УДК 004.045

ПОВЫШЕНИЕ ТОЧНОСТИ КЛАССИФИКАЦИИ НАУЧНЫХ ДАННЫХ ПРИ ИСПОЛЬЗОВАНИИ АНСАМБЛЕВОГО ПОДХОДА

Гусева А.И., Киреев В.С., Филиппов С.А.

Национальный исследовательский ядерный университет «МИФИ», Москва, e-mail: aiguseva@mephi.ru, vskireev@mephi.ru, Stanislav@Philippov.ru

Настоящая статья посвящена перспективам использования таких методов интеллектуальной обработки слабоструктурированных данных, как построение ансамблей алгоритмов для решения задач кластеризации и классификации, в научных рекомендательных и аналитических системах, в системах Business Intelligence, системах для нахождения контента в сети Интернет и оценивается современное состояние рынка таких систем. В статье рассматривается новый метод повышения пертинентности информации в научных рекомендательных системах, научных информационных и аналитических системах, содержащих сведения о научных публикациях. В качестве базы исследования были взяты данные о публикациях по двум российским научным публичным электронным ресурсам – elibrary.ru и cyberleninka.ru. Полученные данные (сведения о более чем 500 тыс. статей) специальным образом были предобработаны для выделения основных слов и терминов. Эффективность предложенных алгоритмических ансамблей достаточно высока и достигает в наилучшем случае 88–90%точности.

Ключевые слова: методы интеллектуальной обработки слабоструктурированных данных, ансамбль алгоритмов, задача многомерной кластеризации, задача многомерной классификации, рекомендательные научные системы, научные аналитические системы, системы BusinessIntelligence, высокопертинентные алгоритмы

IMPROVING THE CLASSIFICATION ACCURACY OF THE SCIENTIFIC DATA WHEN USING ENSEMBLE APPROACH

Guseva A.I., Kireev V.S., Filippov S.A.

National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, e-mail: aiguseva@mephi.ru, vskireev@mephi.ru, Stanislav@Philippov.ru

This article is devoted to prospects of the use of ensembles of algorithms for solving problems of multidimensional clustering and multidimensional classification in scientific Advisory and analytical systems, Business intelligence systems, systems for finding content on the Internet and evaluates the current state of the market of such systems. The article considers a new method of increasing the pertinence of information in the scientific Recommender systems scientific information and analytical systems that contain information about scientific publications. As the base study data were collected on two publications of the Russian scientific public electronic resources elibrary.ru and cyberleninka.ru. Derived data (data about more than 500 thousand articles) was treated to highlight key words and terms. The effectiveness of the proposed algorithmic ensembles is quite high and reaches up to at best 88–90% accuracy.

Keywords: methods for intelligent processing of semi-structured data, ensemble algorithms, the problem of multidimensional clustering, the problem of multivariate classification, Recommender systems, research systems, scientific and analytical systems, Business Intelligence, pertinentnye algorithms

Задача интеллектуальной обработки больших объемов слабоструктурированной информации является в высшей степени актуальной. По версии Gartner, на 2017 г. выделены десять стратегических трендов, которые будут стимулировать развитие четырех важнейших направлений стратегического развития организаций [2]. Первое направление под названием «Всеохватывающая интеллектуализация» (Intelligence Everywhere), охватывает технологии и методы обработки данных, которые включают продвинутое машинное обучение, иискусственный интеллект и позволяют создавать интеллектуальные аппаратные и программные системы, способные учиться и адаптироваться. Второе направление включает в себя технологии, ориентированные на все более тесные связи между реальным и цифровым миром. Третье направление представляет собой объединение платформ и сервисов, необходимых для слияния интеллектуальных цифровых технологий. Четвертое направление включает в себя все аспекты адаптивной архитектуры безопасности.

Одной из областей применения интеллектуальных методов обработки слабоструктурированных научных данных являются системы, формирующие информационные предложения пользователю: научные рекомендательные системы (Recommender Systems), научные информационные системы, аналитические рекомендательные системы, системы Business Intelligence, системы для нахождения контента в сети Интернет.

Одной из мер, позволяющих оценить качество формируемых предложений, является соответствие невысказанной информационной потребности пользователя – пертинентность [36, 13]. Точность измерения данной характеристики является ограничивающим фактором для разработки действительно пертинентных предложений, и потому развитие указанной тематики представляется чрезвычайно важной и перспективной задачей. Проблемы, возникающие при формировании запросов пользователя, могут быть связаны с незнанием набора ключевых слов, однозначно определяющих семантику искомых документов, отсутствием достаточного опыта и квалификации формирования поисковых запросов, либо с отсутствием принятой и устоявшейся терминологии в интересующей предметной области. Все это обусловливает актуальность и значимость исследований, направленных на решение одной из ключевых проблем информационного поиска – проблемы адекватного отображения информационных потребностей пользователей и, как следствие, повышения пертинентности поиска.

Научные рекомендательные системы используются в научных организациях для создания межпредметных связей, в архивах и библиотеках – для сопоставления накопленной информации и навигации среди больших объемов литературы [8].

Системы Business Intelligence (BI) представляют собой совокупность методов и инструментов для перевода необработанной информации в осмысленную, удобную форму. Эти данные используются для бизнес-анализа. Технологии BI обрабатывают большие объемы неструктурированных данных, чтобы найти стратегические возможности для бизнеса. Рекомендательные системы, интегрированные с BI, с точки зрения аналитики выдают рекомендации по выбору инструментов анализа, при формировании отчетов проводят дополнение данными со сходной структурой, проводят сопоставление накопленных данных.

Развитие методов интеллектуальной обработки слабоструктурированных данных является весьма перспективным, так как рынок продуктов BI в последние годы растет. Например, рост выручки крупнейших российских компаний Ай-Теко, РДТЕХ, Прогноз, КРОК, HeliosIT на российском рынке BI в 2013 г.по отношению к 2014 г. составил от 8 до 28%, а в некоторых случаях (Форс, БАРС ГРУП) достиг 60%. По данным IDC за апрель 2013 г., мировые расходы на BI-сервисы будут в среднем увеличиваться на 14,3%, так что в 2016 г. они составят $70,8 млрд. По прогнозам Gartner, до 2016 г. рынок BI систем и аналитических платформ останется одним из наиболее быстро растущих сегментов мирового софтверного рынка. Среднегодовой темп роста этого рынка составит 7% в период с 2011 по 2016 г. К 2016 году объем рынка может достигнуть $17,1 млрд. При этом рынок BI, по данным Gartner от апреля 2012 г., если рассматривать его в совокупности с хранилищами данных и аналитикой в CRM, растёт еще быстрее – на 9% в год [9–10].

Отдельно стоит упомянуть так называемые content discovery platforms (средства для нахождения определенного контента). Одними из наиболее популярных представителей являются Outbrain и Taboola [10]. Outbrain используется крупными брендами как для рекламы своей продукции, так и для публикации рекомендаций со своих сайтов. Среди клиентов такие компании, как TIME, CNN, FastCompany. По информации от самой компании, этот продукт используется более чем на 35000 сайтов и выдает свыше 250 млрд рекомендаций и 15 млрд просмотров страниц в месяц. Эти рекомендации видят свыше 87% пользователей интернета в США.

Состояние проблемы

Оценку результата работы рекомендательных систем можно проводить с точки зрения релевантности и пертинентности [11]. Релевантность определяется как «соответствие полученной информации информационному запросу». Пертинентность как «соответствие полученной информации информационной потребности», т.е. пертинентность – соответствие найденных информационно-поисковой системой документов информационным потребностям пользователя, независимо от того, как полно и точно эта потребность выражена в форме запроса. Пертинентность определяется субъективным восприятием человека. Из-за субъективности пертинентности добиться точного совпадения нельзя: любая поисковая система настраивается на информационные нужды усредненного, а не конкретного пользователя. Недостаточная пертинентность поисков может быть обусловлена такими причинами, как излишняя декомпозиция запросов: информационная потребность пользователя раскрывалась в виде серии из 7–10 очень конкретных

запросов, или обслуживанием по чрезмерно широким запросам: на один запрос абонент получает от сотен тысяч до сотен миллионов документов, веб-страниц, хотя непосредственно соответствует запросу только малая часть информации.

Пертинентность информационного поиска зависит от двух обстоятельств: насколько точно формальный запрос, составленный пользователем, соответствует его информационной потребности и насколько хорошо результаты выдачи информационной системы соответствуют формальному запросу. Существующие методы позволяют создать модели поиска, обладающие высокой релевантностью ответов, т.е. обеспечивающие выдачу информационных предложений, близких по смыслу поисковому запросу. Современные поисковые системы научились достигать достаточно высокого значения этого параметра. Тем не менее с точки зрения пользователя главным критерием оценки результатов поиска информации является ее пертинентность.

Наиболее актуально вопрос составления пертинентного информационного предложения стоит в рекомендательных информационных системах в силу своего назначения. Большинство рекомендательных систем оперирует явным профилем пользователя, однако о том, чтобы оперировать неявными профилями (собранными на основе анализа неявного поведения – дельты времени пребывания на том или ином отрезке текста, факт копирования и т.п.) речи не идет. В то время как коммерческие системы активно внедряют системы анализа неявного поведения, т.к. в типовых ситуациях ожидается (и это уже начинает подтверждать практика, например, в интернет-гипермаркете Amazon), что рекомендательная система сможет демонстрировать результаты поиска информационных предложений еще до того, как человек самостоятельно сформулировал запрос, т.е. осуществлять предсказательный поиск.

Так как пертинентность представляет собой удовлетворённость пользователя результатами информационного поиска, либо подборкой информационных единиц, выданных рекомендательной системой, то для повышения этой характеристики, при формировании выдачи необходимо предсказывать интересы пользователя (информационную потребность). Информацию о пользователе рекомендательная система получает в первую очередь при его регистрации в научной системе. В таком случае, ориентируясь на указанные интересы, можно подобрать объекты из соответствующих разделов. При этом может понадобиться текстовый анализ. Контентная фильтрация формирует рекомендацию на основе поведения пользователя. Например, этот подход может использовать ретроспективную информацию о просмотрах (какие источники читает пользователь и характеристики этих источников). Этот контент может быть определен в ручном режиме или извлечен автоматически на основе других методов подобия. Однако этот подход является достаточно односторонним, так как текущие предпочтения пользователя могут быть не связаны с предшествующими. Также может иметь место проблема полноты и достоверности указанной информации, что приводит к низкой пертинентности рекомендации, которая опирается только на данный способ.

Следующий способ используется в торговых системах и основан на понятии «рыночной корзины», в этом случае анализируются покупки в одном чеке (торговой операции) и находятся наиболее часто встречающиеся предметные наборы. Таким образом, если пользователь в данный момент изучает информационную единицу (ИЕ), находясь на её странице в системе, то ему можно рекомендовать другие единицы по способу «с этим товаром обычно покупают …».

В отличие от указанного выше метода, можно учитывать поведение других пользователей в явном виде. Основная идея данного подхода в выделении паттернов поведения, определении, к какому из них относится текущий пользователь, а затем формирование рекомендации на основе усреднённых интересов пользователей из соответствующего паттерна. Паттерны определяются на основании анализа данных, содержащихся в логах системы. Коллаборативная фильтрация вырабатывает рекомендации, основанные на модели предшествующего поведения пользователя. Эта модель может быть построена исключительно на основе поведения данного пользователя или – что более эффективно – с учетом поведения других пользователей со сходными характеристиками. В тех случаях, когда коллаборативная фильтрация принимает во внимание поведение других пользователей, она использует знание о группе для выработки рекомендаций на основе подобия пользователей. По существу рекомендации базируются на автоматическом сотрудничестве множества пользователей и на выделении

(методом фильтрации) тех пользователей, которые демонстрируют схожие предпочтения или шаблоны поведения. В качестве примера предположим, что вы создаете вебсайт, чтобы предлагать его посетителям рекомендации относительно блогов. На основе информации от многих пользователей, которые подписываются на блоги и читают их, вы можете сгруппировать этих пользователей по их предпочтениям. Например, вы можете объединить в одну группу пользователей, которые читают несколько одних и тех же блогов. По этой информации вы идентифицируете самые популярные блоги среди тех, которые читают участники этой группы. Затем – конкретному пользователю этой группы – рекомендуется самый популярный блог из тех, на которые он еще не подписан и которые он не читает.

Принципы построения рекомендательных систем

Рекомендательные системы (РС) представляют собой программные средства и методы, назначением которых является прогнозирование поведения пользователя в отношении объекта информационного поиска и формирование рекомендаций для объектов, с которыми он еще не встречался [11–12]. Сформированные рекомендации помогают пользователю в различных процессах принятия решений, например, какую научную статью выбрать, какую дисциплину изучить и т.д. Такие рекомендации строятся на основании характеристик этих объектов и (или) профиля самого пользователя. Формирование рекомендаций возможно только на основе полученных данных. Данные, используемые РС, относятся к трем видам объектов: элементы, пользователи, транзакции, т.е. отношения между пользователями и элементами.

Элементами, или объектами информационного поиска может быть научная статья, книга и патент. Элемент может состоять из первичных термов, т.е. информационных единиц – слово, автор, название стиля и т.д. Элементы могут быть характеризованы их сложностью, значением или полезностью. Значение элемента может быть положительным, если элемент полезен для пользователя, или отрицателен, если элемент не является нужным и пользователь принял отрицательное решение при выборе его.

Пользователь системы может иметь свои вкусы и предпочтения. Информацию о пользователях можно собирать разными способами. В пользовательской модели всегда присутствует профиль пользователя, явный или неявный. Явный профиль формируется с помощью заполнения анкет, опросов и т.д. В этом случае пользователь персонализируется. Неявный профиль пользователя формируется за счет учета его действий на сайте.

Под транзакцией понимается зафиксированное взаимодействие между пользователем и РС. Например, журнал транзакций может содержать ссылку на элемент, выбранный пользователем и описанием контекста (например, пользовательский информационный запрос) для того, чтобы сформировать рекомендацию. Такая транзакция может также отражать наличие явной обратной связи, которую пользователь обеспечил, в виде оценки выбранного элемента. Фактически оценки являются самой популярной формой данных транзакции, которые собирает РС. Эти оценки могут быть собраны явно или неявно. В явном наборе оценок пользователя просят обеспечить ее мнение об элементе по рейтинговой шкале. В транзакциях, собирающих неявные оценки, система стремится выводить мнение пользователя на основе его действий. В диалоговых системах, т.е., системах, поддерживающих интерактивный процесс, более усовершенствованные модели транзакции. То есть пользователь может запросить рекомендацию, и система может произвести список предложения. Таким образом, РС на основе дополнительных пользовательских предпочтений предоставляет пользователю лучшие результаты.

В данной статье предлагается исследовать неявный профиль пользователей, формирующийся в научных и образовательных системах и сформулировать правила его обработки в целях уточнения пертинентных ответов в области научной и образовательной информации.

Рекомендательные системы классифицируются как контентные, коллаборативные и гибридные. При контентной фильтрации создаются профили пользователей и объектов. Профили пользователей могут включать демографическую информацию или ответы на определённый набор вопросов. Профили объектов могут включать названия дисциплин и учебных курсов, имена преподавателей и т.п.– в зависимости от типа объекта.

Контентная фильтрация ориентирована на четкую классификацию как пользователей, так и объектов, фигурирующих в информационном предложении. В указанном

случае устанавливается прямое соответствие между пользователями и объектами на основе их характеристик. В целом стратегия хорошо работает в областях с конечным и относительно небольшим количеством критериев оценки, вытекающих из природы вещей, при больших потоках информации и допускает большое количество критериев при небольшом информационном потоке. Основная проблема – классификация и создание новых информационных предложений.

При коллаборативной фильтрации используется информация о поведении пользователей в прошлом – например, информация о заказах или оценках. В этом случае не имеет значения, с какими типами объектов ведётся работа, но при этом могут учитываться неявные характеристики, которые сложно было бы учесть при создании профиля. Основная проблема этого типа рекомендательных систем – «холодный старт»: отсутствие данных о недавно появившихся в системе пользователях или объектах.

Коллаборативная фильтрация в большей мере опирается на анализ траектории пользователя: по каким ссылкам перешел, что и как оценил, на что поставил закладку, где воспользовался социальными кнопками, что искал, а также неявных действий – где задержался дольше, где меньше, анализ совокупных действий и поведения, анализ привычек и т.п. Рассматриваемая стратегия считается на сегодня наиболее перспективной, в т.ч. регулярно проводятся конкурсы среди научных групп для поиска наилучших алгоритмов.

Предлагаемый подход

В данной работе процесс получения рекомендаций конечным пользователем разбивается на несколько шагов. Основными источниками данных для пользователя можно считать его персональные данные, которые он указывает в анкете при регистрации на сайте, а также лог активности, который автоматически записывается системой, исходя из того, какие действия пользователь выполнил на сайте.

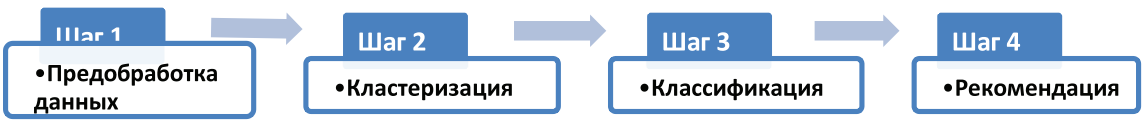
При формировании профиля пользователя, как явного, так и неявного, используемые наборы данных в своей структуре содержат не только количественные, категориальные и бинарные переменные, но среди них могут быть представлены и другие типы – текстовые данные [1]. К ним могут относиться различного рода справки, анкеты, рекомендации и прочие неструктурированные текстовые данные. Текстовые данные имеют особую природу происхождения и требуют особого подхода к обработке и дальнейшей кластеризации и классификации. Этот этап называется предобработкой данных. Этап предобработки включает в себя удаление пунктуации и стоп-слов, приведение символов к нижнему регистру, приведение слов к нормальной форме и бинаризацию данных.

Рекомендации выдаются на основе принадлежности пользователя к какой-то группе пользователей со схожими интересами. Определение таких классов поможет сформировать потребность пользователя более точно, т.е. повысить пертинентность. Данная операция называется кластеризацией. Отнесение нового пользователя к наиболее подходящему классу выполняется с помощью операции классификации. В результате, зная к какому классу относится пользователь, а также то, чем интересуются другие участники со схожими интересами, пользователю выдается рекомендация. Общая схема представлена на рисунке.

Несмотря на всю уникальность и особенность каждого пользователя сети Интернет, существует ряд исследований, доказывающий, что всех пользователей можно разделить на несколько классов [11]. Для последующей классификации, требуется извлечь из исходных данных поведенческие паттерны пользователей, которые можно получить путём применения методов кластерного анализа. Эти методы носят эмпирический характер, и как для работы многих из них, так и для оценки качества полученных решений требуется выдвижение гипотез как о количестве классов, так и об их предварительном составе (или физической интерпретации).

Ансамбль алгоритмов кластеризации

Данные, извлекаемые из неявного профиля пользователя рекомендательных систем, необходимо структурировать,



Последовательность действий для выдачи рекомендации

классифицировать, подвергать тщательному анализу. В этом случае кластерный анализ проводит сегментацию данных через выделение определённых объединений однородных элементов, которые рассматриваются как самостоятельные единицы, обладающие определёнными свойствами [1].

Устойчивость решений в задачах кластеризации может быть повышена благодаря формированию ансамбля алгоритмов и построению с его помощью коллективного решения на основе мнений участников ансамбля, где под мнением алгоритма подразумевается его вариант разбиения данных на кластеры. Данные свойства кластерного анализа особо актуальны при работе в областях с интенсивным использованием данных, когда предметная область слабо формализована, например, для анализа текстовых документов, изображений, построения неявных профилей пользователей рекомендательных систем и т.д. В том случае, если рассматриваемая область содержит различные типы данных, для выделения кластеров необходимо применять не один определённый алгоритм, а набор различных алгоритмов. Ансамблевый (коллективный) подход позволяет снизить зависимость конечного решения от выбранных параметров исходных алгоритмов и получить более устойчивое решение даже при большом количестве шумов и выбросов в данных.

Существуют следующие основные методики получения ансамбля алгоритмов: нахождение консенсусного разбиения, т.е. согласованного разбиения при имеющихся нескольких решениях, оптимального по некоторому критерию, и вычисление согласованной матрицы сходства/различий (co-occurrencematrix) [1].

При формировании окончательного решения используются результаты, полученные различными алгоритмами либо одним алгоритмом с различными значениями параметров, по разным подсистемам переменных и т.д. Применяя ансамбль с различными наборами алгоритмов, в соответствии с их преимуществами и особенностями, можно создать наиболее подходящую схему кластеризации для определённой предметной области. Учитывая тот факт, что выбор конкретной метрики расстояний между объектами является важным фактором, влияющим на результат кластеризации, можно существенно повысить эффективность кластерного анализа.

Предлагаемый авторами ансамбль алгоритмов объединяет два рассмотренных выше подхода и представляет собой сочетание последовательных алгоритмов K-средних, каждый из которых предлагает свое разбиение на основе изменяющейся метрики, и иерархического агломеративного алгоритма, объединяющего полученные решения с помощью особого механизма. Предложенный ансамбль опирается на результаты предварительного исследования исходных данных, которые представляют собой небольшой набор размеченных экспертами объектов. Минимально необходимый процент объема исходной выборки, гарантирующий заданную точность, подлежит дальнейшему изучению. Для определенности, в данной работе используется 0,5 %, что в случае увеличения объема данных, очевидно, должно подлежать пересмотру.

На первом шаге каждый алгоритм K-средних разбивает данные на кластеры, используя свою метрику расстояния. Затем, рассчитывается точность и вес мнения алгоритма в ансамбле по формуле

$$\omega_l = \frac{Acc_l}{\sum_{l=1}^{L} Acc_l}, \qquad (1)$$

где Acc_l – точность алгоритма l, т.е. отношение количества правильно кластеризованных объектов к объему всей выборки, а L – количество алгоритмов в ансамбле.

Для каждого полученного разбиения составляется предварительная бинарная матрица различий размера nxn, где n – количество объектов, необходимое для определения, занесены ли объекты разбиения в один класс. Затем рассчитывает согласованная матрица различий, каждый элемент которой представляется собой взвешенную (с использованием веса из формулы (1)) сумму элементов предварительных матриц. Полученная матрица используется в качестве входных данных для алгоритма иерархической агломеративной кластеризации. Затем с помощью обычных приемов, таких как определение скачка расстояния агломерации, можно выбрать наиболее подходящее кластерное решение.

В данном ансамбле алгоритмов кластеризации были использованы пять K-средних с изменением метрики (см. формулу (1)), как один из наиболее востребованных алгоритмов кластеризации больших данных. Для данных алгоритмов были использованы такие метрики, как:

• Евклидово расстояние

$$p(x, x') = \sqrt{\sum_{i}^{n} \left(x_i - x_i' \right)^2}. \qquad (2)$$

• Манхэттенское расстояние

$$p(x,x') = \sum_{i}^{n} \left| x_i - x_i' \right|. \quad (3)$$

• Расстояние Чебышева

$$p(x,x') = \max\left(\left| x_i - x_i' \right| \right). \quad (4)$$

• Коэффициент Жаккара

$$K(x,x') = \frac{\sum_i^n x_i x_i'}{\sum_i^n x_i^2 + \sum_i^n x_i'^2 - \sum_i^n x_i x_i'}. \quad (5)$$

• Динамическая трансформация временной шкалы (dynamictimewrapping, DTW)

$$DTW(x,x') = \frac{\min\left\{ \sum_{k=1}^{K} d(\omega_k) \right\}}{K}, \quad (6)$$

где K – длина пути между x и x', который вычисляется по специальной матрице трансформаций.

Для получения наилучшего разбиения на кластеры необходимо, как было упомянуто выше, составить бинарную матрицу сходства\различий на каждое L разбиение в ансамбле:

$$H_i = \{h_i(i,j)\}, \quad (7)$$

где $h_i(i, j)$ равен нулю, если элемент i и элемент j попали в один кластер, и 1, если нет.

Следующим шагом в составлении ансамбля алгоритмов кластеризации является составление согласованной матрицы бинарных разбиений.

$$H^* = \{h^*(i,j)\}, \quad (8)$$

$$h^*(i,j) = \sum_{i=1}^{L} w_l h_l(i,j), \quad (9)$$

где w_l – вес алгоритма.

Для формирования наилучшего разбиения по согласованной матрице был выбран алгоритм ближайшего соседа. Для снижения размерности исходных данных был выбран метод главных компонентов (Principal component analysis, PCA). В качестве критерия выбора количества компонент был выбран критерий Кайзера (собственное значение компоненты больше единицы).

Следующим шагом была выявлена точность каждого алгоритма путём сравнения полученного разбиения на два кластера каждым алгоритмом с кластерами, размеченными экспертным способом. После получения значения точности каждого алгоритма, по формуле (1) был рассчитан вес мнения алгоритма.

Ансамбль алгоритмов классификации

Для повышения точности работы алгоритма классификации авторами в [7] была предложена классификация контента.

Основополагающей гипотезой для данной работы было предположение о типах результатов в научных и исследовательских работах. Среди всех научных работ можно выделить несколько групп, которые наиболее полным образом описывают представленные результаты в этих работах. Первоочередным предположением о существовании подобной структуры можно выделить из разделения результатов научных исследований на теоретическую и практическую составляющую. В дополнение к этому можно выделить особый тип работ, который в качестве результата представляет собой обзор чужих результатов научной деятельности. Данный подход согласуется с результатами последних исследований в области классификации пользователей научных систем.

Данная гипотеза была формализована и адаптирована для решения задач классификации в области рекомендательных систем для научных социальных сетей. На ее основе был составлен набор типов пользователей, которые могут быть заинтересованы в получении контента одной из групп результатов научной деятельности. Дополнительно, для каждого типа были выделены словосочетания и ключевые слова, которые наиболее точно описывают каждую из них:

1. Обозреватели – аккумулирует полученный ранее результат и знания о предметной области, отсутствуют собственные разработки – сравнивают подходы и методы, приводят плюсы и минусы, обобщают факты, делают выводы.

2. Аналитики – целью данного типа является проведение сравнительного анализа, получение нового метода на основе ранее известных подходов. Ограничиваются только теоретическими аспектами, формируют теоретическую базу для будущих исследователей-практиков.

3. Практики – данный тип пытается найти практическое применение существующим методам и алгоритмам. Получают действующее решение, готовое для промышленной эксплуатации.

Указанные группы могут встречаться не только в чистом виде, но также возможен их симбиоз с другими группами.

Для решения поставленной задачи классификации авторами предлагается подход, основанный на применении нескольких

различных алгоритмов, голосующих независимо. Данный подход позволяет повысить итоговую точность работы классификатора и снизить вычислительную сложность алгоритмов. Такой набор алгоритмов, как правило, состоит из простых алгоритмов машинного обучения и условно называется ансамбль голосующих алгоритмов. Усиление простых классификаторов – подход к решению задачи классификации (распознавания) путём комбинирования примитивных слабых классификаторов в один сильный. Под силой классификатора в данном случае подразумевается эффективность (качество) решения задачи классификации. При построении ансамбля используются различные комбинации, основанные на указании различных весов алгоритмов, применение одинаковых алгоритмов с разными параметрами, сегментирование данных под разные алгоритмы и т.п. Для построения ансамбля голосующих алгоритмов будут использоваться следующие простые алгоритмы: наивный Байесовский классификатор, метод опорных векторов с линейным ядром, метод опорных векторов с ядром радиальной базисной функции Гаусса и случайный лес деревьев.

Классификатор «наивного» Байеса использует формулу Байеса для расчета вероятности. Идея алгоритма заключается в расчете условной вероятности принадлежности объекта к классу при равенстве его независимых переменных определенным значениям. Цель классификации состоит в том, чтобы понять, к какому классу принадлежит документ, поэтому нам нужна не сама вероятность, а наиболее вероятный класс. Наивный байесовский классификатор объединяет модель с правилом решения. Одно общее правило должно выбрать наиболее вероятную гипотезу; оно известно как апостериорное правило принятия решения (MAP).

Метод опорных векторов (SVM, support vector machine) – набор схожих алгоритмов обучения с учителем, использующихся для задач классификации и регрессионного анализа [12, 13]. Основная идея метода – перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей наши классы. Разделяющей будет гиперплоскость, максимизирующая расстояние до двух параллельных гиперплоскостей. Алгоритм работает в предположении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора. Для решения нелинейных задач используют способ создания нелинейного классификатора, в основе которого лежит переход от скалярных произведений к произвольным ядрам, позволяющий строить нелинейные разделители. Наиболее распространенными ядрами являются: полиномиальное (однородное); радиальная базисная функция; радиальная базисная функция Гаусса; сигмоид.

Классификатор Random Forest (RF) использует ансамбль решающих деревьев. Само по себе решающее дерево не обеспечивает достаточной точности для этой задачи, но отличается быстротой построения. Алгоритм *RF* обучает *k* решающих деревьев на параметрах, случайно выбранных для каждого дерева, после чего на каждом из тестов проводится голосование среди обученного ансамбля. В основе построения этого алгоритма лежит идея о том, что если агрегировать данные от большого количества различных слабых алгоритмов, сведя их в единый ответ, то результат, скорее всего, будет лучше, чем у одного мощного алгоритма. Классификация объектов проводится путём голосования: каждое дерево комитета относит классифицируемый объект к одному из классов, и побеждает класс, за который проголосовало наибольшее число деревьев. Оптимальное число деревьев подбирается таким образом, чтобы минимизировать ошибку классификатора на тестовой выборке. В случае её отсутствия, минимизируется оценка ошибки out-of-bag: доля примеров обучающей выборки, неправильно классифицируемых комитетом, если не учитывать голоса деревьев на примерах, входящих в их собственную обучающую подвыборку.

Предлагаемая модель ансамбля голосующих алгоритмов строится на следующем подходе.

1. Для каждого алгоритма необходимо определить меру однородности классификации по каждому классу на обучающей выборке. При расчёте меры однородности для алгоритмов классификации будет использоваться энтропия (см. формулу (1)) для двух независимых случайных событий «*x*» с «*n*» возможными состояниями (от 1 до «*n*»), где «*p*» – функция вероятности:

$$H(x) = -\sum_{i=1}^{n} p(i)\log_2 p(i). \qquad (10)$$

2. Для каждой статьи рассчитывается вероятность отнесения к классу, и эти значения поочередно умножаются на меру однородности по каждому классу. Из полученных произведений выбирается максимальное значение.

3. В данном случае мера однородности выступает в роли поправочного коэффициента. Результат классификации рассчитывается по формуле (2), где «*h*» – мера однородности алгоритма для класса, а «*p*» – вероятность отнесения объекта к классу:

$$C = \text{argmax}(h_i p_{ij}, i \in (\text{от 1 до 4}). \quad (11)$$

Алгоритм автоматического предложения информационного предложения

Алгоритм автоматического формирования информационного предложения использует поведенческие данные в целях удовлетворения текущей неявной информационной потребности пользователя.

Помимо выявления информационной потребности пользователя, ее можно сформировать, и в этом случае можно использовать понятие долговременного и кратковременного трендов (далее «топа» и «тренда»). К «топовым» информационным единицам можно отнести те, которые наиболее часто ищут или просматривают другие пользователи в течение долгосрочного периода (кварталы и полугодия), кратковременные – те которые являются популярными в течение нескольких дней и недель. Окно просмотра может быть настроено более точно, в зависимости от сферы человеческой деятельности, к которой относятся информационные единицы (ИЕ).

В данной работе для повышения пертинентности предлагается использовать комбинированный подход. Длина рекомендованного кортежа должна учитывать особенности человеческой природы, которые определяют число объектов, которыми может оперировать человек одновременно в 7 ± 2. Сам кортеж будет включать следующие части:

1) ИЕ, полученные по пользователям, похожим на текущего;

2) ИЕ, полученные из наиболее часто встречающихся наборов;

3) ИЕ, относящиеся к долговременным трендам («топы»);

4) ИЕ, относящиеся к кратковременным трендам, актуальным («тренды»).

Пункты (1) и (2) относятся к восстанавливаемой информационной потребности и предполагаются вносящими наибольший вклад в рекомендованный набор (порядка 60–70 %). Пункты (3) и (4) относятся к формируемой информационной потребности и должны занимать 30–40 % от объёма кортежа.

При построении научной или образовательной системы, необходимо учитывать, что они имеют свою специфику. Как видно из описания научных сетей, приведенных выше, пользователь таких систем взаимодействует с информационными единицами – научными результатами.

Результаты исследования и их обсуждение

Для проведения эксперимента по апробации алгоритмов кластеризации и классификации были собраны исходные данные о публикациях по двум российским научным публичным электронным ресурсам – elibrary.ru и cyberleninka.ru. С этой целью была создана специализированная программа-краулер, собравшая основные сведения о статьях – такие как название, автор, ключевые слова, аннотация и т.д. Полученные данные (сведения о более чем 500 тыс. статей) специальным образом были предобработаны для выделения основных слов и терминов, относящихся к разработанной онтологической модели. После нормализации слов часть статей, относящихся к различной тематике, такой как физика, математика, информатика, а также экономика и управление, была размечена с помощью полуавтоматического подхода, включающего метод кластеризации k-means и экспертную оценку.

Для тестирования и оценки ансамбля алгоритмов кластеризации использовалось программное средство Rapid Miner, с помощью которого можно решать как исследовательские, так и прикладные задачи интеллектуального анализа данных, включая анализ текста, анализ мультимедиа, анализ потоков данных, что подходит для тестирования ансамбля алгоритмов кластеризации. Применяя предложенный ансамбль алгоритмов кластеризации, можно повысить достоверность разбиения данных на группы. Существенным является то, что данный метод может применяться в различных областях. Предложенный в данном проекте ансамбль алгоритмов кластеризации нивелирует недостатки метрик расстояний для алгоритмов K-средних, тем самым повышая достоверность разбиения. С помощью предложенного ансамбля удалось обеспечить точность не менее 90 %, а по каждому параметру не менее 80 %.

Для тестирования алгоритма классификации из обработанных данных для эксперимента была извлечена выборка научных публикаций, объём которой составлял 5000 объектов, в равном количестве для каждого направления. Для формирования обучающей и тестовой выборки, исходная выборка разбивалась в процентном соотношении 70:30. Эффективность предложенного алгоритмического ансамбля достаточно высока и достигает в наилучшем случае 88% точности. Тем не менее в эксперименте рассматривались только часть возможных значений параметров алгоритмов и не все возможные тематики статей, поэтому требуются дополнительные исследования в этой области. Предложенный подход показал высокую эффективность в смысле точности классификации научного контента.

Заключение

Проведенное исследование показало, что повышение пертинентности информации в различных информационных системах является крайне актуальной и востребованной задачей. Предложенный метод повышения пертинентности, включающий в себя ансамбли алгоритмов для решения задачи кластеризации и классификации, показал свою работоспособность и перспективность использования. Его успешная реализация позволит существенно повысить качество выполнения запросов пользователей и качество предоставляемых документов. Область применения подобных методов крайне широка, в данном исследовании выделяются научные и аналитические рекомендательные системы, системы BI и системы нахождения контента.

Работа поддержана грантом РФФИ № 15-07-08742.

Список литературы

1. Бочкарёв П.В., Киреев В.С. Разработка ансамбля алгоритмов кластеризации на основе изменяющихся метрик расстояний // Аналитика и управление данными в областях с интенсивным использованием данных: XVIII международная конференция DAMDID/RSDL'2016 (11–14 октября 2016 г., Ершово, Московская область, Россия): труды конференции. – М.: ФИЦ ИУ РАН, 2016. – С. 69–73.

2. Гореткина Е. Десять стратегических технологий 2017 года// PCWeek. – 2016. – № 20 (919). – URL: https://www.pcweek.ru/idea/article/detail.php?ID=189532 (дата обращения 25.11.2016).

3. Гусева А.И., Киреев В.С., Филиппов С.А., Бочкарёв П.В., Кузнецов И.А. Научные и образовательные рекомендательные системы // Информационные технологии в образовании XXI века: сборник научных трудов Международной научно-практической конференции. – М.: НИЯУ МИФИ. 2015. – С. 33–40.

4. Жуликов С.Е., Жуликова О.В. Проблема пертинентности современных информационно-поисковых систем // Вестник Тамбовского государственного университета. – 2013. – № 18. – С. 224–226.

5. Иванова О.Г., Громов Ю.Ю., Дидрих В.Е., Поляков Д.В. Нечеткий подход к определению пертинентности результатов поиска и выбору оптимального запроса // Вестник Воронежского института ФСИН России. – 2011. – № 2. – С. 49–54.

6. Клеменков П.А. Построение новостного рекомендательного сервиса реального времени с использованием NOSQL СУБД // Информатика и ее применения. – 2013. – Т. 7. – № 3. – С. 14–21.

7. Кузнецов И.А., Киреев В.С. Разработка ансамбля алгоритмов классификации с использованием энтропийного показателя качества для решения задачи поведенческого скоринга // Аналитика и управление данными в областях с интенсивным использованием данных: XVIII международная конференция DAMDID/RSDL'2016 (11–14 октября 2016 г., Ершово, Московская область, Россия): труды конференции. – М.: ФИЦ ИУ РАН, 2016. – С. 79–85.

8. Кулагина М.В., Лопатенко А.С. Научные информационные системы и электронные библиотеки. Потребность в интеграции [Электронный источник]. – URL: http://infoculture.rsl.ru/donArch/home/news/ dek/2001/10/2001-10_r_dek-s1.htm (дата обращения 24.08.2016).

9. Обзор: Бизнес-аналитика и большие данные [Электронный источник]. – URL: http://www.cnews.ru/reviews/bi_bigdata_2015/review_table// (дата обращения 24.06.2016).

10. Системы бизнес-аналитики в России 2013 [Электронный источник]. – URL: http://www.rbcgrp.com/files/QlikView_TAdviser2013.pdf (дата обращения 24.06.2016).

11. Guseva A.I., Kireev V.S., Bochkarev P.V., Smirnov D.S., Filippov S.A. The Formation of User Model in Scientific Recommender Systems //I nternational Review of Management and Marketing. – 2016. – № 6(S6). – P. 214–220.

12. Ricci F., Rokach L., ShapiraD., Kantor P.B. Recommender Systems. Handbook, DOI 10.1007/978-0-387-85820-3_1, © Springer Science+Business Media, LLC 2011.

13. Philippov S., Zakharov V., Stupnikov S., Kovalev D. Organization of Big Data in the Global E-commerce Platforms. Ceurworkshopproceedings, Vol-1536 [Selected Papers of the XVII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2015)]. – Obninsk, Russia, October 13-16, 2015. – P. 119–124.