и методов:

(a) поиск устойчивых «регулярностей» в накапливаемых данных мониторинга и

(b) проверка выполнимости заданных систем ограничений.

В свою очередь, первый из этих двух классов представляется уместным разделить на два (существенным образом различающихся по используемым в них моделям, методам и алгоритмам) подкласса, ориентированных на обработку:

(a1) больших (содержащих устойчивые многократно повторяющиеся статистически значимые «регулярности» в поведении наблюдаемых объектов) и

(a2) малых (статистически не значимых) коллекций прецедентов.

Для случая (a1) наиболее активно используемыми математическими техниками ИАД и поддержки при-

[10] Например, медицинскими томографами и т.п.

нятия решений являются регрессионный анализ и кластеризация прецедентов, а также различные варианты технологии байесовского индуктивного вывода. (По-видимому, наиболее полную версию реализации этих математических инструментов анализа данных можно найти в предлагаемых компанией SAS Institute программных комплексах SAS Enterprise Miner и SAS/STAT [31, 32]).Достаточно популярны технологии нейронных сетей (Kohonen self-organizing maps – [31 и др.]). Широкое распространение получили инструменты анализа web-ссылок (Link analysis in Web log data – см., например, [31] и др.), которые ориентированы на извлечение «регулярностей» (устойчиво повторяющихся или же взаимосвязанных сведений о том, кто, когда и с какими запросами посещал соответствующие web-серверы) в хранимых log файлах (или же соответствующих базах данных). Как правило, при анализе событий безопасности подобные инструменты помогают определить: что собственно происходит, и чьими действиями эти события обусловлены.

Отдельного внимания заслуживают модели и инструменты анализа ассоциативных связей. Здесь, наряду с хорошо изученной проблематикой восстановления из эмпирических данных различных видов корреляционных зависимостей (парных корреляций, корреляционных деревьев и т.п. – см., например [31, 33] и др.), все более активно развиваются решения на базе так называемых правил ассоциации (Association Rules – [34, 35] и др.).

Для работы с текстовыми данными используются различные варианты представления документов в виде специального вида графа (извлекаемых из текста автоматизированными средствами) понятий и отношений (см., например, [36] и др.). Затем, на коллекциях таких графов организуются процедуры поиска (data mining'а), целью которых является идентификация «регулярностей» (устойчивых взаимосвязей) в таких подборках данных, позволяющая классифицировать документы по темам, формировать укрупненные представления о содержании тех или иных коллекций документов (их рефератов), соотносить содержание коллекций документов с теми или иными запросами и др. На этих принципах, в частности, строится и реализованная в рамках IBM Watson технология обучения на контекстах [23] и др.

В случае (a2) ситуация складывается не столь благополучно. Многие успешно применимые в случае (a1) модели и технологии при использовании соответствующих алгоритмических процедур анализа данных на малых выборках прецедентов перестают быть математически корректными. Таким образом, вопрос о доверии к порождаемым результатам (о наличии достаточных оснований для их принятия) становится критически важным фактором в принятии решений о применимости тех или иных подходов в конкретных приложениях. Попытки обойти подобные трудности, например, путем использования тех или иных дискретных процедур индуктивного обобщения описаний накапливаемых прецедентов достаточно быстро упираются в проблемы экспоненциально быстро растущих объемов вычислений (заметим, что еще на рубеже 1960-1970-х гг.G. Plotkin [37, 38] показал, что как и в ситуации с методом резолюций, задача порождения минимальных индуктивных обобщений для множеств формул логики предикатов первого порядка имеет экспоненциальные характеристики сложности вычислений) и, как правило, имеют далекое от реальных приложений лишь академическое значение.

Ситуации типа (b) характерны, в частности, для обширного класса задач контроля корректности конфигурирования компьютерных сетей. Здесь в процессе управления сетью необходимо не только обеспечить выполнение требований, задаваемых политикой безопасности, но и сделать это так, чтобы не вносить существенных задержек в процесс собственно информационного обмена. Обычно решения такого типа имеют интегральный характер – часть задач решается аналитически в рамках (технически, как правило, весьма изощренных) математических моделей, в то время как другая, оставшаяся часть, – применением проблемно-ориентированных вычислительных моделей и алгоритмов. Интересные примеры такого типа решений для сетевых технологий нового поколения (в частности – для программно-конфигурируемых сетей) дают, в том числе и работы Open Network Laboratory, Stanford University, USA [39, 40 и др.]. Другой, не менее интересный пример эффективных решений этого типа, – разработанные в московском Центре прикладных исследований компьютерных сетей (ЦПИКС) в сотрудничестве со специалистами факультета вычислительной математики и кибернетики МГУ им. М.В. Ломоносова математические модели, алгоритмы и программные системы формализованной верификации программно-конфигурируемых сетей [3, 4]. Здесь задача контроля корректности исполнения набора базовых требований политики безопасности (в том числе – достижимости вида *точка-точка*, отсутствия циклов, разделения трафика между заданными наборами вершин на непересекающиеся потоки и др.) сперва формализована средствами специально разработанного проблемно-ориентированного расширения языка логики предикатов первого порядка (обеспечивающего выразимость транзитивного замыкания бинарного отношения и формализацию отношения так называемой одношаговой достижимости для вершин графа программно-конфигурируемой сети), далее средствами бинарных решающих диаграмм формализован метод верификации сетей этого типа, который, в свою очередь, воплощен в специальный программный инструментарий (прикладную программную систему верификации сети).

Стандартом de facto в области разработки систем обнаружения и предотвращения вторжений (IDS\IPS) является созданная Мартином Рёшем (Martin Roesch) система анализа сетевого трафика Snort [41]. Система имеет открытый исходный код, а используемые ею для поиска по содержимому передаваемых в сети пакетов данные находятся в открытом доступе [42]. Известен ряд попыток воспроизвести показатели производительности системы Snort в рамках других аналогичных решений, однако, далеко не все из

них могут предложить сопоставимые характеристики быстродействия. Можно предположить, что именно это обстоятельство выступает дополнительным стимулом для разработки новых методов и математических моделей обнаружения вторжений. Интересным примером такого подхода являются работы группы специалистов мехмата МГУ им. М.В. Ломоносова по созданию формальных автоматных моделей для распознавания принадлежности слов регулярному языку, где переформулирование регулярных выражений анализируемого языка исполняется компьютерной системой в автоматическом режиме [43, 44]. Эта технология также позволяет улучшить показатели производительности систем информационной безопасности, ориентированных на противодействие сетевым вторжениям.

Дополнительные возможности при организации ИАД в задачах класса (а) может дать применение математической техники так называемых корректных алгебр над множеством некорректных (эвристических) алгоритмов, предложенной Ю.И. Журавлевым [45] и успешно развиваемой его школой [46, 47 и др.]. Также полезным может быть использование формальных моделей эмпирической индукции на структурных (нечисловых) описаниях прецедентов и методов оптимизации возникающего здесь комбинаторного перебора, разрабатываемых школой В.К. Финна [48-50 и др.].

Завершая краткий обзор подходов, методов и моделей ИАД, представляется целесообразным обратить внимание на ряд существенных ограничений, характерных для рассмотренных технологий. (Хотелось бы надеяться, что именно эти обстоятельства смогут стать одним из стимулов для разработки новых, более совершенных решений). Итак, наиболее существенные «узкие места» – это:

- независимость наблюдаемых в описаниях прецедентов факторов (переменных) для выполнения в рамках ИАД корректного статистического анализа;
- надежность работы с малыми (статистически не значимыми) выборками;
- полнота выполняемого в рамках ИАД каузального анализа (полнота множества выделяемых факторов) целенаправленных вредоносных воздействий;
- использование лишь объяснительных переменных при регрессионной интерполяции обучающей выборки и проблема полноты множества факторов, характеризующих вредоносные воздействия;
- необходимость работать в процессе ИАД в том числе и со структурными (нечисловыми) объектами, наделенными также и числовыми характеристиками;
- необходимость проводить оценку достаточности оснований для принятия результатов ИАД-«экстраполяции» на новые объекты тех эмпирических зависимостей, которые получены в процессе «интерполяции» данных обучающей выборки;
- необходимость преодолевать «проклятие» сложности вычислений, которая обусловлена наличием в рассматриваемой предметной области трудноразрешимых переборных задач, оказывающих критически значимое влияние на требования к производительности соответствующих прикладных программных систем (проблема управления и оптимизации комбинаторного перебора).

ИНТЕГРАЦИЯ ПРОГРАММНЫХ И АППАРАТНЫХ РЕШЕНИЙ КАК ТРЕНД РАЗВИТИЯ ИАД-РЕШЕНИЙ В СИСТЕМАХ ОБЕСПЕЧЕНИИ БЕЗОПАСНОСТИ ОБЛАЧНЫХ СРЕД

Анализ текущей ситуации в рассматриваемой нами области показывает, что все ведущие «игроки» этого сегмента ИТ-рынка параллельно с развитием математических моделей и методов проблемно-ориентированного интеллектуального анализа данных существенное внимание уделяют развитию специальных программно-аппаратных средств поддержки производительности подобных инструментальных решений.

Так, компания SAS Institute наряду с развитием функциональности своих продуктов SAS Enterprise Miner, SAS/STAT, SAS/OR, SAS Constraint Programming Solver и других развивает «сопутствующие» технологии класса High-Performance Procedures [51], задача которых – оптимизировать среду исполнения вычислений и обеспечить максимально возможную производительность базовых решений Компании.

Параллельно с технологиями IBM Watson активно развиваются облачные системно-технические решения компании IBM [52 и др.]. Так, одним из типовых решений для организации облачных ландшафтов в рамках американской национальной программы GENI[11] [53] является программно-аппаратный комплекс GENI-Rack компании IBM.

В архитектуре одного из лидеров соответствующего сегмента глобального рынка промышленных ИБ-решений – решения TippingPoint Security Manager компании HP [13, 54 и др.] – ключевым элементом являются специализированные микросхемы (ASIC). По данным [13], используемый в HP TippingPoint механизм подавления угроз Threat Suppression Engine (TSE) реализован на базе заказных проблемно-ориентированных ASIC-решений. Благодаря сочетанию специализированных ASIC, объединительной панели с пропускной способностью 20 Гбит/с и высокопроизводительных сетевых процессоров механизм TSE обеспечивает полный анализ потока пакетов на уровнях L2—L7, при этом задержка прохождения потока через IPS-систему составляет менее 150 мкс, вне зависимости от количества примененных фильтров.

Решения компании Cisco в области обеспечения безопасности облачных сетевых взаимодействий также (в традиционном для этой Компании стиле) интегрируют профильные проблемно-ориентированные программные и аппаратные компоненты [55].

Наконец, решения для обеспечения безопасности облачных ландшафтов, причем - как разрабатываемые, так и уже предлагаемые рынку под брендом Intel Security объединением компаний McAffe и Intel

[11] GENI –Global Environment for Network Innovations.

[21] и, в частности – прикладные системы класса McAfee Network Security Platform (см., [13] и др.), будут интегрированы с аппаратными архитектурами, формируемыми в том числе и на базе разрабатываемых Intel специализированных chip-set'ов. При этом уже доступны проблемно-ориентированные инструментальные средства поддержки разработки класса Intel DPDK [56], SDK [57] и др. Также уже сегодня в архитектуре серверов класса Seacliff [58] несложно увидеть специальные возможности для эффективного использования прикладных программных ИАД-решений McAffe.

ЗАКЛЮЧЕНИЕ

Предпринятый нами анализ позволяет указать перспективную «нишу» для развития оригинального и актуального для приложений подхода, который мог бы объединить в рамках единого комплекса математических моделей, методов и алгоритмов:

- интеграцию статистических и детерминистских методов[12] ИАД в задачах обеспечения информационной безопасности в облачных средах;
- умение оперировать выборками произвольного размера (как в ситуациях Big Data, так и при работе с малыми – статистически не значимыми – выборками прецедентов);
- умение управлять сложностью вычислений (выбор адекватных языков представления данных и использование процедур оптимизации перебора) при поиске зависимостей в данных обучающей выборки;
- умение оперировать нечисловыми объектами (описаниями прецедентов), содержащими в том числе и числовые значения существенных параметров;
- возможности вести каузальный анализ изучаемых целенаправленных вредоносных воздействий с целью подбора
- оптимального (по задействованным ресурсам) исчерпывающего набора мер по противодействию таким воздействиям.

Реализация перечисленных функциональных характеристик в рамках проблемно-ориентированных инструментальных программных систем дает возможность надеяться на достижение нового, более высокого уровня защищенности перспективных прикладных решений.

СПИСОК ЛИТЕРАТУРЫ

1. Доктрина информационной безопасности Российской Федерации (утверждена Президентом РФ 9 сентября 2000 г., N Пр-1895). – URL: http://www.scrf.gov.ru/documents/5.html
2. Воронина Ю. Данные засекречены // Российская газета. – 2015. – № 984 (10 февраля). – URL: http://www.rg.ru/printable/2015/02/10/ib.html
3. Захаров В.А., Смелянский Р.Л., Чемерицкий Е.В. Формальная модель и задачи верификации программно-конфигурируемых сетей // Моделирование и анализ информационных систем. – 2013. – Т. 20, № 6. – С. 33–48.
4. Захаров В.А., Чемерицкий Е.В. О некоторых задачах реконфигурирования программно-конфигурируемых сетей // Моделирование и анализ информационных систем. – 2014. – Т. 21, № 6. – С. 57–70.
5. ГОСТ Р ИСО/МЭК 15408-1-2008. Информационная технология. Методы и средства обеспечения безопасности. Критерии оценки безопасности информационных технологий. –URL:http://tehnorma.ru/gosttext/gost/gost_4493.htm
6. Эшби У.Р. Введение в кибернетику. – М.: Иностранная литература, 1959. – 432 с. – URL: http://pcp.vub.ac.be/ASHBBOOK.html
7. Denning D. E. An Intrusion Detection Model // Proceedings of the Seventh IEEE Symposium on Security and Privacy. – 1986. – May. – P. 119–131.
8. Lunt T.F. Detecting Intruders in Computer Systems // Proceedings of the 1993 conference on auditing and computer technology. – SRI International. – URL: http://www.researchgate.net/profile/ Teresa_Lunt/publication/2304057_Detecting_Intruders_in_Computer_Systems/links/552e86500cf2acd38cba5c94.pdf
9. Anderson D., Lunt T.F., Javitz H., Tamaru A., Valdes A. Detecting Unusual Program Behavior Using the Statistical Component of the Next-generation Intrusion Detection Expert System (NIDES) // SRI International, Computer Science Laboratory. – Technical Report SRI-CSL-95-06. – May, 1995. – 77 p. – URL: http://www.csl.sri.com/papers/5sri/5sri.pdf
10. Vaccaro H.S., Liepins G.E. Detection of Anomalous Computer Session Activity // The 1989 IEEE Symposium on Security and Privacy (Oakland, CA, USA, May, 1989). – P.280 – 289.
11. Teng H.S., Chen K., Lu S.C-Y. Adaptive Real-time Anomaly Detection Using Inductively Generated Sequential Patterns // IEEE Symposium on Security and Privacy. – May 7-9, 1990. – P. 278-284.
12. Каталог средств защиты информации. – URL: http://zlonov.ru/catalog/
13. Дрозд А. Обзор корпоративных IPS-решений на российском рынке. – URL: http://www.anti-malware.ru/IPS_russian_market_review_2013
14. Intrusion Prevention Systems. – Moxize: IT Solution Discovery & Research. – URL: https://www.moxize.com/Category/Detail/20/intrusion-prevention-systems
15. USENIX Security '14 (23-th USENIX Security Symposium. – August 20-24, 2014. – San Diego, CA). – URL: https://www.usenix.org/ conference/ usenixsecurity14
16. 35-th IEEE Symposium on Security and Privacy. – May 18-21, 2014. – San Jose, CA. – URL: http://www.ieee-security.org/TC/SP2014/index.html

[12] В первую очередь – основанных на машинном обучении на прецедентах, использующем формальные модели эмпирической индукции.

17. Financial Action Task Force[13]. – URL: http://www.fatf-gafi.org/
18. VISA: Fraud Prevention Tools & Real Time Fraud Detection. – URL: http://usa.visa.com/personal/ security/security-program/index.jsp
19. FORTUNE: 100 Best Companies to Work for. SAS Institute. – URL: http://fortune.com/best-companies/ sas-institute-4/
20. SAS Institute (Inc.). Patent applications. – URL: http://www.faqs.org/patents/assignee/sas-institute-inc/
21. Intel Security.– URL: http://www.intelsecurity.com/
22. Clark D. Intel Lead $100 Million Investment into Mirantis // The Wall Street Journal. (August 24, 2015). – URL: http://www.wsj.com/articles/intel-to-lead-100-million-investment-into-mirantis-1440388913
23. Zhu W-D., Foyle B., Gagné D., Gupta V., Magdalen J., Mund I A.S., Nasukawa T., Paulis M., Singer J., Triska M. IBM Watson Content Analytics: Discovering Actionable Insight from Your Content (3-d Edition, July 2014). – IBM Redbooks: IBM Corp. – xxii + 570 p. – URL: http://www.redbooks.ibm.com/abstracts/sg247877.html?Open
24. Lally A., Bagchi S., Barborak M.A., Buchanan D.W., Chu – Carroll J., Ferrucci D.A., Glass M.R., Kalyanpur A., Mueller E.T., Murdock J.W., Patwardhan S., Prager J.M., Welty C.A. WatsonPaths: Scenario-based Question Answering and Inference over Unstructured Information (IBM Research Report RC25489). – IBM Thomas J. Watson Research Center, Yorktown Heights, NY. – 2014. – 20 p. – URL: www.patwardhans.net/ papers/LallyEtAl14.pdf
25. BMC Remedy. – URL: http://www.bmc.com/it-solutions/remedy-itsm.html
26. BMC Software, Eucalyptus, HP, IBM, Intel, Red Hat and SUSE Create Open Virtualization Alliance. – URL: https://openvirtualizationalliance.org/news-events/news/2011/05/bmc-software-eucalyptus-hp-ibm-intel-red-hat-and-suse-create-open
27. HP Open View. Enterprise Security. – URL: http://www8.hp.com/us/en/software-solutions/enterprise-security.html
28. Carasso D. Splunk. – CITO Research. – 2013. – 168 p.
29. Carasso D. Data Mining with Splunk. – URL: http://www.slideshare.net/davidcarasso/datamining5
30. Cohen P. Big Mechanism (DARPA Big Mechanism Program). – URL: http://www.darpa.mil/ program/big-mechanism
31. Data Mining Using SAS Enterprise Miner. A Case Study Approach. – URL: http://support.sas.com/ documentation/cdl/en/emcs/66392/PDF/ default/emcs.pdf
32. SAS/STAT 14.1. User's Guide. High-Performance Procedures. – URL: http://support.sas.com/documentation/ cdl/en/stathpug/68163/PDF/default/stathpug.pdf
33. Pearl J. Causality: Models, Reasoning, and Inference. – Cambridge: Cambridge University Press, 2000. – 451 p.
34. Agrawal R., Imielinski T., Swami A. Mining association rules between sets of items in large databases // Proc. 1993 ACM SIGMOD international conference on Management of data (SIGMOD '93). – N.-Y.: ACM, 1993. – P. 207 – 216.
35. Agrawal R., Srikant R. Fast Algorithms for Mining Association Rules // Proc. 20th int. conf. very large data bases (VLDB). – Morgan Kaufmann. – 1994. – P.487-499.
36. Tkach D. Text Mining Technology: Turning Information Into Knowledge. – IBM White Paper. – 1998. – 20 p. – URL: http://www.math.unipd.it/ ~dulli/corso04/whiteweb.pdf
37. Plotkin G.D. A Note on Inductive Generalization // Machine Intelligence. – 1970. – № 5. – P. 153-164.
38. Plotkin G.D. A Further Note on Inductive Generalization // Machine Intelligence. – 1971. – № 6. – P. 101-124.
39. Kazemian P., Chang M., Zeng H., Varghese G., McKeown N., Whyte S. Real Time Network Policy Checking using Header Space Analysis // Proc. 10th USENIX Symposium on Networked Systems Design and Implementation (April 2-5, 2013, Chicago, IL). – 2013. – P. 99-111. – URL: https://www.usenix.org/system/files /conference/ nsdi13/nsdi13-final8.pdf
40. Kazemian P., Varghese G., McKeown N. Header space analysis: static checking for networks // Proc. 9th USENIX Symposium on Networked Systems Design and Implementation (April 25-27, 2012, San Jose, CA). – .2012. – P. 49-54. – URL: http://yuba.stanford.edu/~peyman/docs/headerspace-nsdi12.pdf
41. Snort. – URL:https://www.snort.org
42. База сигнатур системы Snort. – URL: http://www.snort.org/snortrules/
43. Галатенко А.В. Автоматные модели защищенных компьютерных систем // Интеллектуальные системы. – 2007. – Т. 11, № 1-4. – С. 403-418.
44. Александров Д. Е. Эффективные методы реализации проверки содержания сетевых пакетов регулярными выражениями // Интеллектуальные системы. – 2014. – Т. 18, № 1. – С. 37–60.
45. Журавлев Ю.И. Корректные алгебры над множествами некорректных (эвристических) алгоритмов // Кибернетика. – Часть I. – 1977. – № 4. – С. 5-17; Часть II. – 1977. – № 6. – С. 21-27; Часть III. – 1978. – № 2. – С. 35-43.
46. Журавлев Ю.И., Рязанов В. В., Сенько О. В. «Распознавание». Математические методы. Программная система. Практические применения. – М.: Фазис, 2006. – 176 с.
47. Рудаков К.В. О некоторых универсальных ограничениях для алгоритмов классификации // Журнал вычислительной математики и мате-

[13] Группа разработки финансовых мер борьбы с отмыванием денег.

матической физики. – 1986. – Т.26, №11. – С. 1719–1730.

48. Автоматическое порождение гипотез в интеллектуальных системах / ред. В.К.Финн. – М.: Либроком, 2009. – 528 с.
49. Финн В.К. Индуктивные методы Д.С. Милля в системах искусственного интеллекта // Искусственный интеллект и принятие решений. Часть I. – 2010. – №3. – С.3 -21; Часть II. – 2010. – №4. – С.14 -40.
50. Забежайло М.И. О некоторых возможностях управления перебором в ДСМ-методе // Искусственный интеллект и принятие решений. Часть I. – 2014. – № 1. – С. 95 -110; Часть II. – 2014. – № 3. – С.3 – 21.
51. Base SAS. High-Performance Procedures. – URL: http://support.sas.com/documentation/cdl/en/prochp/68141/PDF/default/prochp.pdf
52. IBM Cloud Services. – URL: http://www-935.ibm.com/services/us/en/it-services/cloud-services/
53. GENI: Exploring Networks of the Future. – URL: https://www.geni.net
54. HP TippingPoint. – URL: http://www8.hp.com/ru/ru/software-solutions/network-security/index.html
55. Cisco Cloud Security White Papers. – URL: http://www.cisco.com/c/en/us/products/security/cloud-web-security/white-paper-listing.html
56. Intel DPDK: Data Plane Development Kit. – URL: http://dpdk.org/
57. ADI QuickStart SDN Development Kit (SDK) – URL: https://www.sdxcentral.com/products/adi-engineering-gigabit-sdn-quickstart-development-kit/
58. Intel запускает SDN платформу Seacliff Trail. – URL: http://servernews.ru/tags/sdn-платформа

Материал поступил в редакцию 15.09.15.

Сведения об авторах

ГРУШО Александр Александрович – доктор физико-математических наук, профессор, ведущий научный сотрудник Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (ФИЦ ИУ РАН)
e-mail: grusho@yandex.ru

ЗАБЕЖАЙЛО Михаил Иванович – доктор физико-математических наук, ведущий научный сотрудник ВИНИТИ РАН, Москва
e-mail: m.zabezhailo@yandex.ru

ЗАЦАРИННЫЙ Александр Алексеевич – доктор технических наук, профессор, заместитель директора Федерального исследовательского центра «Информатика и управление» РАН (ФИЦ ИУ РАН), Москва
e-mail: alex250451@mail.ru

ПИСКОВСКИЙ Виктор Олегович – кандидат физико-математических наук, ведущий специалист Центра прикладных исследований компьютерных сетей, Москва
e-mail: vpvp80@yandex.ru

БОРОХОВ Сергей Владимирович – старший научный сотрудник Федерального исследовательского центра «Информатика и управление» РАН, Москва
e-mail: sborokhov@yandex.ru

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

УДК 004.81 : 510.64

Н.Д. Лыфенко

Об одном подходе к классификации текстовых данных, основанном на идеях Д.С. Милля

Рассматривается задача автоматической классификации текстовых документов на естественном языке. Для классификации применяется метод, основанный на идеях Д.С. Милля, использующий общие принципы (но не технические детали) ДСМ-метода автоматического порождения гипотез. Описываются эксперименты и оценивается качество работы системы, построенной для реализации описанной методики. При оптимальном подборе опций точность предлагаемого подхода превышает точность других методов.

Ключевые слова: *системы классификации текстов, машинное обучение, интеллектуальный анализ данных, обработка естественного языка, ДСМ-метод*

ВВЕДЕНИЕ

В настоящее время остается актуальной задача автоматической классификации текстовых документов [1]. Ее значимость только увеличивается в связи с экспоненциальным ростом объема текстовой информации, представленной в Интернете, и в связи с развитием технологий обработки больших массивов данных, в том числе документов на естественном языке.

Интерес к этой проблеме не ослабевает. Ведутся коммерческие разработки, например, Irosoft, Artyl's, TextAnalyst, проводится множество исследования и научных конференций [2–4], разрабатываются новые теории представления текста, создаются интегрированные среды разработки, например, RapidMiner, Gate, позволяющие понизить уровень погружения новых пользователей в компьютерную лингвистику.

Одними из популярных и эффективных алгоритмов классификации текстов являются: алгоритм k-ближайших соседей [5], машина опорных векторов [6], наивный байесовский классификатор [7], комбинация простых классификаторов (бустинг) [8]. Однако высокая результативность систем классификации текстов также зависит от качества предварительной обработки документа и выбора модели представления текста.

Во многих системах подобного рода для представления текста в виде вектора признаков используют хорошо зарекомендовавшие себя статистические методы [9], не учитывающие синтаксическую и семантическую связь слов в тексте, и оперирующие частотой встречаемости слов в тексте. Однако наиболее семантически значимые термы, вес которых названные методы стремятся увеличить, могут иметь невысокую частоту и, соответственно, значение в векторе признаков. В таком случае эти термины могут оказаться нерелевантными, что приводит к снижению точности алгоритма классификации. Поэтому в настоящей статье предложен модифицированный метод n-грамм, позволяющий учитывать контекст термов в тексте, а также методика взвешивания термов и варианты методов классификации, которые базируются на идеях Д.С. Милля.

В ходе сравнительного анализа методик токенизации было установлено, что использование предложенного метода приводит к увеличению точности классификации (хотя и не очень большому), но при этом позволяет существенно сократить пространство признаков, и, следовательно, уменьшить вычислительную сложность алгоритмов классификации.

КАНОНЫ ДЖОНА СТЮАРТА МИЛЛЯ. ДСМ-МЕТОД И ПРИМЕНЕНИЕ ИДЕЙ Д.С. МИЛЛЯ К АНАЛИЗУ СТОХАСТИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ

Задача настоящего раздела – показать, каким образом из неформальных правил индуктивной логики Д.С.Милля [10], используя идеи ДСМ-метода В.К.Финна [11], можно получить неформальные правила индукции, подходящие для анализа стохастических данных.

Британский философ, логик, экономист и политолог Джон Стюарт Милль опубликовал свою книгу «Система логики, силлогистической и индуктивной» (A System of Logic, Ratiocinative and Inductive) в 1843 г. Последнее издание русского перевода этой книги вышло в 2011 г. [12]. В середине 1980-х гг. В.К.Финн предложил метод интеллектуального анализа данных, основанный на некоторой формализации правил

индуктивных рассуждений (канонов) Д.С. Милля. Этот метод в честь Джона Стюарта Милля был назван ДСМ-методом автоматического порождения гипотез [13].

Не следует думать, что упомянутая выше формализация была буквальным переводом неформальных правил Д.С.Милля на язык современной математической логики, правила ДСМ-метода В.К.Финна используют лишь общие идеи Д.С.Милля. Формулировка этих правил оригинальна и достаточно сложна. В работах В.К.Финна правила ДСМ-метода формулируются на языке некоторой бесконечнозначной логики с J-операторами Россера-Тьюркетта [14]. В формулировку правил входит так называемое «условие исчерпываемости», которое позволяет определять эффективные (насколько это возможно) алгоритмы анализа данных.

С.О.Кузнецов обнаружил, что обсуждаемые правила можно сформулировать на языке соответствий Галуа [15] и использовать для построения алгоритмов ДСМ-метода развитый аппарат анализа формальных понятий (Formal Concept Analysis) [16].

Современное представление о ДСМ-методе изложено в большой статье В.К. Финна [17].

Однако в настоящей работе не будет применена техника ДСМ-метода, а будут использованы только идеи, лежащие в основе классического ДСМ-метода и непосредственно каноны Милля, адаптированные для случая стохастических зависимостей. Поэтому в нашей статье отсутствуют строгие формулировки правил ДСМ-метода по В.К.Финну. В ней содержится только неформальное описание этих правил, где некоторые важные для техники ДСМ-метода части (например, условие исчерпываемости) будут опущены.

Заметим, что ДСМ-метод предлагался как метод поиска закономерностей детерминистского характера (правил без исключений). Это логико-комбинаторный метод, который хорошо работает на небольших объемах данных и испытывает трудности, когда объем данных вырастает. Поэтому некоторые идеи Д.С.Милля использовались в ДСМ-методе почти буквально. Существенное усложнение формулировки правил было связано прежде всего с алгебраическими и алгоритмическими проблемами. В случае стохастических закономерностей такой буквальный перенос правил Милля уже невозможен. Необходимо будет сначала описать неформально стохастический аналог ДСМ-правил.

ДСМ-метод работает с формализованными описаниями *объектов* предметной области. Такое описание должно представлять *структуру* объекта. Предполагается, что объекты могут обладать или не обладать *целевым свойством*. Целевое свойство может быть *составным*, т.е. представлять собой на самом деле *множество свойств*.

Обычно объект ДСМ-метода представлен в виде множества *атомов*. Такое множество удобно интерпретировать как битовый вектор, а атомы отождествить с *признаками*. Можно расширить этот битовый вектор за счет добавления поля с целевым свойством (тоже считать его признаком). В этом случае мы получаем представление объектов, с которым обычно работают разнообразные методы машинного обучения (интеллектуального анализа данных).

Таким образом, данные для ДСМ-метода могут быть представлены в виде битовой матрицы, в которой объектам соответствуют строки, а признакам (включая целевое свойство) – столбцы.

ДСМ-метод работает по следующему алгоритму:

1) анализируются данные, представляющие собой структурированные описания объектов предметной области, и формируются *гипотезы о возможных причинах наличия* целевого свойства у этих объектов; возможная причина должна быть *фрагментом* (подмножеством) описания объекта; понятно, что фрагмент также может быть представлен в виде битовой строки;

2) аналогичным образом находятся *причины отсутствия целевого свойства*; это тоже должны быть фрагменты описания структуры объекта;

3) с помощью гипотез о возможных причинах формируются *предсказания* о наличии или отсутствии целевого свойства у тех объектов, для которых это было неизвестно;

4) к исходным данным *добавляются* описания тех объектов, у которых было предсказано наличие или отсутствие целевого свойства;

5) шаги 1–4 повторяются до тех пор, пока не будет достигнуто «*условие насыщения*», т.е. такая ситуация, когда уже не удается получить новые предсказания;

6) проверяется, можно ли *объяснить* наличие или отсутствие целевого свойства с помощью гипотез о возможных причинах; если можно, то говорится, что исходные данные удовлетворяют условию «*каузальной полноты*»; нарушение условия каузальной полноты является поводом для модификации данных и изменении параметров ДСМ-метода; после модификации данных происходит переход к шагу 1.

В настоящей работе мы не будем касаться вопросов, связанных с модификацией данных для ДСМ-метода. Подробный анализ таких вопросов в связи с обнаружением эмпирических закономерностей и их классификацией содержится в работе В.К.Финна [18]. Поэтому среди перечисленных этапов работы ДСМ-метода нас будут интересовать шаги 1–3, так как использование предсказаний как новых данных в этой работе также использоваться не будет.

Д.С.Милль в книге [10] сформулировал ряд методов индуктивных рассуждений: метод сходства (Method of Agreement), метод различия (Method of Difference), соединенный метод сходства и различия (Joint Method of Agreement and Difference), метод остатков (Method of Residues), метод сопутствующих изменений (Method of Concomitant Variations). Первоначально ДСМ-метод использовал только правила, построенные на основе метода сходства (и, отчасти, соединенного метода сходства и различия). Формализация остальных правил (канонов) Д.С. Милля была проделана В.К. Финном сравнительно недавно в работе [19]. Система анализа данных, использующая формализации различных канонов Д.С.Милля , была разработана А.Ю. Волковой [20].

Далее будут неформально описаны следующие правила индукции: метод сходства Д.С.Милля [10], соединенный метод сходства и различия Д.С. Милля [10], упрощенные правила индукции ДСМ-метода (с запретом на контрпример и без запрета на контрпример) [21], правила индукции, ориентированные на извлечение стохастических закономерностей.

Метод сходства (первый канон Д.С. Милля) [10]. *Если два или более случаев подлежащего исследованию явления имеют общим лишь одно обстоятельство, то это обстоятельство — в котором только и согласуются все эти случаи — есть причина (или следствие) данного явления.*

Соединенный метод сходства и различия (третий канон Д.С.Милля) [10]. *Если два или более случаев возникновения явления имеют общим лишь одно обстоятельство и два или более случаев невозникновения того же явления имеют общим только отсутствие того же самого обстоятельства, то это обстоятельство, в котором только и разнятся оба ряда случаев, есть или следствие, или причина, или необходимая часть причины изучаемого явления.*

Прежде чем дать неформальное описание правил индукции ДСМ-метода В.К.Финна, введем некоторые термины, использование которых сократит формулировку и коротко опишем особенности подхода В.К. Финна к обнаружению причин целевого свойства.

ДСМ-метод работает с сущностями трех сортов: *объектами*, *фрагментами объектов* и *целевыми свойствами*. *Положительным примером* (для целевого свойства) будем называть объект, обладающий целевым свойством. *Отрицательным примером* будем называть объект, не обладающий целевым свойством. В правилах Д.С. Милля положительному примеру соответствует «случай возникновения явления», а отрицательному примеру – «случай невозникновения явления».

В качестве возможных причин как наличия, так и отсутствия целевого свойства будут рассматриваться *фрагменты* объектов. Причина наличия свойства будет называться *положительной причиной* или плюс-причиной. Причина отсутствия свойства будет называться *отрицательной причиной*, минус-причиной или антипричиной свойства.

Подчеркнем, что в ДСМ-методе и причины наличия, и причины отсутствия свойств – это факты *наличия* (а не отсутствия) в объекте определенных фрагментов. С этим обстоятельством связаны некоторые технологические проблемы, которые, впрочем, легко разрешаются.

Упрощенные правила индукции ДСМ-метода (без запрета на контрпример). (Плюс-правило) *Пусть f – общий фрагмент двух или более положительных примеров для целевого свойства и f не является общим фрагментом для двух или более отрицательных примеров для этого же свойства. Тогда f – возможная причина наличия целевого свойства.*

(Минус-правило) *Пусть f – общий фрагмент двух или более отрицательных примеров для целевого свойства и f не является общим фрагментом для двух или более положительных примеров для этого же свойства. Тогда f – возможная причина отсутствия целевого свойства.*

Если рассматривать плюс-правило, то первое условие посылки этого правила такое же, как в методе сходства (и объединенном методе сходства и различия) Д.С.Милля, а второе условие несколько слабее, чем в соединенном методе сходства и различия. Минус-правило формулируется симметрично.

Упрощенные правила индукции ДСМ-метода (с запретом на контрпример). (Плюс-правило) *Пусть f – общий фрагмент двух или более положительных примеров для целевого свойства и f не входит ни в один отрицательный пример для этого свойства. Тогда f – возможная причина наличия целевого свойства.*

(Минус-правило) *Пусть f – общий фрагмент двух или более отрицательных примеров для целевого свойства и f не входит ни в один положительный пример для этого свойства. Тогда f – возможная причина отсутствия целевого свойства.*

В правилах с запретом на контпример второе условие в посылке более жесткое и быстрее проверяемое, чем в правилах без запрета на контрпример. Но оно, опять же, не совпадает со вторым условием из соединенного метода сходства и различия.

Таким образом, правила индукции ДСМ-метода, предложенного В.К. Финном [13], вообще говоря, отличны от канонов Милля, но используют те же идеи сходства и различия.

Использование идей ДСМ-метода для анализа статистических закономерностей

Теперь попробуем предложить неформальные описания правил, подходящих для извлечения закономерностей из стохастических данных.

Заметим, что правила ДСМ-метода «говорят», например, о «двух или более положительных примерах» и «отсутствии отрицательных примеров». В случае стохастических данных так определенно нельзя сказать. Неформально следует описать правила, используя «нечеткие» ограничения, такие как «много» или «мало», которые впоследствии формализуются с помощью статистических критериев.

Неформальное описание аналогов правил ДСМ-индукции для стохастических данных. (Плюс-правило) *Пусть существует достаточно много случаев вхождения фрагмента f в положительные примеры (для целевого свойства) и достаточно мало случаев вхождения фрагмента f в отрицательные примеры. Тогда f – возможная причина наличия целевого свойства.*

(Минус-правило) *Пусть существует достаточно много случаев вхождения фрагмента f в отрицательные примеры (для целевого свойства) и достаточно мало случаев вхождения фрагмента f в положительные примеры. Тогда f – возможная причина отсутствия целевого свойства.*

Легко заметить, что приведенные выше неформальные правила суть «нечеткие» варианты правил ДСМ-метода с запретом на контрпример.

Правила ДСМ-индукции, которые могли бы использовать статистические соображения, вводились и ранее, например, в работах [22–25]. Строгое формально-логическое описание таких правил должно использовать обобщенные статистические кванторы в стиле П. Гаека и Т. Гавранека [26] (см. также [22–25]).

Однако во всех перечисленных работах предполагалось использование стандартной техники ДСМ-метода, которая для большого объема данных неудобна из-за ее высокой вычислительной сложности. В настоящей работе предлагается совершенно другой, оригинальный, подход, основанный на формировании «типичного (обобщенного) положительного примера» и «типичного (обобщенного) отрицательного примера». Технические детали этого подхода будут изложены позже.

Опишем теперь неформальные правила порождения предсказаний о наличии или отсутствии целевых свойств у тех объектов, для которых это неизвестно. Такие правила в ДСМ-методе называются *правилами аналогии*. Это название связано с тем, гипотеза о том, что объект обладает свойством, выдвигается в том случае, когда он похож на два или более объекта, обладающие этим свойством, и не похож на объекты этим свойством не обладающие. А сходство объектов как раз и выражается их общим фрагментом, который был определен как причина (наличия или отсутствия) целевого свойства.

Упрощенные правила аналогии ДСМ-метода. (Плюс-правило) *Пусть неизвестно, обладает ли объект o целевым свойством. Пусть также объект o содержит какую-либо причину наличия целевого свойства и не содержит ни одной причины отсутствия целевого свойства. Тогда объект o обладает целевым свойством.*

(Минус-правило) *Пусть неизвестно, обладает ли объект o целевым свойством. Пусть также объект o содержит какую-либо причину отсутствия целевого свойства и не содержит ни одной причины наличия целевого свойства. Тогда объект o не обладает целевым свойством.*

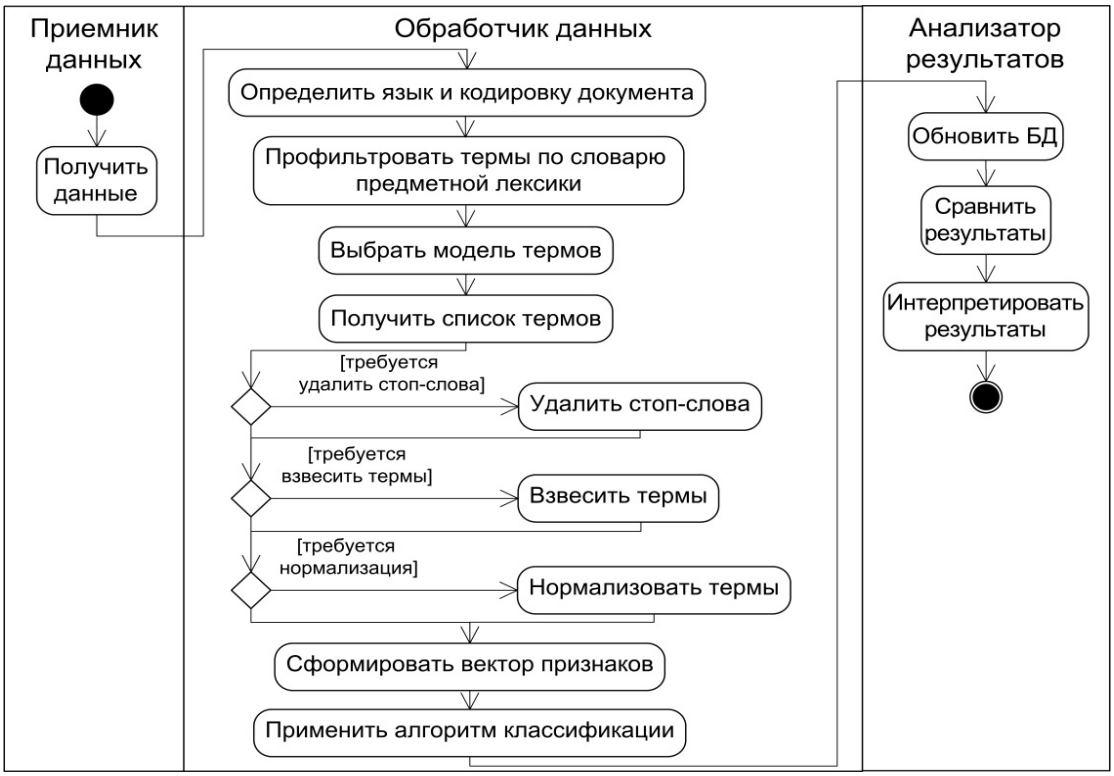
Для стохастического случая правило аналогии можно сформулировать точно так же, если предполагать использование традиционной ДСМ-техники.

В нашей же работе сходство объекта o с объектами, обладающих целевым свойством, будет оцениваться как мера близости объекта o к обобщенному объекту – представителю класса объектов, обладающих целевым свойством. Будет считаться, что объект обладает свойством, если эта мера превышает некоторый порог. Конкретная формулировка правил аналогии для рассматриваемого в настоящей работе случая будет дана ниже.

Завершая раздел, следует отметить, что в нем были изложены лишь идеи и неформальные описания правил. Конкретная их интерпретация, уточнение и формализация будут проделаны в разделе «Алгоритм классификации, основанный на идеях Д.С.Милля».

КОНЦЕПТУАЛЬНАЯ СХЕМА СИСТЕМЫ

На рисунке представлена концептуальная схема системы, решающая задачу автоматической классификации текстовых документов на естественном языке.



Концептуальная модель системы автоматической классификации документов

В настоящей статье нас будут интересовать следующие этапы анализа текста: нормализация, выделение и взвешивание термов, классификация.

Нормализация

В качестве объекта исследования во многих проектах по компьютерной лингвистике выступают тексты на иностранном языке, поэтому большая часть технологий, ориентированных на язык документа, является для русского языка слабо развитой. Для этапа нормализации, приведения входного слова к нормальной форме, используется морфологический анализ. Обычно этот этап реализуется с помощью алгоритма, оперирующего списком правил, морфологического словаря или некоторого комбинированного подхода.

Одними из наиболее популярных решений для нормализации слов русского языка являются: стеммер Snowball [26], идеи алгоритма которого базируются на стеммере Портера, использующего набор морфологических правил конкретного языка, лингвистические процессоры из программного пакета ДИАЛИНГ[1], программа mystem[2] [27], морфологический анализатор pymorphy2, использующий словари из OpenCorpora [28].

В качестве базового решения мы используем словарь, созданный на основе COM-объекта, реализующего морфологический анализ в проекте АОТ [29], но в настройках системы есть возможность выбрать инструментарий для нормализации: pymorphy2, mystem или упомянутый выше словарь. После нормализации с помощью словаря для слов, не вошедших в словарь, строится множество гипотетических нормальных форм: производится поиск максимальной подстроки в строке, состоящей из слова, записанного в обратном порядке. Для остальных слов, которые не удалось нормализовать подобным способом, применяется алгоритм стеммера Портера.

Данными для эксперимента служил корпус из более 52000 текстовых файлов, суммарным объемом 85 МБ и включающий 12 млн словоформ. Результаты эксперимента представлены в табл. 1.

В результате было установлено, что большая часть лексем содержится в словаре, а не вошедшие в него слова – это слова на иностранном языке или ошибки, порожденные на этапе токенизации.

Таблица 1

Эксперимент по нормализации

Метод нормализации	Количество словоформ (шт)	Количество словоформ (%)
словарь	9568917	77,9
максимальная подстрока	382344	3,2
алгоритм Портера	2320302	18,9

Отбор признаков (feature selection)

Для работы с текстовыми документами обычно используют их векторное представление (vector space model) и модель «мешка слов» (bag of words) для интерпретации текста[30]. Наиболее популярными признаками для классификации являются:

- *n*-граммы [31, 32], т.е. последовательности символов длины *n*. Подобное упрощенное представление не учитывает ни синтаксической, ни семантической информации, но при этом использует простую математическую модель [33];
- последовательности слов, как частный случай n-грамм;
- предложения;
- самые частотные последовательности символов.

В текущем исследовании признаками являются термы, которые представляют собой последовательность слов длины 1,2 или 3, состоящую из имён существительных, прилагательных и глаголов. Особенностью данного подхода является попытка учёта левого и правого контекста рассматриваемого слова.

В любом случае, мы будем предполагать, что перед тем как производить классификацию, мы должны сформировать и зафиксировать *словарь термов*, конечное, но достаточно большое, множество термов, которое мы будем в этой работе обозначать через T. Для последующей классификации мы отбираем статистически значимые термы, имеющие высокую документную частоту в выборке. Для каждого слова анализируется слова, стоящие слева и справа от него, которые в дальнейшем сформируют сам терм. Далее для каждого $t \in T$ подсчитывается документная частота терма t в классе D_c по следующей формуле:

$$df\left(t, D_c\right) = \frac{\left|\{dD_c \mid td\}\right|}{\left|D_c\right|}, \qquad (1)$$

где d – документ из класса D_c, $\left\{d \in D_c | t \in d\right\}$ – множество документов из класса D_c, содержащих терм t, через $\left|A\right|$ обозначается мощность множества A.

На основе сравнительного анализа, данного в [34], можно говорить, что документная частота является простым, вычисляемым за линейное время, но в тоже время эффективным признаком для классификации.

[1] Решение реализовано в виде COM-объекта, доступного на сайте aot.ru (на момент написания статьи сайт не функционировал)

[2] Относительно свободная версия программы mystem появилась гораздо позже, чем описанный выше лингвистический процессор, который с 2010 г. распространяется на условиях лицензии LGPL, а его исходные коды доступны в SVN-репозитории.

Таким образом, подсчитываем отношение числа документов, в которых встречается терм, ко всему множеству документов из этого класса.

Далее отбираем термы с весом, не превосходящим заданное значение, установленное эмпирически и в среднем не превосходящее 10% от всех *n*-грамм в классе. Такой подход позволяет выбрать существенные, по крайне мере статистически значимые термы, имеющие большой вес и влияющие на классификацию, а также сократить пространство признаков за счет отсутствия шумовых термов.

В свою очередь, классическими способами получения численных значений у признаков являются: частотный и бинарный подходы, мера tf-idf с различными вариантами, самая частотная последовательность [35], учет взаимной информации (IG), статистика хи-квадрат и др.

Представление текста в виде вектора признаков

Пусть t – терм из некоторого конечного множества термов T, d – документ из некоторого класса документов. Вес терма t в документе d обозначим через $w(t,d)$ и будем определять по формуле:

$$w(t,d)=\begin{cases}1, t\in d,\\ 0, t\notin d.\end{cases} \qquad (2)$$

Таким образом, вектор весов, представляющий документ, будет бинарным вектором. Для удобства в дальнейшем будем отождествлять векторы с кортежами, индексированными множеством термов T, т.е. будем представлять вектор в виде отображения $v: T\to\{0,1\}$ (бинарный вектор), $v: T\to\mathbf{R}$ (вещественный вектор, здесь $\mathbf{R}$ – множество вещественных чисел) или как частный случай вещественного вектора – вектор частот, отображение $v: T\to[0,1]$.

Обозначим через d^v документ, представленный вектором v. В этом случае, очевидно, верно соотношение

$$v(t)=w(t,d^v), \qquad (3)$$

для любого терма $t\in T$.

Для более точных предсказаний будем учитывать не только наличие вхождения термов в документ, но и отсутствие таких вхождений. Аналогичный технический прием часто используется при практическом использовании ДСМ-метода. Для этого расширим исходное множество термов T до множества $\hat{T}=T\cup\tilde{T}$ за счет добавления «антитермов» из множества $\tilde{T}=\{\tilde{t}\ |t\in T\}$, для которых

$$v(\tilde{t})=1-v(t). \qquad (4)$$

АЛГОРИТМ КЛАССИФИКАЦИИ, ОСНОВАННЫЙ НА ИДЕЯХ Д.С.МИЛЛЯ

Индукция (фаза обучения)

Возьмем некоторую предметную область. Рассмотрим обучающую коллекцию документов для бинарной классификации. В этой коллекции выделим два класса:

- класс документов, являющихся положительными примерами, т.е. документов, относящихся к некоторой *интересующей нас подобласти* рассматриваемой предметной области, обозначим этот класс через D_{pos},
- класс документов, *не относящихся* к интересующей нас подобласти, т.е. класс отрицательных примеров, обозначим его через D_{neg}.

Фаза обучения будет соответствовать *этапу индукции* в ДСМ-методе. В этой фазе формируются два *вектора-экземпляра* для двух классов документов D_{pos} и D_{neg}.

Вектор-экземпляр – это обобщенный (родовой) вектор, представляющий целый класс документов. Вектор-экземпляр класса D_{pos} должен в неявной форме содержать все возможные *причины наличия* целевого свойства, а вектор-экземпляр класса D_{neg}, должен в неявной форме содержать все возможные *причины отсутствия* целевого свойства. Через Pos будем обозначать вектор-экземпляр класса D_{pos}, а через Neg – вектор-экземпляр класса D_{neg}.

Методы формирования векторов-экземпляров классов могут быть разными. Общая идея – усилить и подчеркнуть факт достаточно частого вхождения терма в документы класса. Если терм *часто* входит в документы класса D_{pos} и *редко* – в документы класса D_{neg}, то он должен получить в векторе Pos высокий вес. Этот подход вполне соответствует неформальному правилу индукции: фрагмент, претендующий на то, чтобы быть возможной причиной наличия свойства, должен содержаться в достаточно большом количестве положительных примеров и в достаточно малом количестве отрицательных примеров. Высокий вес как раз имеют термы, являющиеся элементами возможных причин: они часто содержатся в положительных примерах (документах класса D_{pos}) и редко – в отрицательных примерах (документах класса D_{neg}). Заметим, что термы, содержащиеся в положительных и отрицательных примерах примерно с одинаковой частотой, должны получить вес, близкий к 0. Такой вес должны, например, получить стоп-слова.

Вектор-экземпляр Neg формируется симметрично. Большой вес в этом векторе должны получить термы, часто встречающиеся в D_{neg} и редко – в D_{pos}.

Пусть

$$df_{neg}(t) = df(t, D_{neg}), \quad (5)$$

$$df_{neg}(t) = df(t, D_{neg}), \quad (6)$$

где $t \in T$. Для $\tilde{t} \in \tilde{T}$ положим

$$df_{pos}(\tilde{t}) = 1 - df_{pos}(t), \quad (5\text{ a})$$

$$df_{neg}(\tilde{t}) = 1 - df_{neg}(t). \quad (6\text{ b})$$

Конкретные формулы для вычисления Pos и Neg могут быть разными. Приведем несколько возможных вариантов.

Вариант 1.

$$Pos^{1}(t) = df_{pos}(t) - df_{neg}(t), \quad (7)$$

$$Neg^{1}(t) = df_{neg}(t) - df_{pos}(t), \quad (8)$$

где $t \in \hat{T} = T \cup \tilde{T}$.

В этом случае возможны отрицательные значения координат векторов-экземпляров. Если частоты вхождения терма в положительные и отрицательные примеры примерно равны, то вес терма будет близок к 0. Чтобы веса всех термов заведомо были бы неотрицательными, можно использовать Вариант 1+.

Вариант 1+.

$$Pos^{1+}(t) = df_{pos}(t) \dot{-} df_{neg}(t), \quad (9)$$

$$Neg^{1+}(t) = df_{neg}(t) \dot{-} df_{pos}(t), \quad (10)$$

где $t \in \hat{T}$. Здесь $\dot{-}$ – усеченная разность, т.е.,

$$x \dot{-} y = \begin{cases} x - y, & x \geq y, \\ 0, & x < y. \end{cases} \quad (11)$$

Вариант 2.

$$Pos^{2}(t) = \frac{df_{pos}(t) + \varepsilon}{df_{neg}(t) + \varepsilon} - 1, \quad (12)$$

$$Neg^{2}(t) = \frac{df_{neg}(t) + \varepsilon}{df_{pos}(t) + \varepsilon} - 1, \quad (13)$$

где $t \in \hat{T}$. Здесь $\varepsilon \in (0,1]$ – «параметр чувствительности». С одной стороны, наличие слагаемого ε не позволяет возникнуть ошибке деления на 0, с другой стороны, величина ε задает «чувствительность к малым частотам». Чем меньше ε, тем меньшие частоты могут оказывать влияние на значения координат вектора. Нетрудно убедиться в том, что координаты векторов Pos и Neg для Варианта 2 принадлежат отрезку $\left[-\frac{1}{1+\varepsilon}, \frac{1}{\varepsilon}\right]$. Эти координаты могут быть отрицательными, а в случае примерного равенства частот $df_{pos}(t)$ и $df_{neg}(t)$ координаты обсуждаемых векторов будут близки к 0. Чтобы добиться неотрицательных координат в любом случае, можно воспользоваться Вариантом 2+.

Вариант 2+.

$$Pos^{2+}(t) = \frac{df_{pos}(t) + \varepsilon}{df_{neg}(t) + \varepsilon} \dot{-} 1, \quad (14)$$

$$Neg^{2+}(t) = \frac{df_{neg}(t) + \varepsilon}{df_{pos}(t) + \varepsilon} \dot{-} 1, \quad (15)$$

где $t \in \hat{T}, \varepsilon \in (0,1]$.

Наконец, можно предложить версию Варианта 2+, в которой координаты векторов Pos и Neg будут нормированы.

Вариант 2Norm.

$$Pos^{2\text{Norm}}(t) = \left(\frac{df_{pos}(t) + \varepsilon}{df_{neg}(t) + \varepsilon} \dot{-} 1 \right) \cdot \varepsilon, \quad (16)$$

$$Neg^{2\text{Norm}}(t) = \left(\frac{df_{neg}(t) + \varepsilon}{df_{pos}(t) + \varepsilon} \dot{-} 1 \right) \cdot \varepsilon, \quad (17)$$

где $t \in \hat{T}, \varepsilon \in (0,1]$.

Аналогия (фаза тестирования)

Фаза тестирования соответствует рассуждениям по аналогии ДСМ-метода. В этой фазе формируются предсказания: принадлежит ли новый вектор требуемому классу, т.е. обладает ли он целевым свойством.

Неформальное правило аналогии для нашего случая можно сформулировать следующим образом: *Объект обладает целевым свойством, если он существенно больше похож на положительные примеры (для этого свойства), чем на отрицательные*.

Обобщенным представителем класса положительных примеров является вектор-экземпляр Pos, а обобщенным представителем класса отрицательных примеров является вектор-экземпляр Neg. Отношение сходства нового объекта с положительными и отрицательными примерами выражается через меру сходства векторного представления этого объекта и векторов экземпляров Pos и Neg.

В нашем случае объекты – это документы, а в качестве меры сходства будем использовать косинус угла между векторными представлениями документов, выраженный через скалярное произведение, и модули векторов.

Обозначим через $\mathrm{sim}\left(\vec{a},\vec{b}\right)$ меру сходства между векторами $\vec{a}$ и $\vec{b}$. Тогда

$$\mathrm{sim}\left(\vec{a},\vec{b}\right)=\frac{\left(\vec{a},\vec{b}\right)}{\left|\vec{a}\right|\cdot\left|\vec{b}\right|}. \quad (18)$$

Тот факт, что вектор v, представляющий документ d^v, существенно более похож на вектор-экземпляр Pos, чем на вектор-экземпляр Neg, можно выразить следующей формулой

$$\mathrm{sim}\left(v,Pos\right)-\mathrm{sim}\left(v,Neg\right)\geq p^{+}, \quad (19)$$

где $p^{+}\in\left(0,1\right]$ – некоторое пороговое значение.

Симметрично, формулой

$$\mathrm{sim}\left(v,Neg\right)-\mathrm{sim}\left(v,Pos\right)\geq p^{-}, \quad (20)$$

где $p^{-}\in\left(0,1\right]$, будет выражен тот факт, что новый вектор v существенно более похож на отрицательные примеры, чем на положительные.

Таким образом, можно сформулировать следующие правила отнесения нового документа, представленного вектором v к тому или иному классу:

- $d^v\in D_{pos}$, если справедлива формула (19);
- $d^v\in D_{neg}$, если справедлива формула (20).

Нетрудно убедиться, что условия (19) и (20) одновременно истинными быть не могут. Кроме того, возможна ситуация, когда некоторые документы не могут быть с уверенностью отнесены ни к одному из классов. Это типичная ситуация для ДСМ-метода. Она означает, что нам осталось неизвестным, обладает или не обладает новый объект целевым свойством.

Возможны и другие правила для определения принадлежности к классам D_{pos} и D_{neg}.

Обозначим через $SPos$ бинарный вектор, определенный следующим образом

$$SPos\left(t\right)=\mathrm{sign}\left(Pos\left(t\right)\right)=\begin{cases}1, & Pos\left(t\right)>0,\\ 0, & Pos\left(t\right)\leq 0.\end{cases} \quad (21)$$

Аналогично, через $SNeg$ обозначим бинарный вектор определенный следующим образом:

$$SNeg\left(t\right)=\mathrm{sign}\left(Neg\left(t\right)\right)=\begin{cases}1, & Neg\left(t\right)>0,\\ 0, & Neg\left(t\right)\leq 0.\end{cases} \quad (22)$$

Через $\overline{SPos}$ обозначим инверсию вектора $SPos$, т.е.

$$\overline{SPos}\left(t\right)=\neg\left(SPos\left(t\right)\right)=\begin{cases}1, & SPos\left(t\right)=0,\\ 0, & SPos\left(t\right)=0.\end{cases} \quad (23)$$

Аналогично, через $\overline{SNeg}$ обозначим инверсию вектора $SNeg$:

$$\overline{SNeg}\left(t\right)=\neg\left(SNeg\left(t\right)\right)=\begin{cases}1, & SNeg\left(t\right)=0,\\ 0, & SNeg\left(t\right)=0.\end{cases} \quad (23a)$$

Введем следующие обозначения:

$$C_{pos}\left(v\right)=\left(SPos,v\right), \quad (24)$$

$$C_{neg}\left(v\right)=\left(SNeg,v\right), \quad (25)$$

$$\overline{C_{pos}}\left(v\right)=\left(\overline{SPos},v\right), \quad (26)$$

$$\overline{C_{neg}}\left(v\right)=\left(\overline{SNeg},v\right). \quad (27)$$

Здесь через $\left(a,b\right)$ обозначено скалярное произведение векторов a и b. Очевидно, что $C_{pos}\left(v\right)$ – это мощность пересечения множеств термов, представленных векторами $SPos$ и v. Аналогично могут быть интерпретированы и скалярные величины $C_{neg}\left(v\right)$, $\overline{C_{pos}}\left(v\right)$, $\overline{C_{neg}}\left(v\right)$.

На основе сказанного выше можно сформулировать следующие правила, определяющие принадлежность неизвестного вектора v классам D_{pos} и D_{neg}:

- $d^v\in D_{pos}$, если $C_{pos}\left(v\right)+\overline{C_{neg}}\left(v\right)>0,5+\theta_{pos}$ и $C_{neg}\left(v\right)+\overline{C_{pos}}\left(v\right)<\theta_{neg}$ (28)
- $d^v\in D_{neg}$, если $C_{neg}\left(v\right)+\overline{C_{pos}}\left(v\right)>0,5+\theta_{neg}$ и $C_{pos}\left(v\right)+\overline{C_{neg}}\left(v\right)<\theta_{pos}$, (29)

где θ_{pos} и θ_{neg} – пороговые значения для соответствующих классов, которые выбираются при настройке системы.

Наличие таких порогов представляется необходимым, так как при решении задачи четкой классификации зачастую вероятность принадлежности документа к классу может быть около 50%, что позволяет говорить о некоторой неуверенности принятия решения об отнесении к какому-либо классу. Пороги позволяют сформировать класс, в котором будут содержаться документы, о принадлежности которых система не может сообщить с достаточно высокой уверенностью. Эмпирически было установлено, что наиболее релевантные результаты система выдает при значении порогов в следующих диапазонах: $\theta_{pos}\in\left[0;5\right]$, $\theta_{neg}\in\left[0;7\right]$.

Следует отметить, что при таком подходе для представления текста в виде вектора признаков ника-

ким образом не используется статистическая информация о частоте встречаемости термов, основным критерием служит только факт наличия или отсутствия некоторого множества причин (термов) для классификации. Документная частота применяется только на этапе формирования эталонного вектора для каждого класса.

В дополнение к описанным выше, некоторые простые правила классификации можно сформулировать следующим образом:

- $d^v \in D_{pos}$, если $C_{pos}(v) > C_{neg}(v)$; (30)
- $d^v \in D_{neg}$, если $C_{neg}(v) > C_{pos}(v)$. (31)

Более гибкие правила можно получить, включив в формулы вес положительных и отрицательных предсказаний:

- $d^v \in D_{pos}$, если $C_{pos}(v) \cdot w_{pos} > C_{neg}(v)$; (32)
- $d^v \in D_{neg}$, если $C_{neg}(v) \cdot w_{neg} > C_{pos}(v)$, (33)

где w_{pos} и w_{neg} – весовые коэффициенты для положительных и отрицательных предсказаний.

ЭКСПЕРИМЕНТЫ

Для получения релевантных результатов классификации требуется точная настройка работы системы, которая включает следующий набор булевых опций:

– использование стоп-слов в векторе признаков;

– фильтрация термов;

– формирование термов, состоящих только из именных групп;

– нормализация слов;

– фильтрация цифр.

Не для всех задач классификации удаление стоп-слов оказывает положительное влияние на увеличение точности работы алгоритма классификации. В работах [36, 37] было показано, что использование стоп-слов в качестве признаков (наряду с другими) для решения задачи определения авторства документа повышает точность классификации в среднем на 8%.

Для сокращения пространства признаков и формирования более осмысленного, репрезентативного и содержательного набора признаков классификации, целесообразно не учитывать низкочастотные термы, вес которых ниже установленного порога для конкретной выборки. В качестве такого значения может выступать среднее значение частоты встречаемости терма в коллекции, вычисляемое по формуле:

$$\theta(D) = \frac{\sum_{t \in D_c} df(t, D_c)}{|D_c|} .$$

В силу особенностей естественного языка основной «смысл предложения», т.е. набор термов (признаков для классификации), содержится в подлежащем, которое чаще всего выраженно в русском языке именем существительным, поэтому для задачи тематической классификации поиск именных групп в предложении и, соответственно, более детальный анализ такого терма представляется перспективным.

Классическим этапом при формировании вектора признаков классификации стала нормализация слов, которая позволяет отобразить разные словоформы в одну лексему. Чаще всего для задач автоматической классификации использование этого этапа приводит к повышению точности и полноты классификации, позволяет сократить пространство признаков, что для некоторых алгоритмов классификации уменьшает время работы.

В ходе экспериментов на данных из спортивных предметных областей стало очевидно, что некоторые числительные, употребляемых в тексте в виде чисел, требуют особого подхода. Так, например, число 11 без контекста вряд ли может служить весомой причиной для отнесения к какому-либо классу, тоже можно сказать и про прилагательное *метровый.* Но если сформировать терм «11-метровый», то, вероятнее всего, документ, содержащий такой терм, относится к спортивной тематике.

Основная задача, которая решается во время проведения экспериментов, состоит в нахождении наиболее значимых опций, применяемых для настройки системы, которые, в свою очередь, позволяют изменять качество и число рассматриваемых признаков для классификации. Основными настройками являются[3] (согласно условным обозначениям):

- использование стоп-слов в векторе признаков (isSW);
- фильтрация термов (isFilter);
- формирование термов состоящих только из именных групп (isNP);
- нормализация слов (isNorm);
- фильтрация цифр (isNum).
- Набор меток методов классификации:
- c – использование значение косинуса угла, описано условиями (19) и (20),
- e – расширение вектора признаков документа, формулы для принятия решения о принадлежность или не принадлежности документа классу описаны в соответствующем разделе (28) и (29),
- w – взвешенное количество пересечений, вычисляемое по (32) и (33),
- u – количество пересечений, вычисляемое по формулам (30) и (31).

Последующие серии экспериментов устроены следующим образом: для текстовой коллекции документов будут подсчитаны различные метрики эффективности методов (точность, аккуратность, полнота, общая ошибка, ошибка для каждого класса) в зависимости от используемых настроек системы.

Для эксперимента была собрана тестовая коллекция из 6012 текстовых документов. 3874 – документа для положительного класса и 2138 документа – для отрицательного класса соответственно. Перед

[3] Для каждой опции приведено ее краткое обозначение, используемого в дальнейшем исследовании.

системой стояла задача провести тематическую классификацию 2575 документов, которые еще не имеют метки класса. Под положительными документами подразумеваются тексты, относящиеся к политическим темам, отрицательной коллекции принадлежат тексты, не имеющие отношения к политике, т.е. класс «остальное», в котором содержатся документы про спорт, экономику, происшествия, культуру, здоровье. Этот массив документов был получен автоматическим скачиванием новостной ленты с интернет-издания газеты «Аргументы и факты» в период с 10.02.2014 по 10.08.2014. Для каждого документа метка класса была известна, исходя из данных RSS-ленты. Далее полученная выборка была разделена на обучающую и тестовую выборки согласно описанным ранее правилам и было проведено обучение классификатора по формулам варианта 1,т.е.(7), (8).

Как было сказано, для оценки качества работы системы используются классические метрики оценки: *точность, полнота, ошибка, f-мера* [38].

Результаты проведённых экспериментов представлены в табл. 2 для наиболее важных комбинаций настроек системы (isNorm, isSW, isFilter, isNP, isDigit) и рассматриваемых методов (c,e,u,w).

Таблица 2

Результаты оценки качества системы

isNorm	isSW	isFilter	isNP	isDigit	Cls	Ошибка	Точность	Полнота	F-мера
1	0	0	0	0	c	19,00	0,65	0,92	0,74
1	1	0	0	0	c	15,00	0,64	0,91	0,75
1	0	0	1	0	c	17,00	0,65	0,90	0,75
1	0	1	0	0	c	19,00	0,68	0,85	0,75
1	1	1	0	0	c	15,00	0,64	0,91	0,75
1	1	1	1	0	c	54,00	0,33	0,75	0,46
1	0	0	0	1	c	17,00	0,63	0,90	0,73
1	0	0	0	0	e	18,00	0,62	0,92	0,74
1	1	0	0	0	e	16,00	0,64	0,91	0,75
1	0	0	1	0	e	17,00	0,64	0,90	0,75
1	0	1	0	0	e	18,00	0,63	0,90	0,74
1	1	1	0	0	e	17,00	0,65	0,89	0,75
1	**1**	**1**	**1**	**0**	**e**	**18,00**	**0,78**	**0,69**	**0,73**
1	0	0	0	1	e	18,00	0,62	0,92	0,74
1	0	0	0	0	u	14,00	0,62	0,95	0,75
1	1	0	0	0	u	13,00	0,63	0,95	0,75
1	0	0	1	0	u	14,00	0,62	0,95	0,75
1	0	1	0	0	u	13,00	0,62	0,95	0,75
1	1	1	0	0	u	13,00	0,62	0,95	0,75
1	1	1	1	0	u	64,00	0,35	0,46	0,74
1	0	0	0	1	u	22,00	0,58	0,74	0,65
1	0	0	0	0	w	18,00	0,62	0,92	0,74
1	1	0	0	0	w	16,00	0,64	0,91	0,75
1	0	0	1	0	w	17,00	0,64	0,90	0,75
1	0	1	0	0	w	18,00	0,62	0,91	0,74
1	1	1	0	0	w	16,00	0,65	0,90	0,75
1	1	1	1	0	w	56,00	0,26	0,78	0,39
1	0	0	0	1	w	18,00	0,62	0,92	0,74

Следует подчеркнуть, что для методов w,u,c уменьшение размера вектора признаков за счёт нормализации, учёта стоп-слов, фильтрации и использования только именных групп существенно снижает точность классификации (в среднем на 0,3). Для предложенного метода данный показатель, напротив, увеличивается (0,65/0,78), что позволяет говорить о перспективность используемого подхода.

Следует отметить, что в данном случае решается не классическая задача бинарной классификации, а некоторое ее усложнение. Не отнесение к одному из классов не является причиной для отнесения к другому, так как, как было сказано выше, в нашем исследовании используется третий класс так называемых сомнительных документов.

В результате проведенных экспериментов удалось установить, что точность и полнота предлагаемого подхода к классификации соизмерима с результатами для вариантов *c,u,w*, а в некоторых случаях превосходит эту оценку.

В дальнейшем предполагается провести серию экспериментов на репрезентативных данных других предметных областей, модифицировать алгоритмы принятия решения о принадлежности документа к классу и реализовать алгоритм ДСМ-системы более близкий к классическому.

СПИСОК ЛИТЕРАТУРЫ

1. Sebastiani F. “Text categorization”, // Text Mining and its Applications / ed. Alessandro Zanasi. – Southampton: WIT Press, UK, 2005 – P. 109–129.
2. TextAnalyst. – URL: http://www.megaputer.com /site/textanalyst.php (дата обращения 12.03.2015).
3. Irosoft . Automatic document classification module docutheque entreprise. – URL: http://www.irosoft.com/en/communiques-presse/irosoft-adds-automatic-document-classification-module-docutheque-entreprise (дата обращения 10.04.2015).
4. Automatic Document Classification with Artsyl’s docAlpha. – URL: http://www.artsyltech.com/da_classification.html (дата обращения 26.04.2015).
5. Yang Y. An evaluation of statistical approaches to text categorization // Information Retrieval. – 1999. – Vol.1 (1-2). – P. 69–90.
6. Joachims T. Text categorization with suport vector machines: Learning with many relevant features // In Proceedings of European Conference on Machine Learning. – 1998. – P. 137–142.
7. McCallum, K. Nigam. A comparison of event models for naive bayes text classication // In AAAI-98 Workshop on Learning for Text Categorization – 1998.
8. Schapire R. E., Singer Y. Boostexter: A boosting-based system for text categorization // Machine Learning. – 2000. – №39 – P. 135–168.
9. Bai J., Nie J.-Y. Using language models text classification // In Proceedings of Asia Information Retrieval Symposium, Beijing – 2004.
10. Mill J.S. A System of Logic, Ratiocinative and Inductive. – NY.: Harper & Brothers, 1882. – 1176 с.
11. Финн В.К. Базы данных с неполной информацией и новый метод автоматического порождения гипотез // Диалоговые и фактографические системы информационного обеспечения. – М., 1981. – С. 153–156.
12. Милль Д.С. Система логики силлогистической и индуктивной. – М.: ЛЕНАНД, 2011. – 832 с.
13. Финн В.К. О машинно-ориентированной формализации правдоподобных рассуждений в стиле Ф.Бэкона – Д.С.Милля // Семиотика и информатика. – 1983. – Вып. 20. – С. 35–101.
14. Rosser J.B., Turquette A.R. Many-valued logics. – Amsterdam: North-Holland, 1951.
15. Кузнецов С.О. ДСМ-метод на языке соответствий Галуа // Научно-техническая информация. Сер 2. – 2006. – № 12. – С. 1–7.
16. Ganter B., Wille R. Formal Concept Analysis: Mathematical Foundations. – Berlin: Springer-Verlag, 1999.
17. Финн В.К. Эпистемологические основания ДСМ-метода автоматического порождения гипотез // Научно-техническая информация. Сер.2. – 2013. – Часть I, № 9, С.1–29; Часть II, №12, С.1–26
18. Финн В.К. Об определении эмпирических закономерностей посредством ДСМ - метода автоматического порождения гипотез // Искусственный интеллект и принятие решений. – 2010. – № 4. – С. 41–48.
19. Финн В.К. Индуктивные методы Д.С. Милля в системах искусственного интеллекта. // Искусственный интеллект и принятие решений. – 2010. – Часть I, № 3. – С. 3–21; Часть II, № 4. – С. 14–40,
20. Волкова А.Ю. Алгоритмизация процедур ДСМ-метода автоматического порождения гипотез // Научно-техническая информация. Сер 2. – 2011. – № 5. – С. 6–12.
21. Аншаков О.М. ДСМ-метод: теоретико-множественное объяснение // Научно-техническая информация. Сер. 2. – 2012. – № 9. – С. 1–19.
22. Григорьев П.А. Об одном методе автоматического порождения гипотез, схожем с ДСМ-методом: применение статистических соображений // Научно-техническая информация. Сер. 2. – 1996. – № 5–6. – С. 52–55.
23. Григорьев П.А. Sword-системы или ДСМ-системы для цепочек, использующие статистические соображения // Научно-техническая информация. Сер. 2. – 1996. – № 5–6. – С. 45–51.
24. Аншаков О.М. Обобщенные кванторы, определяемые с помощью шаблонов. Часть I // Научно-техническая информация. Сер. 2. – 2000. – № 11. – С. 5–17.
25. Аншаков О.М. Обобщенные кванторы, определяемые с помощью шаблонов. Часть II // Научно-техническая информация. Сер. 2. – 2001. – № 5. – С. 35–48.
26. Гаек П., Гавранек Т. Автоматическое образование гипотез: Математические основы общей теории. – М.: Наука, 1984. – 280 с.

27. Porter M.F. “Snowball: Alanguage for stemming algorithms”. 2001
28. Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. MLMTA – 2003
29. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts: 4th International Conference, (AIST 2015) – Yekaterinburg, Communications in Computer and Information Science, Springer – 2015.
30. Автоматическая Обработка Текста. – URL: https://www.aot.ru (дата обращения 06.02.2015).
31. Salton G., Allan J., Buckley C. Automatic structuring and retrieval of large text files // Communications of the ACM. – 1994 – Vol.37(2)
32. Cavnar W. B., Trenkle J. M.: N-Gram-Based Text Categorization // Proceedings of

SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. – 1994 – P. 161–175.

33. Dunning T.: Statistical Identification of Languages // Comp. Res. Lab. Technical Report, MCCS. – 1994. – P. 94–273.
34. Salton G. Automatic text processing: the transformation, analysis, and retrieval of information by computer. – Boston: Addison-Wesley Longman Publishing, 1989.
35. Yang, Y., & Pedersen, J. O. A comparative study on feature selection in text categorization // Proc. of ICML-97. – 1997. – P. 412-420.
36. Ahonen-Myka H. Finding All Maximal Frequent Sequences in Text // Proceedings of the 16th International Conference of Machine Learning, ICML-99 Workshop on Machine Learning in Text Data Analisys. – 1999. – P.11-17.
37. Rohith K Menon, Yejin Choi. Domain Independent Authorship Attribution without Domain Adaptation // Proceedings of Recent Advances in Natural Language Processing. – Hissar, Bulgaria, 2011. – P. 309–315
38. S. Raghavan A. Kovashka R. Mooney Authorship Attribution Using Probabilistic Context-Free Grammars // In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-2010). – 2010. – P. 38–42.
39. Агеев М.С., Кураленок И.Е. Официальные метрики РОМИП'2004 // Российский семинар по Оценке Методов Информационного Поиска (РОМИП 2004). – Пущино, 2004

Материал поступил в редакцию 12.08.15.

Сведения об авторе

ЛЫФЕНКО Николай Дмитриевич – аспирант отделения интеллектуальных систем в гуманитарной сфере Российского государственного гуманитарного университета, Москва
e-mail: lyfenkonick@ya.ru

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 81’42

Т.И. Ключенко

Методологическая ценность «библиотечной математики» в междисциплинарном дискурсе*

Математика и математические методы в библиотечно-информационной теории и практике являются важнейшими инструментами познания структуры и состава профессиональных систем, выявления закономерностей и механизма их функционирования.

Ключевые слова: *библиотечно-информационная деятельность, математические методы, классификация, закономерности, методология, информационно-коммуникационные технологии, информатизация, мобильные коммуникации*

Во всем мире получает распространение концепция, суть которой в том, что «сердцем» информационного общества должна стать современная библиотека как одна из его инфраструктурных составляющих. Библиотечные специалисты в области информационно-коммуникационных технологий (ИКТ) обеспечивают жизнедеятельность инфраструктуры информационного общества, именно той организации, где информация аккумулируется, изучается, хранится, систематизируется и предоставляется всем. В условиях новой парадигмы информационной среды, когда изменяются не просто информационные потоки, но и сами способы отслеживания появления новой информации, способы ее отбора, хранения, классификации, индексации и представления в доступ, руководитель современной библиотеки должен находить тот уровень информационной обеспеченности и сервиса, который позволит ему сохранить свой читательский контингент и привлечь новый. Поэтому в современных технически оснащенных библиотеках, которые, по сути, все чаще выполняют функции автоматизированных библиотечно-информационных центров, требуются принципиально новые методологические, технические и технологические подходы. Такие подходы могут разрабатывать специалисты, способные интегрировать идеи из различных областей науки, оперировать междисциплинарными категориями, комплексно воспринимать междисциплинарный процесс.

Педагогическая деятельность в области «библиотечной математики» как профессионально ориентированного направления дискретных математических моделей позволяет эффективно использовать математические знания в современных компьютерных технологиях для преподавания библиотечно-информационных дисциплин [1]. Необходимость расширенного толкования образования в области библиотечно-информационной деятельности обусловлена еще и тем, что эта деятельность, по сути, по образному выражению Я.Л. Шрайберга является двигателем общественного прогресса, механизмом обеспечения осмысленного и уверенного превращения нашего общества в общество информационное. Даже на начальных ступенях библиотечно-информационного университетского образования библиотечно-информационные дисциплины продолжают ограничиваться формализованными данными о законах-зависимостях, закономерностях-тенденциях, принципах, заповедях и нормативах, библиотечно-информационных нормах и законодательных отношениях в области авторского и интеллектуального права, необходимых в дальнейшем для практической библиотечно-информационной деятельности.

Однако в теории библиотечно-информационных наук понимание законов природы и сущности процессов информационной деятельности опирается не только на догму теории этой области знания, но и на весь методологический инструментарий, способный воплотить требования библиотечно-информационной деятельности, богатство социального опыта и рациональные достоинства разума в строгих и математически выверенных библиотечно-информационных конструкциях, позволяющих увидеть своеобразную логику «библиотечной математики» [2, 3].

Разработанный нами набор критериев для классификации математических методов [4, 5], ориентирован на необходимость постоянно воспроизводить измерения с учетом специфики объекта познания библиотечно-информационной деятельности, в качестве которого выступает деятельность социального субъекта. Предпринятый нами в свое время анализ областей применения математических методов в библиотечной теории и практике позволил констати-

* Работа выполнена при финансовой поддержке Российского гуманитарного научного фонда, проект 14-03-12004

ровать, что вероятностно-статистические методы занимают значительное место в моделировании библиотечно-информационных процессов и явлений [2, 3]. Это обусловлено, прежде всего тем, что процессы, связанные с общением и управлением в библиотеке, являются стохастическими, отличаются исключительной сложностью и изменчивостью. Они складываются под влиянием многих объективных и субъективных условий. Особенности процессов библиотечно-информационной деятельности вызывают дополнительные трудности их моделирования. Эти трудности заключаются, во-первых, в необходимости построения многократных математических моделей; во-вторых, в учете стохастичности при формировании этих моделей; в-третьих, в наличии многих качественных признаков, которые сложнее поддаются количественному описанию; в-четвертых, в изменчивости процессов библиотечно-информационной деятельности, которая не может улавливаться в пределах одной математической модели.

Как справедливо отмечается в работе [6], изменения во внешней среде происходят постоянно (принимаются новые законы, вводятся новые порядки финансирования, разрабатываются и внедряются новые технологии, трансформируются культурные нормы, ценности и т. д.), при этом возникает необходимость систематического поиска рациональных путей и способов приспособления – адаптации библиотек к меняющимся условиям. Развитие информационных технологий обусловливает появление новых механизмов адаптации, которые, наряду с традиционными, требуют изучения, анализа и систематизации. Адаптация в работе [6] рассматривалась как процесс, состоящий из трех блоков: адаптант (библиотека) – адаптивная связь – внешняя среда. В качестве метода изучения адаптивных связей между этими блоками был выбран корреляционно-регрессионный анализ, позволяющий точно установить степень влияния на основной (базовый) библиотечный процесс таких факторов, как численность пользователей.

Важнейшим методом исследования «алгебры естественной библиотечно-информационной деятельности» является математическое моделирование, с помощью которого осуществляется тестирование фрагментов эталонной деятельности путем сличения ее с соответствующими аспектами математической модели реальной библиотечно-информационной деятельности. Поскольку библиотековедение и библиографоведение не преуспели в полной мере в познании законов библиотечно-информационной деятельности [7], пришлось, основываясь на требованиях к научному закону в философии науки [8], попытаться сформулировать собственное определение закона, необходимое нам для дальнейших рассуждений. В нашей интерпретации под законом в библиотечно-информационной деятельности мы понимаем построение символической и стохастической моделей реальной действительности, изоморфных определенным инвариантным связям и отношениям, объективно имеющим место между объектами, явлениями и процессами в этой деятельности. Накопленный в ходе наших исследований опыт позволяет констатировать, что содержание закона библиотечно-информационной деятельности формулируется сначала как гипотеза, объясняющая наблюдаемые факты. Эта гипотеза доказывается в ходе экспериментальной проверки выводов из нее для определенных критических условий. Чтобы гипотеза была контролируемой, она должна однозначно определять форму, объекты и пределы действия предполагаемой закономерности. Такая определенность достигается, если утверждаемые инварианты характеризуются количественно или структурно. И это уместно, прежде всего, когда познающий субъект сталкивается с новыми явлениями, законы которых не найдены.

Ценность разработанной Д.Ю. Тепловым в 60-е г. XX в. матрицы информационного обмена, исходной основой которой было предположение, что рассеяние публикаций в значительной мере отражает процессы обмена информацией между отраслями, состояла в том, что с ее помощью удалось получить новое знание о явлении рассеяния. Трактовка рассеяния как отдачи информации отраслью позволила ввести в рассмотрение и обратный процесс приема информации в отрасль, названный Д.Ю. Тепловым комплексированием информации [4].

Построенная Д.Ю. Тепловым математическая модель не потеряла своей актуальности и сегодня. Во-первых, как яркий пример подхода к моделированию процессов библиотечно-информационной деятельности на базе теоретических рассуждений с привлечением необходимого статистического материала. Во-вторых, в связи с изменением статуса информации (экономический ресурс, элемент производительных сил, товар) в новых условиях социально-экономического развития общества, когда управление информацией становится важным стимулирующим фактором этого процесса. Внедряемые повсеместно информационно-коммуникационные технологии и стремительно развиваемые информметрия и вебометрия предоставляют широкие возможности для анализа матриц информационного обмена при значительно большем разбросе в повторе конкретных данных с использованием современных средств матричного анализа (высчитать коэффициент, например, прямого и полного информационного обмена, построить динамические модели и т. д.).

Математическая обработка эмпирических данных – важное звено в количественном познании материальных явлений. Но математика – это не только удобный и эффективный инструментарий или средство символической иллюстрации различных конкретных теорий. В силу своего почетного места в системе культуры и классификации наук математика позволяет прийти к выражению таких соотношений, которые нелегко или даже невозможно построить иным путем. Это обусловлено тем, что в союзе с естественными и гуманитарными науками математика разрабатывает общую методологию познания, опираясь на идеи эволюции, системности и самоорганизации.

Проведенное нами в 70-е г. XX в. исследование [5] имело цель рассмотреть документальный поток в его взаимосвязи с информационными потребностями, в его обусловленности социальной и индивидуальной мотиваций на основе выдвинутой гипотезы о возникновении этого потока как реакции на информационный дефицит. В результате математической обработ-

ки полученных данных методом выравнивания с определением численных значений коэффициентов было установлено, что между интенсивностью использования информационного потока и информационным дефицитом существует прямо пропорциональная зависимость. В итоге было подтверждено предположение о возникновении документального потока как реакции на информационный дефицит, а изучение причин, обусловивших особенности функционирования этого потока, позволило установить взаимосвязь и взаимозависимость информационных потребностей и проявлений их особенностей в возникающем потоке литературы.

Наша ссылка на проведенные в XX в. исследования [1, 2, 3] свидетельствует о важности владения навыками применения математических методов при исследовании конкретного явления, зависящего от узкого круга легко реализуемых и воспроизводимых условий. Реализуемость этих условий крайне важна для практического применения, а воспроизводимость есть существенный элемент принятой нами методологии тщательно объяснить область применения выявленной закономерности. Здесь мы приходим к понятию модели.

Дальнейшие наши исследования в этом направлении [2, 3, 9] привели к следующим заключениям. Библиотека как объект моделирования представляет собой пример сложной системы, в математическом описании которой заведомо заключена нелинейность. Исходные параметры библиотеки связаны между собой сложными и заранее неизвестными исследователю зависимостями. Изменение одного из параметров влечет за собой изменение других. Как следствие возникает потребность при создании адекватной математической модели реальной библиотеки прибегать к методике многократного эксперимента, который дает возможность изучать сложные многопараметрические системы. Это обусловлено необходимостью закладывать в математическую модель в качестве исходных данных не просто готовые аналитические зависимости (что как раз и вызывает сомнение относительности сложных нелинейных систем), а зависимости, установленные эмпирически,

Библиотечно-информационное мышление как важнейшая составляющая интеллектуальной культуры, ближайшим образом родственно математическому рассуждению. Рассматривая университетское образование библиотечно-информационных специалистов в таком междисциплинарном дискурсе, можно заключить, что из развития библиотечно-информационной науки нельзя исключать конструктивную составляющую этой науки с точки зрения «математического владения» ею.

Безусловно, есть немало таких библиотечно-информационных проблем, как, например, проблема прав и обязанностей, которые очень сложны и деликатны и потому трудно поддаются обстоятельному анализу, не говоря уже об их математической формализации. Однако непредубежденному профессионалу, библиотечно-информационному специалисту анализ современной библиотечно-информационной практики и проведенное нами многолетнее исследование показывают, что математика все чаще становится действенным инструментом исследования библиотечно-информационных объектов [2, 3, 6]. Как известно, наибольшая методологическая ценность современной математики в развитии познания состоит в том, что на ее абстрактном языке выражается внутренняя организация хорошо формализованных знаний и проводится теоретический анализ в наиболее развитых областях науки.

Опираясь на законы Ранганатана, Брэдфорда, Чубарьяна и др. [10], которые лежат в основе организации и функционирования библиотек, можно определить набор показателей, характеризующих их работу, выбрать оптимальную архитектуру компьютерных систем, разработать технологию их построения, стратегию и практику вовлечения IT-инфраструктуры библиотек в сервисы «cloud computing» (облачные вычисления). Решение этих проблем в библиотечно-информационной деятельности обусловлено широкой информатизацией, развитием информационно-коммуникационных технологий. Появляются новые пользователи библиотечно-информационных ресурсов (представители деловых кругов, информационные посредники, информационные работники, онлайновые пользователи и т. д.), новые авторы (баз полного текста, электронных журналов, гипермедиа и т. д.), изменяются функции, статус библиотек. Все это приводит к возникновению рынка библиотечно-информационных продуктов и услуг, базирующихся на компьютерных библиотечно-информационных системах, на мобильных коммуникациях. Для того чтобы иметь «собственное лицо» на международном и нарождающемся отечественном рынке, необходимо разрабатывать и создавать уникальные «пионерские» и рентабельные автоматизированные библиотечные системы, технологии, рабочие места, продукты и услуги.

В Республике Татарстан с 2008 г. реализуется республиканская целевая программа «Электронный Татарстан». Она предполагает объединение усилий по информатизации различных отраслей жизни региона и создание единых баз данных различных ресурсов. По этой причине выбор Автоматизированной библиотечной информационной системы (АБИС) требовал единого сервера для всех библиотек Республики и программного обеспечения, как для читателей, так и для библиотечных работников. Одно из основных требований, которому должна соответствовать такая система в Татарстане – это возможность обеспечивать поиск книг на татарском языке, который используется жителями Республики наравне с русским. Сегодня единые сервисные мощности расположены в Дата-центре Правительства Республики Татарстан. К зданиям Национальной библиотеки проведены оптические линии связи, все республиканские библиотечные сети получили высокоскоростной доступ к Интернету, и это позволяет пользователям работать с АБИС как облачной структурой. Каждая библиотека имеет на сервере свой раздел, работает с ним через Интернет, и все читатели могут пользоваться любым каталогом любой библиотеки и даже электронными копиями книг, размещенными в Сводном электронном библиотечном каталоге в разделе «Электронные книги»[1] [11, 12].

[1] Создаваемая структура находится в постоянном развитии. В 2010 г. в нее вошли четыре библиотеки Республики Та-

Инновации в информационно-коммуникационных технологиях стимулируют необходимость принимать рациональные экономичные решения – эксплуатировать АБИС на облачной платформе на основе аутсорсинга. Библиотеки сегодня размещают свою информацию и используют социальные сети, построенные на облачных платформах, осваивают блоги, мобильные системы, планшетники и портативные ПК, букридеры, что приводит, естественно, к трансформации всей библиотечно-информационной деятельности [10]. Чтобы соответствовать библиотеке XXI в., нынешнему поколению библиотечных работников в условиях интенсивной трансформации библиотечно-информационной деятельности предстоит решать серьезные задачи: каким образом обслуживать онлайнового и обычного читателя с учетом обеспечения необходимого минимума платных услуг; какой должна быть стратегия и тактика формирования фонда современной библиотеки и как организовать учет фонда онлайновых подписок; каким образом в электронных каталогах и других компонентах информационной навигации следует отражать удаленные ресурсы, особенно при использовании систем открытого доступа; каким должен быть новый подход к статистике, планированию и оценке качества библиотечно-информационного обслуживания; каким образом обеспечить мобильного читателя необходимыми библиотечно-информационными услугами; каким образом определить эффективность (выгоду) участия библиотеки в корпоративных системах (с обязательным пониманием, как получить от этого выгоду). Решение перечисленных многоаспектных задач (список далеко не исчерпывающий) актуализирует методологическую ценность «библиотечной математики» в условиях междисциплинарного дискурса [13].

Идеал методологической ценности «библиотечной математики» будущего – это единство научных исследований и гуманистических ценностей, единство социальных целей математического познания и этических принципов человечества.

СПИСОК ЛИТЕРАТУРЫ

1. Галявиева М.С., Елизаров А.М., Ключенко Т.И. Фундаментальная математическая подготовка, ее роль и значение в обучении гуманитария вуза культуры и искусств // Вестник Казанского государственного университета культуры и искусств. – 2004. – №1. – С. 58-61.
2. Ключенко Т.И. Математизация библиотечного образования. – Казань: Медицина, 2001. – 104 с.
3. Ключенко Т.И. Математика в гуманитарном вузе сегодня. – М.: ВИНИТИ, 2006. –176 с.
4. Теплов Д.Ю. Библиографическая (вторичная) информация по технической литературе в СССР: дис. д-ра. пед. наук. – Л.: Ленингр. гос. ин-т кульутры им. Н.К. Крупской, 1969. – 505 с.
5. Ключенко Т.И. Изучение документальных потоков по естествознанию и технике и проблемы библиографии: сб.науч.тр. – Л.: Ленинг. гос. инст культуры им. Н.К. Крупской,1983. – С.77-85.
6. Макеева О.В. Механизмы адаптации публичных библиотек в условиях меняющихся социокультурных практик населения: автореф. дис. канд. пед. наук. – Новосибирск, 2012. – 25с.
7. Соколов А.В. Законы, закономерности, заповеди в документологии // Вестник Челяб. Гос. Акад. культуры и искусств. – 2010. – №3(23). – С. 20-24.
8. Кондаков Н.И. Логический словарь-справочник. – М., 1957. – 720 с.
9. Бикеев Ш.С., Дрешер Ю.Н., Ключенко Т.И. Прогнозирование тенденций развития отрасли или проблемы как одна из сфер современной библиографической деятельности // Научная библиотека в современном социокультурном контексте: Международ. науч. конф., 12-15 окт. 1993 г.: тез доклада / АН Украины, ЦНБ им. В.И.Вернадского. – Киев, 1993. – С. 219-220.
10. Соколов А.В. Детерминизм и деонтология в документной коммуникационной системе (постановка проблемы) // Вестник Челяб. Гос. Акад. культуры и искусств. – 2008. –№4(16). – С. 6-34.
11. Елизарова Р.У. Информационно-коммуникационные технологии: готовность библиотек к электронному развитию: Научно-практическое пособие. – М.: Литера, 2013. –136 с.
12. Камалетдинов Р.К. Автоматизированные библиотечные информационные системы как средство освоения и внедрения ИКТ: опыт Республики Татарстан // Вестник Казанского государственного университета культуры и искусств. – 2012. – № 4. – С. 16-25.
13. Шрайберг Я.Л. Электронная книга, будущее библиотеки и общественное сознание: попытка осмысления и предвидения // Конференция «Крым», 2013 год. – М.: ГПНТБ России, 2013. – 72 с.

Материал поступил в редакцию 28.04.15.

Сведения об авторе

КЛЮЧЕНКО Тамара Ивановна – доктор педагогических наук, профессор, заведующая кафедрой информатики Казанского государственного института культуры
e-mail: kluchenkoT@rambler.ru

тарстан: Национальная библиотека РТ, Республиканская детская библиотека, ЦБС Зеленодольского района РТ и Айшинская сельская библиотека. Весной 2012 г. в проекте были уже 13 ЦБС с двадцатью филиалами, в 2013 г. – 2017. Работа продолжится до тех пор, пока все библиотеки не вольются в «Единый сводный электронный каталог Республики Татарстан».

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 811.161.1’322.2 ’ 373.42’374

Е.И. Большакова, И.А. Большаков

Аффиксальный критерий паронимии для построения компьютерного словаря паронимов русского языка

Описывается аффиксальный критерий паронимии, предложенный в результате исследования наиболее крупного печатного словаря паронимов русского языка и формализующий понятие паронимии для построения компьютерных словарей. Паронимами считаются пары слов единого корня и одной части речи, у которых различия в аффиксах находятся в строго установленных рамках. На основе критерия построен объемный компьютерный словарь русских паронимов, предназначенный в первую очередь для автоматизированного исправления паронимических ошибок в текстах. Намечены пути уточнения критерия и предложена общая структура компьютерных словарей паронимов.

Ключевые слова: *паронимы, паронимия, компьютерный словарь паронимов, паронимические ошибки, аффиксальный критерий паронимии*

ВВЕДЕНИЕ

Понятие паронимии, связанное с внешним сходством слов, возникло в лингвистике довольно давно. В английском языке слово *paronymous* известно с середины XVII века. Однако если сейчас собрать с интернет-сайтов десяток англо- и франкоязычных определений слов-паронимов, то единства в понимании не обнаружится. В большинстве определений указывается совпадение корня слов (*wise – wisdom*), но не указывается явно их принадлежность к одной части речи. В других определениях фигурирует совпадение звучания при различии смысла и орфографии (*hare – hair*), хотя такие слова принято относить к омофонам. Упоминаются также никак не уточняемые единство происхождения и различие окончаний слов-паронимов. В словаре [1] слово *paronymous* имеет два разных смысла, лишь один из которых имеет отношение к сходству слов.

При такой ситуации становится понятным отсутствие в западноевропейской лексикографии словарей паронимов с описанием различий их значений и указанием соответствующих диагностирующих контекстов. Отдельные сведения о паронимах содержатся лишь в словарях и пособиях по общему словоупотреблению, в частности, в [2] – для английского языка, в [3] – для французского языка.

Русская лексикография в части паронимов выглядит заметно выигрышнее: за последние десятилетия вышло три содержательных словаря русских паронимов [4–6]. В них много диагностирующих контекстов, а словарь [4] указывает смысловые различия паронимов особенно детально. Для смысловых значений типичной паронимической пары характерно то, что они не совпадают (как у синонимов), не противопоставляются (как у антонимов), при этом внешнее сходство слов не означает их полного совпадения (как у омонимов), к примеру: *зрительный – зрительский*.

В предисловиях к этим словарям содержатся уточнения понятия паронимии, причем единым является требование **одинаковости корней и частей речи** слов-паронимов (*абонент – абонемент, задобрить – сдобрить*). В отношении же черт внешнего сходства слов предлагаемые уточнения разнятся. Так, словарь [5] требует от паронимов одинаковое место ударения (например, *сытый – сытный*) и, тем самым, одинаковое число слогов, в пику этому в [4] предлагается, например, пара *показ – показание* с весьма разным числом слогов. Различается и понимание соотношения семантики слов-паронимов и их корней, например, в [6] не признаются полноправными паронимами синонимы типа *специфический – специфичный* и слова с омонимичными корнями типа *ножевой – ножной*.

Таким образом, все три словаря русских паронимов не дают строгого определения паронимии, необходимого как для автоматизации построения более

полных словарей паронимов (наиболее представительный словарь В. Красных [6] включает лишь 2600 паронимов), так и для определенных приложений по автоматической обработке текстов. К таким приложениям относится задача автоматического исправления ошибок в текстах.

Внешнее сходство слов является источником разнообразных ошибок, встречающихся в текстах. Ошибки, при которых одно слово заменяется на другое существующее слово, на него похожее, но отличное по смыслу (например, *word – world*), в западной лингвистике называются малапропизмами [7, 8]. К малопропизмам относятся паронимические ошибки, т.е. неправомерные замены слов на слова с тем же корнем и той же части речи, как, например, при использовании словосочетания *разговорчивый жанр* вместо *разговорный жанр*. Важным приложением словарей паронимов является подбор кандидатов на исправление таких ошибок.

В настоящей работе понятие паронимии уточняется для задачи автоматического исправления паронимических ошибок в русскоязычных текстах. В результате исследования наиболее полного печатного словаря паронимов [6], предложен формальный критерий паронимии, названный **аффиксальным**. Паронимами считаются те пары слов одного корня и одной части речи, у которых различия в аффиксах находятся в фиксированных рамках (раздельно в префиксах и суффиксах). С использованием предложенного критерия автоматически построен компьютерный словарь паронимов, который превышает по объему все известные печатные словари русских паронимов, а также включает все однокоренные паронимы из компьютерного словаря [9]. Исходными для построения словаря данными были группы однокоренных слов русского языка, взятые из системы КроссЛексика [10].

В нашей работе намечаются также пути уточнения аффиксального критерия, позволяющие исключить из числа паронимов внешне малопохожие пары слов (такие как *взнос – переноска*). Обсуждается общая структура компьютерных словарей паронимов.

ФОРМАЛИЗАЦИЯ ПОНЯТИЯ ПАРОНИМИИ

Критерий паронимии должен учитывать факторы разной природы: внешнее, формальное сходство слов, их семантическое различие, а также морфосинтаксические свойства.

Для прикладной задачи автоматического исправления паронимических ошибок нам требуется учитывать не только часть речи, но и другие морфологические свойства слов. Важно, чтобы паронимы словаря (используемого для подбора исправляющих слов) отвечали принципу *морфологической инвариантности контекста*: замена в тексте одного паронима другим без внесения каких-либо иных правок не нарушает морфологическую правильность текста, хотя может изменить его смысл. Принцип инвариантности облегчает исправление паронимических ошибок. Так, если в тексте встретилось словосочетание *безусловный рефлектор,* и тем или иным способом выявлена его ошибочность, то в построенном с учетом указанного принципа словаре паронимов будут сразу найдены всевозможные замены ошибочного слова и среди них *рефлектор → рефлекс*. Этой заменой текст исправляется без какого-либо его редактирования (так как сохраняет морфологическую правильность контекста), другая же замена *рефлектор → рефлексия* потребовала бы пересогласования прилагательного *безусловный* по роду (см. подробнее в [7, 8]). Таким образом, сформулированный выше принцип не допускает паронимию слов *рефлектор* и *рефлексия*.

Для соблюдения принципа инвариантности нами были приняты следующие соглашения, уточняющие часть речи слов-паронимов.

Расщепление существительных по числу и роду: для паронимов по отдельности рассматриваются четыре подгруппы существительных: муж. рода ед. числа, жен. рода ед. числа, сред. рода ед. числа и множ. числа (для множ. числа род считается нерелевантным). Согласно принципу морфологической инвариантности, представители разных подгрупп паронимами быть не могут.

Отделение причастий от глаголов: причастия играют в текстах синтаксическую роль прилагательных и имеют ту же парадигму, поэтому причастия обоих видов и залогов включаются в прилагательные. Однословные степени сравнения прилагательных считаются отдельными прилагательными. Все это допускает паронимические пары в таких группах слов, как {*старый, стареющий, старейший, устаревающий...*}.

Отделение деепричастий от глаголов: русские деепричастия образуют при глаголах примерно те же зависимые обстоятельственные группы, что и наречия (*уйти → торопясь / торопливо*), и поэтому включаются в их группу.

Разделение глаголов и причастий по возвратности: наличие или отсутствие возвратной частицы *ся/сь* делит глаголы и причастия на две несопоставляемые группы, поскольку эта частица существенно меняет модель управления слова, тем самым делая его непохожим на все иные слова той же части речи без частицы.

Расщепление глаголов по виду: два вида русского глагола могут быть существенно разными морфологически, к тому же у них есть различия в сочетаемости, например, можно *делать прыжки*, но нельзя *сделать прыжки*. Поэтому совершенный и несовершенный виды одного глагола относятся нами соответственно к двум разным группам.

Предлагаемый нами формальный критерий паронимии слов русского языка допускает пары слов-паронимов только в рамках одной из рассмотренных выше групп и учитывает их аффиксное сходство при одинаковом корне, например: *отъ**езд** – по**езд**ка* (единый корень, как правило, влечет их смысловую близость). Однако понимание одинаковости корня также требует уточнения.

Корневые морфемы могут иметь несколько алломорфов, которые отличаются буквенно (***дух** – **душ**а; **лиц**о – **лич**ина*). С другой стороны, слова могут иметь омонимичные корни:{***бур**ый, **бур**ный, **бур**овой*}. В нашей задаче были приняты следующие решения.

Учет алломорфизма корня: слова с разными алломорфами одного корня относятся к одной группе однокоренных слов (***отч***ество – ***отеч***ество), и лишь тогда, когда алломорфы корня оказываются слишком далекими по буквенному составу, они формируют разные группы. К примеру, алломорфы *лож/лаг* Vs. *клад/клас* образуют группы {*по**ложить, на**ложить, по**лагать*...} и {***клас**ть, вы**клад**ывать, на**клад**ывать*...}.

Объединение омонимичных корней: в одну группу включаются слова с омонимичными корнями, например: {***бур**ый, **бур**ный, **бур**овой*}. Объединяются в группу и однокоренные слова, омонимичные корни которых имеют хотя бы один одинаковый алломорф {***душ**а, **дух**ота* + ***душ***}, {*за**плак**анный, **плач**ущий* + ***плат**ный, у**плач**енный* + ***плат**яной, **полот**няный*}. Лишь тогда, когда объединенная группа оказывалась слишком обширной, она разбивалась на 2-3 группы с неперекрывающимися алломорфами корня.

Следующее упрощающее решение касалось многокоренных слов, а также слов с префиксоидами типа *много, едино, мульти*... и суффиксоидами типа *летн, этажн*...

Опущение многокоренных слов, слов с префиксоидами и суффиксоидами: не считаются паронимами слова *зловредный, злокачественный, злонамеренный, этажный* и *многоэтажный*, *летний* и *многолетний*, *законный* и *закономерный*. Исключение сделано для слов, имеющих два одинаковых склеенных корня, но разные аффиксы, например: *благотворный* и *благотворительный*.

Кроме перечисленных уточнений наш формальный критерий слов-паронимов учитывает их *аффиксное расстояние*, причем отдельно – в префиксах (приставках) и в суффиксах. Оно оценивается парой целых чисел (*Np*, *Ns*), где *Np* – расстояние в префиксах, т.е число различающихся префиксов, вычисляемое как минимальное количество элементарных операций редактирования [10] цепочки префиксов (их удаление, вставка или замена), переводящих цепочку префиксов одного слова в цепочку префиксов другого слова. Аналогично определяется число *Ns* для суффиксов. Окончания слов не учитываются, поскольку они определяются последним словообразовательным суффиксом или корнем слова. Например: *со-автор – автор-ит-ет*: *Np=1, Ns=2*.

Ограничения на значения (*Np*, *Ns*) для нашего критерия паронимии были установлены по результатам статистического обследования словаря [6], в ходе которого использовалась информация о морфемном разборе слов, взятая из системы КроссЛексика [10].

ИСХОДНЫЕ СЛОВАРНЫЕ ДАННЫЕ И ИХ МОРФОЛОГИЧЕСКАЯ ОБРАБОТКА

Исходной лексической базой для определения пар слов, удовлетворяющих формальному критерию паронимии, были словарные данные системы КроссЛексика – группы слов, имеющих одинаковый корень и относящихся к одной части речи: существительные, глаголы, прилагательные и наречия (последние – при уточненном их понимании, описанном выше). Омонимия слов в группах не учитывается (так как омонимы имеют одинаковую морфемную структуру).

Все слова групп были подвергнуты морфемному разбору вручную, т.е. расчленены на префиксы, корень, суффиксы и окончание. Выделять суффиксы было особенно сложно. В частности, было не ясно, как задавать окончания в инфинитивах; присоединять ли так называемые тематические гласные *а, и, е* к суффиксам причастий *ющ* и *вш*. Было понятно, что склеивание некоторых морфов, различаемых лингвистами, отнесет к паронимам множество не слишком похожих слов, а расщепление морфов сильно отдалит в пространстве аффиксов даже похожие слова. Мы не учитывали аффиксный алломорфизм, и в ряде случаев склеивали смежные суффиксы.

В число префиксов включена частица *не*, пишущаяся слитно, которая вместе с префиксами *а, анти, контра, против* формирует антонимы сравниваемых слов.

Ниже представлены примеры групп для существительных, глаголов и прилагательных после морфемного разбора (префиксам предшествует знак «–», корню «+», суффиксу «–», окончанию «*», перед частицей *сь/ся* также ставится «–»). В любом слове префиксов не более трех, суффиксов – не более шести, а корень, окончание и возвратная частица единственны.

Укажем статистические характеристики указанных групп слов.

Общий объем словарных данных: порядка 2,5 тыс. групп однокоренных слов, охватывающих около 26 тыс. уникальных слов, среди них:

существительных	42,2 %
глаголов	21,9 %
прилагательных	33,7 %
наречий	2,2 %

Пример 1

```
+АВТОР*
+АВТОР-ИЗ-АЦИ*Я
+АВТОР-ИТ-АР-Н-ОСТ*Ь
+АВТОР-ИТ-ЕТ*
+АВТОР-ИТ-ЕТ-Н-ОСТ*Ь
+АВТОР-ИТ-ЕТ*Ы
+АВТОР-СТВ*О
+АВТОР*Ы
-СО+АВТОР*
-СО+АВТОР-СТВ*О
-СО+АВТОР*Ы
```

(A)

```
+БЕД-Н*ЕТЬ
+БЕД-ОВ*АТЬ
+БЕД-СТВ-ОВ*АТЬ
-НА+БЕД-СТВ-ОВ*АТЬ-СЯ
-О+БЕД-Н*ЕТЬ
-О+БЕД-Н*ИТЬ
-О+БЕД-Н*ЯТЬ
-О+БЕД-Н*ЯТЬ-СЯ
-ПО+БЕД-СТВ-ОВ*АТЬ
-ПРИ+БЕД-Н*ИТЬ-СЯ
-ПРИ+БЕД-Н*ЯТЬ-СЯ
```

(B)

```
-НЕ+СИСТЕМ-АТ-ИЗ-ИР-ОВ-АНН*ЫЙ
-НЕ+СИСТЕМ-АТ-ИЧ-ЕСК*ИЙ
+СИСТЕМ-АТ-ИЗ-ИР-ОВ-АНН*ЫЙ
+СИСТЕМ-АТ-ИЗ-ИР-УЮЩ*ИЙ
+СИСТЕМ-АТ-ИЧ-ЕСК*ИЙ
+СИСТЕМ-АТ-ИЧ-Н*ЫЙ
+СИСТЕМ-Н*ЫЙ
```

(C)

Средний объем группы составляет 9.5 слов. Число однокоренных пар около 302 тыс., среди них примерно 150 тыс. уникальных однокоренных пар слов с совпадающими частями речи (в уточненном их понимании) и совпадением наличия/отсутствия частицы *ся/сь* у прилагательных, глаголов и наречий. Наиболее часто встречающиеся префиксы: *за-* , *по-* , *у-* ; суффиксы: *-н-* , *-ся* , *-к-* . Различных суффиксов примерно в 3,5 раза больше, чем префиксов. Максимальное расстояние пар слов в префиксах равно 3 (*вносимый –непроизносимый*) , а в суффиксах – 6 (*линейный – линеаризированный*).

Для построения словаря паронимов специальной программой дополнительно были вычислены необходимые морфологические категории всех слов составленных групп однокоренных слов. Для глаголов, прилагательных и наречий определялась только часть речи, а для существительных – еще число и род (мужской, женский, средний или общий – только для множественного числа).

СТАТИСТИЧЕСКОЕ ОБСЛЕДОВАНИЕ СЛОВАРЯ В. КРАСНЫХ

Аффиксное сходство однокоренных слов изучалось на материале словаря В. И. Красных [6], являющегося наиболее полным из печатных словарей паронимов русского языка и послужившего фактически образцом лингвистического понимания паронимии. Словарь содержит 1100 так называемых паронимических рядов из 2–7 слов (всего 2600 паронимов). Слова каждого ряда относятся к одной части речи (существительные, глаголы или прилагательные), тем самым они неявно считаются равноправными внутри своего ряда, т.е. любое из них паронимично всем остальным.

Все возможные пары слов из одного паронимического ряда обследовались визуально. При сравнении существительных одного ряда не учитывались пары, различные по роду и/или числу, но добавлялись множественные числа для тех существительных, которые таковые имеют. В итоге сравнивалась, например, пара *показ – показание*, в отличие от пары *показы – показания*. В глагольные ряды добавлялись глаголы другого вида, если таковой у них существует. Для всех рассмотренных таким образом пар слов подсчитывались расстояния *Np* и *Ns*, морфемный состав слов брался из уже подготовленных групп однокоренных слов.

Всего в словаре В. Красных было насчитано 3297 пар паронимов, статистика их аффиксного расстояния представлена вторым столбцом таблицы. Наибольшее число пар различаются только одним аффиксом, причем больше пар различается только одним суффиксом. Далее по количеству идут пары, различающиеся двумя аффиксами. Неожиданно большое количество пар слов оказалось на минимальном расстоянии (0, 0), т.е. когда слова имеют одинаковый морфный состав. Сюда попали существительные с алломорфизмом корня (*невежа – невежда*), глаголы и прилагательные с алломорфными корнями или разными окончаниями (*воскресать – воскресить – воскрешать*).

Поскольку набор из семи расстояний: (0, 0), (0, 1), (1, 0), (0, 2), (1, 1), (1, 2), (0, 3), выделенных в таблице полужирным шрифтом, покрывает почти 99% всех рассмотренных пар, он был принят в качестве **аффиксального критерия паронимии**, который можно записать в виде формулы:

$$(Np = 0)\ \&\ (Ns \le 3) \square (Np = 1)\ \&\ (Ns \le 2)$$

со следующей формулировкой: либо префиксы в сравниваемой паре слов одинаковы, а различий в суффиксах не более трех, либо у них один различный префикс, а различий в суффиксах не более двух**.** Как видим, согласно В. Красных, внешняя схожесть слов предпочитает пары слов с одинаковыми началами и допускает больше различий в суффиксах, чем в префиксах.

Таким образом, для построения компьютерного словаря мы применяем формальный критерий: паронимами считаются слова одной части речи и единого корня (в уточненном их понимании), удовлетворяющие указанному выше ограничению на аффиксное расстояние.

Заметим, что представленные в таблице пары, не отвечающие формальному критерию, как правило, взяты из групп однокоренных слов системы КроссЛексика. Обычно они внешне несходны, как пара *запредельность – разделенность*.

ПОСТРОЕНИЕ СЛОВАРЯ ПАРОНИМОВ

Формальный критерий паронимии применялся в ходе программного вычисления пар паронимов исходя из групп однокоренных слов, подвергнутых морфемному разбору и морфологической категоризации,

Каждая группа из *M* слов преобразуется в *M* статей: одно слово исходной группы становится головным для статьи, а *M–1* остальных, подчиненных слов упорядоченно следуют за ним. Вычисляются значения *Np* и *Ns* для всех пар из головного и подчиненного слова. После отсева подчиненных слов, не отвечающих формальному критерию паронимии с головным, все статьи, в которых осталось хотя бы одно подчиненное слово, включаются в словарь паронимов. Для существительных дополнительно учитывается род и число слов, для глаголов – вид и возвратность, и если подчиненное слово отличается от головного по этим параметрам, оно автоматически исключается из статьи.

Ниже – см. Пример 2 – приводим итоговые статьи, сформированные на основе группы (А). Для слова *авторизация* паронимов не нашлось, хотя другие существительные женского рода в группе есть.

Группа (С) однокоренных слов из семи прилагательных переходит в следующие семь статей (см. Пример 3), где число слов, паронимичных головному, колеблется от одного до четырех.

Приведенные примеры показывают, что число паронимов у разных слов из одной группы слов с одинаковым корнем и частью речи может существенно различаться.

Статистика аффиксного расстояния (в процентах)

Np, *Ns*	Словарь Красных	Группы Кросс-Лексики	Примеры пар паронимов из словаря Красных и пар однокоренных слов из системы КроссЛексика
0, 0	4,3	1,3	*невежа–невежда, отечество–отчество, воскресать– воскрешать, осветить–осветлить, отбегать–отбежать, заспавший–засыпавший, временный–временной*
0, 1	31,4	8,0	*абонент–абонемент, доносить–донашивать, маленький–малый, прогулы–прогулки, двигатель–движитель*
1, 0	30,1	39,3	*вход–выход, входить–выходить, индукторный–редукторный, ходить–сходить, выйти–пойти*
0, 2	16.6	6,7	*манера–манерность, автономия–автономность, стрелковый–стреляный, центрировать–централизовать*
1, 1	14,3	25,8	*кондуктор–продукт, проведать–выведывать, означенный–назначаемый*
2, 0	0,1	1,4	*ход–перерасход, извещение–оповещение, означить–переназначить, неискушенный–укушенный*
0, 3	2,8	0,9	*аварийность–авария, актерствовать–активизировать, актовый–активирующий, авторизованный–авторский*
1, 2	0,3	10,9	*болезнь–заболевание, актировать–дезактивировать, агитируемый–сагитированный*
2, 1	0,0	1,3	*запредельность–разделенность, вещание–оповещение, надуманный–понапридумавший*
3, 0	0,0	0,0	*деление–перераспределение, задумывать–понапридумывать, уместный–несовместный*
Проч.	0,4	4,5	*мерзость–омерзительность, политизированность– аполитичность, опубликованный–публицистический, материалистический–материвший*

Пример 2

```
АВТОР
 АВТОРИТЕТ
 СОАВТОР

АВТОРИТАРНОСТЬ
 АВТОРИТЕТНОСТЬ

АВТОРИТЕТ
 АВТОР
 СОАВТОР

АВТОРИТЕТНОСТЬ
 АВТОРИТАРНОСТЬ

АВТОРИТЕТЫ
 АВТОРЫ
 СО+АВТОРЫ

АВТОРСТВО
 СОАВТОРСТВО

АВТОРЫ
 СОАВТОРЫ
 АВТОРИТЕТЫ

СОАВТОР
 АВТОР
 АВТОРИТЕТ

АВТОРСТВО
 СОАВТОРСТВО

СОАВТОРЫ
 АВТОРИТЕТЫ
 АВТОРЫ
```

Пример 3

```
НЕСИСТЕМАТИЗИРОВАННЫЙ
  СИСТЕМАТИЗИРОВАННЫЙ =ANT

НЕСИСТЕМАТИЧЕСКИЙ
  СИСТЕМАТИЧЕСКИЙ =ANT
  СИСТЕМАТИЧНЫЙ ~ANT

СИСТЕМАТИЗИРОВАННЫЙ
  НЕСИСТЕМАТИЗИРОВАННЫЙ =ANT
  СИСТЕМАТИЗИРУЮЩИЙ

СИСТЕМАТИЗИРУЮЩИЙ
  СИСТЕМАТИЗИРОВАННЫЙ
  СИСТЕМАТИЧЕСКИЙ
  СИСТЕМАТИЧНЫЙ

СИСТЕМАТИЧЕСКИЙ
  НЕСИСТЕМАТИЧЕСКИЙ =ANT
  СИСТЕМАТИЗИРУЮЩИЙ
  СИСТЕМАТИЧНЫЙ ~SYN
  СИСТЕМНЫЙ

СИСТЕМАТИЧНЫЙ
  НЕСИСТЕМАТИЧЕСКИЙ ~ANT
  СИСТЕМАТИЗИРУЮЩИЙ
  СИСТЕМАТИЧЕСКИЙ ~SYN
  СИСТЕМНЫЙ

СИСТЕМНЫЙ
  СИСТЕМАТИЧЕСКИЙ
  СИСТЕМАТИЧНЫЙ
```

Подсчитанная в ходе построения словаря статистика аффиксного расстояния всех рассмотренных пар слов представлена в третьем столбце таблицы. Для сопоставимости с данными словаря В. Красных были исключены наречия (их всего 2,2%). Косинус второго и третьего столбцов таблицы, рассматриваемых как числовые векторы, равен 0,79, что дает дополнительное обоснование принятого аффиксального критерия паронимии.

Общее количество уникальных паронимических пар, удовлетворяющих формальному критерию, составило порядка 135 тыс., количество статей (= головных слов) – 22 тыс., среднее число паронимов в статье – 8,8. Статей стало меньше, чем слов в исходных группах однокоренных слов, так как при их формировании произошел отсев слов. Для существительных коэффициент отсева равен 3,73; для глаголов – 1,79; для прилагательных – 1,46. Среднее число паронимов на статью оказалось довольно большим из-за объемных групп глаголов, например, исходная группа из 44 глаголов с корнем *бег/беж* даже после размежевания по возвратности дает набор из 36 невозвратных глаголов-паронимов: *бегать, бежать, вбегать,... сбежать, убегать* и соответственно 36 статей словаря паронимов с 35 паронимами в каждой.

ЭКСПЕРТНАЯ ОЦЕНКА СЛОВАРЯ И ПУТИ УТОЧНЕНИЯ КРИТЕРИЯ

Построенный словарь паронимов существенно превосходит по объему словарь Красных, так что неизбежно возникает вопрос, насколько соответствуют паронимии в ее интуитивном лингвистическом понимании те найденные по формальному критерию пары, которые не входят в этот словарь. В связи с этим была проделана экспертная выборочная проверка таких пар: визуально изучались несколько случайно выбранных фрагментов построенного словаря, охватывающие примерно 650 пар слов. Большое число пар вполне можно считать паронимами (*авторизованный – авторский*), однако, не во всех парах слова внешне похожи (например: *аккредитация – кредитка*), в то же время есть внешне похожие слова, которые обычно к паронимам не относятся (*река – речушка*, *аппетитный – неаппетитный*).

Дополнительно рассматривались пары слов, не попавшие в построенный словарь согласно принятому критерию: среди них есть паронимы из словаря Красных, например, *материалистический – материальный*, а также другие пары, которые можно отнести к паронимам (*извещение – оповещение*).

Тем не менее, мы полагаем предложенный аффиксальный критерий паронимии хорошим первым приближением, и не исключаем дальнейшего его ужесточения, чтобы отсеять внешне мало похожие пары, или смягчения, чтобы допустить пары с одинаковыми конечными суффиксами типа *мерзость – омерзительность*. Для соответствующего уточнения критерия необходимо учитывать дополнительные факторы.

Для определения этих факторов были проведены эксперименты по машинному обучению [12], с использованием метода опорных векторов (SVM). Реализованный машинный классификатор использовал 11 классифицирующих признаков, включающих не только аффиксное расстояние слов (отдельно в префиксах и суффиксах), но и такие формальные признаки, как разность длин слов, редакционное расстояние [11] в буквах между корнями слов и словами в целом, количество совпадающих букв в начале и конце слов и другие. В качестве обучающего множества были взяты 260 пар примеров: положительных (характерные пары паронимов) и отрицательных. В последние были включены внешне непохожие слова, не подходящие под наш критерий (*ходули – перерасходы*), а также подходящие, но которые хотелось бы исключить из паронимов (*абажур – абажурчик*, *понятный – непонятный*).

Построенный на базе машинного классификатора и групп однокоренных слов словарь паронимов включал уже только 80 тыс. пар слов, при этом степень покрытия словаря Красных (т.е. доля построенных паронимов, входящих в словарь В. Красных, от общего числа паронимов этого словаря) была несколько ниже (96%), чем у словаря, построенного по исходному формальному критерию. Повторная выборочная оценка построенных пар слов показала, что, несмотря на значительный отсев пар (по сравнению с построенным словарем из 135 тыс. пар слов), и в целом положительные улучшения (исключены все пары, приведенные выше как отрицательные примеры), были также исключены и пары паронимов с небольшими длинами слов: *век – вечность*, *ледяной – льдистый*. В то же время в словарь были включены некоторые пары, которые едва ли можно считать паронимами (например, *табурет – табуретка*).

Дальнейшие эксперименты с машинным обучением, при которых увеличивалось число обучающих примеров, не привели к существенному улучшению ситуации. Поэтому полагаем основой для уточнения понятия паронимии предложенный аффиксальный критерий, но дополненный как формальными так и неформальными признаками слов, подбираемыми в зависимости от конкретного приложения итогового словаря паронимов (заметим попутно, что исходным ресурсом для его построения могут браться все те же группы однокоренных слов). Для нашей прикладной задачи исправления паронимических ошибок более критична полнота словаря (отбор нужного паронима из предложенного списка делает человек), поэтому вполне допустим и словарь из 135 тыс. слов.

Укажем некоторые пути уточнения критерия паронимии: учет аффиксного алломорфизма и семантики аффиксов (в частности, отрицательных префиксов типа *не-*, *анти-* и уменьшительно-ласкательных суффиксов); ослабление аффиксального критерия для отдельных частей речи (например, для прилагательных включение пар с аффиксным расстоянием (0,4): *активаторный – активированный*); учет различий в сочетаемости слов. Последнее важно для разграничения семантики слов и используется в печатных словарях паронимов в качестве диагностических контекстов употребления паронимов, вместе с их толкованиями. Например, в словаре [6]:

Проездной: Дающий право на проезд каким-либо видом транспорта. *Проездной билет, Проездные документы*.

Проезжий: Предназначенный или годный для проезда. *Проезжая дорога. Проезжая часть улицы*.

Согласно положениям дистрибутивной семантики [13] значения слов близки, если они употребляются в схожих контекстах, что может быть использовано для выявления синонимов и антонимов среди обнаруженных по критерию паронимии пар слов. С этой целью был проведен эксперимент, в котором участвовало 500 пар прилагательных (как самых многочисленных представителей паронимов) и сочетающиеся с ними существительные (взяты из системы, описанной в [10]). Для каждой пары прилагательных вычислялся коэффициент K их смысловой близости, зависящий от числа сочетающихся слов с первым словом (N_1), со вторым словом(N_2) и одновременно с ними обоими (N_{12}):

$$K = \frac{N_{12}}{\sqrt{N_1 \times N_2}} \quad .$$

Вычисляемый таким образом коэффициент равен нулю, если множества сочетающихся слов не пересекаются, и равен 1, если множества полностью совпадают. Согласно нашим экспериментальным оценкам, значения коэффициента близости, превышающие 0,3, сигнализируют о синонимии или антонимии исследуемых прилагательных (так, K=0,47 для *эклектический – эклектичный*, и K=0.57 для *правильный – неправильный*).

СТРУКТУРА И СОСТАВ СЛОВАРЕЙ ПАРОНИМОВ

Приведенные примеры статей построенного компьютерного словаря показывают, что число паронимов у разных слов из одной группы слов с одинаковым корнем и частью речи может существенно различаться. Это означает, что понятие паронимического ряда (используемое, в частности, в словаре Красных) как совокупности равноправных взаимно паронимичных слов несостоятельно. Если слово *X* = *несистематизированный* паронимично слову *Y* = *систематизированный*, а слово *Y* паронимично слову *Z* = *систематизирующий*, то это не значит, что *X* паронимично *Z*, т.е паронимия не является транзитивным отношением слов. Однако возможны цепи слов произвольной длины, где любые два смежных слова паронимичны, например, *несистематизированный – систематизированный – систематизирующий – систематический – систематичный – системный.*

При отказе от понятия паронимического ряда в качестве базового следует взять понятие паронимической пары, и словарь паронимов можно строить двумя способами.

1. Список пар, упорядоченных по первому слову. Там, где у первого слова есть несколько паронимов, возникающие пары упорядочиваются по второму слову. Чтобы исключить внутренние отсылки, каждая пара включается в такой словарь дважды, с разным порядком следования слов в паре. Примерно так строятся словари антонимов – см., например, [14].

2. Упорядоченный список словарных статей со структурой «головное слово – упорядоченный список его паронимов». Именно этот способ реализован в нашем компьютерном словаре, и он представляется более предпочтительным. Действительно, в каждой статье вместе со списком паронимов удобно расположить иные сведения о головном слове, например, для прилагательного:

- Толкование и перевод на английский язык;
- Список паронимов;
- Сведения о сочетаемости (=диагностические контексты) по разделам:
  - Наречия, определяющие головное слово;
  - Определяемые им существительные;
  - Какими глаголами управляется головное слово;
  - Какими прилагательными/причастиями оно управляется;
  - Какими наречиями/деепричастиями оно управляется;
  - Модель управления (включая функцию именного сказуемого);
  - Частотные сочиненные пары, в которые входит головное слово.

В ограниченном объеме эти сведения даются в печатных словарях [4, 6] при сравнении слов-паронимов, более полно эти связи собраны в компьютерном словаре [10].

Отметим, что применяемый нами формальный критерий паронимии допускает, что паронимы могут быть синонимами (*патетический – патетичный*) или антонимами (*типичный – атипичный*), это не исключается и в словаре Красных. При этом возможны:

- синонимы абсолютные (могут заменять сравниваемое слово в любом контексте) и относительные (годны для замен в некоторых контекстах);
- антонимы абсолютные (взяты из словарей антонимов) и относительные (синонимы абсолютных антонимов и антонимы их синонимов).

В последнем приведенном примере статей построенного словаря паронимов показаны статьи, автоматически размеченные (средствами системы КроссЛексика) символами синонимии и антонимии.

Если не считать абсолютных синонимов, совокупности диагностирующих контекстов у таких пар слов чаще всего различны, и поэтому разумно хранить всех их в словаре паронимов. Это касается и абсолютных синонимов, имеющих разные паронимы. И только тогда, когда абсолютные синонимы образуют изолированную пару типа *апельсиновый – апельсинный*, их в паронимический словарь включать нецелесообразно.

ЗАКЛЮЧЕНИЕ

Проведенное статистическое исследование наиболее крупного печатного словаря русских паронимов позволило выявить нам параметры аффиксного сходства (в префиксах и суффиксах по отдельности) слов-паронимов и предложить формальный аффиксальный критерий паронимии. Паронимами считаются пары слов с одинаковым корнем и одной части речи, у которых различия в префиксах и суффиксах находятся в фиксированных рамках. Согласно этому критерию построен компьютерный словарь русских паронимов, по объему превышающий все известные печатные

словари. Основное приложение этого словаря – подбор слов-замен для автоматизированного исправления паронимических ошибок в текстах. Программные эксперименты по анализу состава построенного словаря показали пути дальнейшего уточнения аффиксального критерия, что может потребоваться для других приложений. Предложена новая структура словарей паронимов, опирающаяся на понятие паронимической пары и равно подходящая для компьютерных и печатных словарей.

СПИСОК ЛИТЕРАТУРЫ

1. Merriam Webster's Collegiate Dictionary. – Merriam-Webster Inc., 1993.
2. Fowler H.W. Dictionary of Modern English Usage. – Wordsworth Editions Ltd, 1994.
3. Péchoin D., Dauphin B. Dictionnaire des difficultés du français d'aujourd'hui. – Larousse, 2001.
4. Бельчиков Ю.А., Панюшева М.С. Словарь паронимов современного русского языка. – М.: Русский язык, 1994. – 455 с.
5. Вишнякова О.В. Словарь паронимов русского языка. – М.: Русский язык, 1984. – 352 с.
6. Красных В.И. Толковый словарь паронимов русского языка. – М.: Астрель, АСТ, 2007. – 589 с.
7. Bolshakov I.A., Gelbukh A. On Detection of Malapropisms by Multistage Collocation Testing. // Proc. 8th Intern. Conference on Applications of Natural Language to Information Systems NLDB´2003, June 2003, Burg, Germany / eds.A. Düsterhöft, B. Talheim. – GI-Edition, LNI V. P-29, Bonn, 2003. – P. 28-41.
8. Большакова Е.И., Большаков И.А. Автоматическое обнаружение и автоматизированное исправление русских малапропизмов // Научно-техническая информация. Сер. 2. – 2007. – № 5. – С. 27-40.
9. Гусев В.Д., Саломатина Н.В. Электронный словарь паронимов: версия 2 // Научно-техническая информация. Сер. 2. – 2001. – № 7. – С. 26-33.
10. Большаков И.А. КроссЛексика: универсум связей между русскими словами // Бизнес-информатика. – 2013.– № 3 (25) – С. 12-19.
11. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклад АН СССР – 1965. – Т. 163, № 4. – С. 845-848.
12. Броварь И.В. Методы построения компьютерного словаря морфемных паронимов: дипломная работа. – М.: МГУ им. М.В.Ломоносова, ВМК. – 51 с.
13. Лингвистический энциклопедический словарь / под ред. В. Н. Ярцевой – М.: Советская энциклопедия, 1990. – 685 с.
14. Михайлова О.А. Словарь антонимов русского языка. – М.: Эксмо, 2007.

Материал поступил в редакцию 15.09.15.

Сведения об авторах

БОЛЬШАКОВА Елена Игоревна – кандидат физико-математических наук, доцент факультета вычислительной математики и кибернетики МГУ им. М.В. Ломоносова
e-mail: eibolshakova@gmail.com

БОЛЬШАКОВ Игорь Алексеевич – доктор технических наук, профессор, независимый исследователь
e-mail: iabolshakov@gmail.com

УДК 811.161.1’37

Р.И. Розина

Семантические переходы в разговорной речи*

Рассматриваются семантические переходы, результатом которых является появление разговорных и сленговых значений русских глаголов. Проводится сопоставление с семантическими переходами, лежащими в основе разговорных и сленговых значений английских глаголов. Различия между семантическими переходами, характерными для русской и английской разговорной речи, связываются с межкультурными различиями.

Ключевые слова: *разговорная речь, сленг, семантический переход, метафора, глагол, русский язык, английский язык, язык и культура, язык и общество*

Рассмотрим несколько регулярных семантических переходов как от литературного значения к разговорному, так и в рамках разговорной речи. Термин «семантический переход» я использую вслед за Анной А. Зализняк как более широкий, чем термин «полисемия», потому что понятие семантического перехода позволяет включить в область рассмотрения не только два сосуществующих значения одного слова, но и значения, которые слово имело и теряло в ходе семантической эволюции, диалектные значения и значения слов, связанных отношениями морфологической деривации [1, с. 35].

Не всегда удается доказать, есть или нет семантический переход между значениями. Яркий пример – слово *квасить*. Разговорное значение *квасить* ‘пить много и долго’ никакими компонентами не связано с литературным значением *квасить* ‘готовить что-л. к употреблению, вызывая брожение, делая кислым’, например, *квасить капусту*[1]. Поэтому можно предположить что между этими значениями нет отношения деривации и что имеет место не семантический переход, а омонимия двух слов: *квасить* (литературного) и *квасить* (разговорного). Между тем есть немецкий глагол quassen ‘много есть и пить’, который дал в английском глагол *quaff*[2] ‘много есть и пить’ и, возможно, был заимствован и русским языком.

Еще один пример – слово *жесть*, которое может выражать как положительную, так и отрицательную оценку, ср.:

А у нас вообще, Вера Ванна, во второй школе просто ***жесть*** *была женщина* (Разговор о школе // Из коллекции Ульяновского университета).

У меня, кстати, в этом семестре экзамен еще будет по истории… ***жесть***. (Разговоры ульяновских студентов // Из коллекции Ульяновского университета).

Непонятно, каким образом на основе значения ‘очень тонкий стальной лист’ могло возникнуть значение оценки. Возможно, что в этом значении слово *жесть* – усечение слова *жесткач* ‘что-л. очень интенсивное или страшное, ужасное, неприятное’ (*Драка была –* ***жесткач****, я думала, такое только по телевизору показывают*)[3].

Рассмотрим некоторые встречающиеся в разговорной речи регулярные семантические переходы.

Несколько разговорных существительных реализуют семантический переход ‘обувь, характерная для какой-л. социальной группы, – глупый, простодушный или тупой человек’: *валенок*, *лапоть*, *сапог*. Понятно, что здесь пропущено звено ‘обувь, характерная для какой-л. социальной группы, – человек, представитель этой социальной группы – характерное свойство такого человека’. Пропущено звено и в семантическом переходе, который реализуют такие существительные, как *деревня* и *тундра*: ‘отдаленное или глухое место – неотёсанный человек’. Его же демонстрирует суффиксальное производное *деревенщина*: *деревня* ‘глухое место’ – *деревенщина* ‘неотесанный, необразованный человек’. Полная форма этого перехода – ‘отдаленное или глухое место – человек, живущий в этом месте, – неотёсанный человек’. Этот же переход, но с отсутствующим первым

* Работа выполнена в рамках проекта, поддержанного Российским гуманитарным научным фондом (грант № 14-04-00604а).

[1] Глагол *квасить* образован от названия напитка *квас*, которое возводится к лат. cāseus ‘сыр’ из *kvātjo; др.-инд. kváthati ‘кипятить, варить;’ kvāthas ‘отвар’; гот. hraÞō ‘пена’ [2].

[2] Предполагается, что двойное f – следствие неправильной передачи на письме двойного s, что вполне реально, учитывая особенности готического шрифта [3].

[3] Мысль о том, что слово *жесть* может быть связано со словом *жесткач*, была высказана в устном выступлении А.Д. Шмелевым. Еще одно предположение – о том, что слово *жесть* ведёт своё происхождение из жаргона музыкантов, исполняющих тяжёлый рок, – было высказано М.Я. Дымарским в личном общении: в этом жаргоне слово *жестокий*, а затем *жёсткий* использовалось для выражения положительной оценки.

звеном, просматривается в таком существительном, как *чалдон* ‘житель Сибири – глупый человек’[4].

С точки зрения регулярных семантических переходов интерес представляют прилагательные, описывающие превышающий норму размер органов и частей тела человека. Для этих прилагательных характерна полисемия ‘имеющий орган или часть тела больше нормы – имеющий способность, связанную с хорошим функционированием этого органа или части тела’. Примеры: *глазастый* ‘способный замечать все детали происходящего вокруг себя’, *головастый* ‘умный, сообразительный’, *рукастый* ‘умелый’. Однако эта связь нерегулярна: такие прилагательные, как *лопоухий, ушастый* и *носатый* не имеют значения ‘обладающий хороший слухом’ или ‘обладающий хорошим обонянием’. Возможно, причина в том, что запах и слух воспринимаются не внешней, видимой частью этих органов, а их внутренними поверхностями. Слово *лобастый* не значит ‘умный’: *лобастым* может быть щенок, ребенок, взрослый человек, собака и волк – и контексты, в которых употребляется это слово, могут быть как положительными, так и отрицательными.

У некоторых прилагательных этого ряда физические значения отошли на задний план, а на первое место выступили функциональные значения: слово *рукастый* малоупотребительно в своем основном значении ‘с длинными большими руками’; *зубастый* ‘способный зло и остроумно ответить’ в своем основном значении ‘с острыми сильными зубами’ не употребляется по отношению к человеку – зубастым может быть только животное (*зубастая щука, зубастый волк*); слово *языкатый* ‘острый на язык, способный хорошо, метко ответить’ вообще не имеет физического значения ‘с большим языком’ – скорее всего, потому, что язык не виден и не может использоваться для характеристики внешности человека.

Очень большое количество семантических переходов реализуют глаголы звучания. Один из широко известных случаев семантического перехода – соотношение ‘звук – движение, которое сопровождается этим звуком’, например, *тарахтеть* (*мотор* ***тарахтит*** *– по шоссе* ***тарахтела*** *косилка*), *свистеть* (*ветер* ***свистит*** *– пули* ***свистят*** *по степи*); *жужжать* (*шмель* ***жужжит*** *– мимо* ***прожужжал*** *шмель*). В разговорной речи используется близкий к этому переход ‘издать громкий короткий звук – совершить действие, которое сопровождается этим звуком’, в частности ‘ударить’: *ахнуть* (***ахнуть*** *от восторга –* ***ахнуть*** *кого-л. мордой об стол*); *бабахнуть* (*пушки* ***бабахнули*** *–* ***бабахнуть*** *кого-л. кулаком по спине*); *бухнуть* (*пушка* ***бухнула*** *–* ***бухнуть*** *кого-л. кулаком в грудь*).

Эти же и некоторые другие глаголы звучания реализуют другой семантический переход: ‘издать громкий короткий звук – бросить что-л. с шумом или упасть’ (*статуэтка с полки* ***бабахнулась***; ***брякнула*** *цепь у калитки –* ***брякнул*** *связку ключей на стол; я щас в обморок* ***брякнусь***; ***бухнула*** *сумку на пол –* ***бухнулась*** *на диван*). Механизм этого перехода понятен: центральный и единственный компонент основного значения в производном вытесняется на периферию, приобретая статус следствия действия. Это звено связано еще с двумя – ‘сказать что-л. неуместное в разговоре’: *Дай только клятву, что ты не* ***бабахнешь*** *всё это на общем собрании; Смотри, не* ***брякни*** *что-нибудь такое в разговоре; Андрей Кураев умудрился* ***бухнуть****… в этом своём интервью две замечательные совершенно вещи* (из коллекции НКРЯ) и ‘положить (использовать) слишком большое количество чего-л. во что-л. (в чем-л.)’: ***бухнула*** *слишком много соли в суп*; ***бабахнула*** *двойную дозу валокордина*; ***грохнула/ухнула*** *все деньги на шмотки*.

В отношении семантических переходов возникает вопрос, свойственны ли они только разговорной речи и характерны ли они только для русского языка. В этой статье мы попытаемся ответить на второй вопрос.

Сходный с только что рассмотренным семантический переход реализуется в английском языке во фразеологизме *to drop a brick* ‘сказать что-л. неуместное’, букв. ‘уронить кирпич’, и *to drop a clanger* с тем же значением (*clanger* – нечто тяжелое, сделанное из металла, например, колокол от *clang* ‘звук, который издает тяжелый металлический предмет’).

Другой регулярный семантический переход ‘сделать грязным – испортить репутацию, опозорить’ реализуют глаголы загрязнения: *грязнить*, *замазать, замазаться, замарать, замараться, запачкать, заплевать*. Ср.: *Посмотри, ты рукав* ***замазала*** *– Уже после самоубийства Серго Сталин решил меня* ***замазать*** *участием в репрессиях.*

Этот переход встречается и в других языках, ср. укр. *забруднити* ‘испачкать’ и ‘лишить моральной чистоты, опозорить’; *замазати*, *заплямувати* ‘запятнать’ и ‘опорочить’; *заплювати*, *запакостити*, *зчорнити; а также* англ. *besmirch, besmear, foul, smear, soil, stain, spatter, sully, defile.* Из этого не следует, что данный переход характерен только для разговорной речи, ср. англ. глаголы *besmirch, besmear,* имеющие книжный характер.

Рассмотрим теперь семантические переходы, которые реализуются в одном из самых многочисленных тематических классов разговорной речи – в глаголах выпивки.

Глаголы выпивки делятся на две группы – глаголы, описывающие разовое действие 'выпить одним глотком' (*ахнуть, бухну́ть, вмазать, врезать, дербалызнуть, дерябнуть, дёрнуть, долбануть, жахнуть, загрузиться, заложить, замахнуть, зашибить, кирнуть, клюкнуть, махнуть, накатить, опрокинуть, поддать, принять, раздавить, тяпнуть, хватить, хлобыстнуть, хлопнуть, хряпнуть…*), и глаголы, описывающие длительный процесс выпивания или хроническое пьянство (*бухать, глушить, гудеть, жрать, закладывать (за воротник), заливать, зашибать, квасить, керосинить, кирять, лакать, стаканить, хлебать, хлестать…*).

Многие глаголы первой группы реализуют семантический переход ‘ударить – выпить’. Это такие глаголы, как *вмазать, врезать, долбануть, жахнуть, тяпнуть* (неясно, от какого именно значения:

[4] В своём исходном значении *чалдон* – название первых русских поселенцев в Сибири, потомков ссыльных, каторжан и разбойников, которым в фольклоре приписывалась глупость и лень.

‘резко ударить острым предметом – тяпкой или топором’ – или ‘укусить’), *хватить*. А два других глагола, *дёрнуть и опрокинуть,* реализуют другой тип перехода, ‘сделать резкое движение – выпить’. Неясным остается, чем мотивировано значение ‘выпить’ у таких глаголов, как *дербалызнуть*, *дерябнуть*, *хлобыстнуть*, *хряпнуть* и *клюкнуть*. Между тем оказывается, что *дерябнуть* имеет диалектное происхождение: в целом ряде диалектов, в частности, в вологодском, владимирском, рязанском, этот глагол имеет значение ‘сильно ударить’, ‘хватить’ [2, 4]. С глаголом *дерябнуть* связан глагол *дербалызнуть*, значение ‘ударить’ у которого указано в словаре Ушакова. *Хлобыстнуть* – глагол, который в ряде диалектов (калужском, тамбовском, рязанском) значит ‘хлестнуть, ушибить, ударить’. Что касается глагола *клюкнуть*, то Фасмер указывает на его связь с болгарским *клюкам* «стучу, толкаю», т.е. здесь опять присутствует идея удара. О глаголе *клюкать* В.В. Виноградов писал, что это областное народное слово, обозначавшее состояние опьянения и широко употреблявшееся в южно-великорусских говорах (курском, орловском, тульском). При этом оно вытеснило старое *куликать*, которым пользовались в обиходном языке XIII в. [5, с. 919–920). Это значение сформировалось у диалектного глагола *клюкать* с исходным значением ‘*клевать, бить, постукивать клювом’*, в севернорусских говорах – ‘*слегка ударять по чему-нибудь топором*’**.**

В.В. Виноградов указывает на существование еще одного глагола, имеющего значение ‘выпить’, который сейчас не употребляется, – это глагол *клюнуть,* ср.:

И что ***не клюнувши*** *и чарки и другой,*

В суд ни один из них не ступит и ногой.

Таким образом, все эти глаголы (*дерябнуть*, *дербалызнуть* и *клюкнуть*, так же как вышедшее из употребления *клюнуть*) реализуют семантический переход ‘ударить – выпить’.

Что касается глагола *замахнуть*, то Даль приводит пример *замахнуть лапоть*, *закинуть*. Таким образом, *закинуть* реализует тот же семантический переход, что *дёрнуть* и *опрокинуть*: ‘резким движением переместить объект – выпить’.

Посмотрим теперь, насколько эта группа глаголов выпивки характерна именно для русского языка.

В английском языке не так много глаголов выпивки[5]. Это *bevvy, booze, tipple, tope, indulge* и *swig*. Из них *bevvy* – глагол, образованный усечением в сочетании с суффиксацией от существительного *beverage* ‘напиток’. Глаголы *booze, swig* ‘пить алкоголь быстро и большими глотками’ образованы по конверсии от существительных *booze* и *swig* ‘алкогольный напиток’. Глаголы *tipple* и *tope* ‘много пить’– непроизводные, возможно заимствования.

Разговорных и сленговых обозначений алкогольных напитков в английском языке тоже совсем немного по сравнению с русским: *juice*, букв. ‘сок’; *booze,* букв. ‘выпивка, спиртное’; *bottle*, букв. ‘бутылка’; *the hard stuff*, букв. ‘крепкая субстанция’; *Dutch courage,* букв. ‘голландская храбрость’; *hooch* (*hootch*), ‘алкогольный напиток, обычно виски низкого качества или полученное незаконным путём’[6]; *lush* (сленг), ‘алкоголь’; *falling-down juice* (сленг), букв. ‘сок, от которого падают’; *nip,* ‘небольшое количество алкоголя’.

Несколько лексем используется для обозначения алкоголика: это *soak* (от глагола *to soak* ‘пропитать жидкостью’), *lush* (сленг) от существительного *lush* ‘спиртное’ и *wino* от *wine* ‘вино’.

Совсем другая картина возникает, если обратиться к прилагательным. Их очень много: *intoxicated, букв.* ‘находящийся в состоянии интоксикации’[7]; *loaded* (сленг, преимущественно канадский и американский), букв. ‘загруженный’; *tight* ‘навеселе, под мухой, на взводе’, букв. ‘натянутый’; *canned* (сленг), букв. ‘законсервированный’; *flying* (сленг), букв. ‘летящий’; *bombed* (сленг), букв. ‘разбомбленный’; *stoned* (сленг), букв. ‘побитый камнями’; *wasted* (сленг), букв. ‘растраченный понапрасну’; *smashed* (сленг), букв. ‘разбитый’; *hammered* (сленг), букв. ‘убитый, побитый молотком’; *steaming* (сленг), букв. ‘дымящийся, испускающий пар’; *wrecked* (сленг), букв. ‘разбитый, потерпевший аварию’; *soaked*, букв. ‘промокший, пропитанный’; *out of it* (сленг), букв. ‘вне’; *plastered*, букв. ‘оштукатуренный’; *blitzed* (сленг), букв. ‘стукнутый’; *boozed-up* (сленг), букв. ‘наспиртованный’; *lit up* (сленг), букв. ‘освещенный’; *stewed* (сленг), букв. ‘тушёный’; *pickled*, букв. ‘маринованный’*; bladdered* (сленг) от bladder ‘мочевой пузырь’; *under the influence*, букв. ‘под воздействием’; *sloshed* (сленг), букв. ‘забрызганный’; *tipsy* ‘навеселе, подвыпивший’; *maudlin*, букв. ‘слезливый, плаксивый’; *well-oiled* (сленг), букв. ‘хорошо смазанный’; *legless* ‘безногий’; *paralytic* ‘парализованный’; *mullered* (сленг), букв. ‘размазанный’; *trashed* (сленг), букв. ‘никчемный, такой, которому место на свалке’; *tiddly* (сленг), букв. ‘нетвердо держащийся на ногах’; *zonked* (сленг), букв. ‘ошалевший, одурманенный’; *blotto* (сленг) ‘вдребезги пьяный, одурманенный’ от *blot* ‘ставить пятна, грязнить’ или от *blotter* ‘промокашка’; *fuddled* ‘подвыпивший’; *tanked up* (сленг) ‘накачавшийся пивом’, букв. ‘наполненный до краев’; *pie-eyed* (сленг), букв. ‘с глазами, как пирожки’.

Итак, оказывается, что хотя слова, связанные с выпивкой, занимают значительное место как в русской, так и в английской разговорной речи, в центр внимания в каждом из этих языков попадают разные её аспекты.

В русской разговорной речи имеется явный «крен» в сторону обозначения спиртных напитков и их поглощения, в то время как в английской – в сторону описания состояния опьянения.

[5] Списки слов составлялись на основе тезауруса [6].

[6] Усечение от Hutchinoo, названия индейского племени, делавшего напиток.

[7] Подобрать этим прилагательным русские эквиваленты по большей части невозможно. Все они в своих переносных значениях описывают состояние опьянения разной степени (в основном сильного), а различаются своими буквальными значениями, которые мы и приводим. Для производных слов в некоторых случаях, когда невозможно найти перевод, мы указываем мотивирующее слово.

СПИСОК ЛИТЕРАТУРЫ

1. Зализняк Анна А. Семантический переход как объект типологии // Вопросы языкознания. – 2013. – № 2. – С. 32–51.
2. Фасмер М. Этимологический словарь русского языка. – М.: Прогресс, 1986. – URL: http://www.etymojnline.com.
3. Online Etymological Dictionary. – URL: http://www.etymonline.com.
4. Даль В.И. Толковый словарь живого великорусского языка. – М.: Русский язык, 2002.
5. Виноградов В.В. История слов. – М.: Изд-во ИРЯ, 1999.
6. Collins English Thesaurus. – URL: http://www.collinsdictionary.com/english-thesaurus.

Материал поступил в редакцию 22.08.15.

Сведения об авторе

РОЗИНА Раиса Иосифовна – доктор филологических наук, ведущий научный сотрудник Института русского языка им. В.В. Виноградова РАН
e-mail: rarozina@yandex.ru

ВНИМАНИЮ ЧИТАТЕЛЕЙ!

С 2000 года ВИНИТИ РАН вошел в состав Управляющего совета Консорциума Универсальной десятичной классификации (УДК). Институт в качестве единственного в России владельца лицензии на распространение печатных и электронных (на CD-ROM) изданий УДК на русском языке возобновил полное издание таблиц УДК.

ВИНИТИ РАН предлагает издания:

1. Таблицы УДК

УДК. Том I Общая методика применения УДК. Вспомогательные таблицы. Основные таблицы. Общий отдел. Алфавитно-предметный указатель к Общему отделу (только электронное издание)

УДК. Том II **1/3** Философия. Психология. Религия. Богословие. Общественные науки (только электронное издание)

УДК. Том III **5/54** Математика. Естественные науки (только электронное издание)

УДК. Том IV **55/59** Геологические и биологические науки

УДК. Том V **6/61** Медицинские науки (только электронное издание)

УДК. Том VI (часть 1) 6/621 Прикладные науки. Технология. Инженерное дело (только электронное издание)

УДК. Том VI (часть 2) 622/629 Техника. Инженерное дело (только электронное издание)

УДК. Алфавитно-предметный указатель к т. VI (1 и 2 части) (только электронное издание)

УДК. Том VII **63/65** Сельское хозяйство. Домоводство. Управление предприятием

УДК. Том VIII **66** Химическая технология. Химическая промышленность. Пищевая промышленность. Металлургия. Родственные отрасли

УДК. Том IX **67/69** Различные отрасли промышленности и ремесел. Строительство

УДК. Том X **7/9** Искусство. Спорт. Филология. География. История.

УДК. Изменения и дополнения. Выпуск 2 (к т.т. 1-3) (только электронное издание)

УДК. Изменения и дополнения. Выпуск 3 (к т.т. 1-6) (только электронное издание)

УДК. Изменения и дополнения. Выпуск 4 (к т.т. 1-7)

УДК. Изменения и дополнения. Выпуск 5 (к т.т. 1-10)

2. Государственный рубрикатор научной и технической информации (ГРНТИ) в 2-х томах, издание шестое, 2007.

Для подписки необходимо направить заявку для оформления счета по адресу:

125190, Россия, Москва, ул. Усиевича, 20, НМО ВИНИТИ

Телефон: 8-499-155-42-52

Факс: 8-499-943-00-60 (для НМО)

E-mail: typo@viniti.ru