ИПМ им.М.В.Келдыша РАН

Абрау-2016 • Труды конференции

А.С. Мохов

**Методика использования
программного комплекса TextCat для
обработки и анализа научных
документов**

Рекомендуемая форма библиографической ссылки

Мохов А.С. Методика использования программного комплекса TextCat для обработки и анализа научных документов // Научный сервис в сети Интернет: труды XVIII Всероссийской научной конференции (19-24 сентября 2016 г., г. Новороссийск). — М.: ИПМ им. М.В.Келдыша, 2016. — С. 274-278. — doi:[10.20948/abrau-2016-18](https://doi.org/10.20948/abrau-2016-18)

Методика использования программного комплекса *TextCat* для обработки и анализа научных документов

А.С.Мохов

НИУ «Московский энергетический институт»

Аннотация. В статье рассматриваются функциональные возможности программного комплекса TextCat, предназначенного для обработки и классификации двуязычных (русско-английских) текстовых документов. Приведены результаты классификации на трех группах выборок, как по техническим наукам, так и по естественно-научным и гуманитарным.

Ключевые слова: Интеллектуальный анализ текстовых данных, информационный поиск, классификация двуязычных библиографических документов, цифровые библиотеки.

1. Введение

В последние десятилетия существенно увеличились объемы накопленной информации и возросли потребности практики в разработке программно-алгоритмических средств обработки и анализа текстовых данных. Наиболее высокие темпы роста документального потока наблюдаются в научно-технической области. Усиливающаяся информационная перегрузка существенно расширяет число организаций и отдельных пользователей, готовых для повышения эффективности своей деятельности активно использовать процедуры интеллектуального анализа текстов (Text Mining). Как следствие, активизируется разработка систем Text Mining, которые в зависимости от размера обрабатываемого документального массива делятся на три основные группы.

1. Системы, ориентированные на поиск и анализ больших данных (Big Data), включая Web-трафик, сведения из социальных сетей, сообщения электронной почты.

2. Системы, предназначенные для автоматизации документооборота и поддержки принятия решений в крупных корпорациях, банках, государственных организациях, университетах, новостных агентствах.

3. Системы обработки информации для удовлетворения потребностей и снижения информационной перегрузки отдельных пользователей (или небольших коллективов пользователей).

Наибольшая потребность в персональных системах сложилась в научном сообществе, представители которого работают в условиях значительной информационной перегрузки. Для снижения этой перегрузки пользователь нуждается в специальном математическом обеспечении, способном выполнять функции «интеллектуального ассистента». Такой «интеллектуальный ассистент» должен автоматизировать ряд рутинных операций по получению, обработке и предварительному анализу научной информации. В качестве пользователей этих систем выступают специалисты-предметники, научные сотрудники, преподаватели, эксперты, программисты, аспиранты. Для удовлетворения информационных потребностей данной категории пользователей на кафедре Управления и информатики НИУ «Московский энергетический институт» разработан программный комплекс (ПК) *TextCat*.

2. Основные функциональные возможности программного комплекса *TextCat*

ПК *TextCat* предназначен для проведения предварительной обработки текстовых данных, обучения классификаторов, осуществления классификации и фильтрации документов. Отличительными особенностями ПК является использование наряду с известными классификаторами оригинальных высокоточных методов классификации, разработанных автором, а также возможности обработки двуязычной (русско-английской) информации и учета местоположения терминов в различных разделах документов [1,2]. ПК состоит из следующих основных модулей:

Модуль загрузки и обновления текстовых данных. В нем реализуется ввод документальной информации из внешних источников и ее преобразование во внутренний формат. Предусмотрены процедуры удаления полных дубликатов и автоматической загрузки документов с указанных Интернет-адресов (например, сайтов журналов и конференций, цифровых библиотек).

Модуль предварительной обработки данных. В нем осуществляется обработка русскоязычных, англоязычных, двуязычных (русско-английских) документов, проводится отсечение стоп-слов, стемминг, взвешивание терминов (для этого предусмотрены широко известные операции *tf*- и *tf-idf* взвешивания, а также взвешивание на основе совстречаемости термина и класса в документах выборки). В модуле проводится анализ структуры документа и предоставляется возможность взвешивания терминов с учетом местоположения в различных разделах текста [2].

Модуль обучения классификаторов. В нем из документов, имеющих во внутренней базе данных, формируются обучающие и экзаменационные выборки, которые далее используются для настройки параметров методов классификации. В ПК реализовано 15 методов, часть из них относится к хорошо известным классификаторам (метод центроидов – МЦ, метод ближайших соседей – метод к-БС, наивный байесовский метод – НБМ), остальные классификаторы являются профильными методами или их

модификациями (профильные методы – это методы, в которых рассчитывается формальный объект – профиль класса, способный характеризовать все остальные элементы класса при классификации новых документов) [1,2]. Для определения профилей проводится выявление наиболее информативных терминов для каждого класса с помощью специальных процедур (статистического χ^2 - критерия, критерия взаимной информации, эвристического критерия (коэффициент Жаккара) и комбинированного подхода, являющегося сочетанием вышеперечисленных критериев). В модуле для пользователя предлагаются рекомендуемые значения настраиваемых параметров (длина профиля, число ближайших соседей, мера близости и т.п.), предусмотрена возможность просмотра сформированных в ходе обучения профилей классов (терминов, упорядоченных по значениям весов).

Модуль синтеза коллективов решающих правил (КРП). Модуль позволяет из имеющихся 15 классификаторов формировать коллективы решающих правил, состоящие из трех, пяти или семи членов. Отбор метода в КРП проводится на основе расчета двух показателей: точности и разнородности (способности методов, входящих в КРП, делать ошибки на разных документах массива). Наиболее предпочтительным является объединение в едином коллективе самых точных (и одновременно) самых разнородных классификаторов.

Модуль классификации и фильтрации. В нем предусматривается использование ранее обученных классификаторов для обработки новых данных (решение различных задач пользователя, например, категоризация русскоязычных (или англоязычных) публикаций, фильтрация статей, получаемых из профильных журналов, выявление дубликатов).

Модуль интерфейса, который обеспечивает взаимодействие пользователя с другими модулями.

3. Методика использования ПК *TextCat*

ПК *TextCat* разработан для проведения высокоточной классификации научных публикаций в различных предметных областях (инженерно-технической, естественно-научной, социально-экономической, гуманитарной). Последовательность действий по его применению представлена в следующей методике.

Для использования программного комплекса необходимо запустить установочную программу, устанавливающую ПК *TextCat* на локальный компьютер, после чего пользователь может осуществлять все остальные действия, необходимые для дальнейшей классификации документов.

В соответствии с информационной потребностью пользователя и областью профессиональных интересов формируются обучающие и экзаменационные выборки, используя возможности цифровых библиотек eLibrary и КиберЛенинка (или документальные массивы, предоставленные

пользователем). Выборки могут состоять из русскоязычных, англоязычных и двуязычных (русско-английских) документов.

После загрузки выборок в базу программного комплекса, В ПК автоматически проводится предварительная обработка данных (удаление стоп-слов, стемминг) и взвешивание терминов. Документы представляются в виде частично структурированной модели, учитывающей местоположение терминов в различных разделах документов.

Далее пользователь может провести обучение и рассчитать ошибку для трех групп методов:

- «классические» классификаторы (МЦ, метод к-БС, НБМ),
- известные профильные методы (статистические РО-профиль и Q-профиль, теоретико-информационный НМИ-профиль и эвристический J-профиль и профиль Сокала-Сниса (С-С-)),
- разработанные автором модификации профильных методов (UNI1, UNI2, UNI3, UNI5, UNI6, два обобщенных профильных метода учитывающих местоположение термина в документе)

В случае близких значений ошибок, когда разные методы оказываются лучшими на различных экзаменационных выборках, для выявления наиболее точного метода применяется непараметрический статистический критерий Уилкоксона.

В случае если, используя указанные методы, не удалось достичь точности, удовлетворяющей пользователя, формируются коллективы решающих правил, состоящие из пяти классификаторов (отбираются наиболее точные и разнородные методы, обученные на предыдущих шагах). При необходимости дополнительно повысить точность синтезируется КРП, состоящий из семи классификаторов. При этом, в коллектив могут отбираться как методы обученные на одноязычных выборках (русских или английских), так и на двуязычных выборках.

Если точность сформированных КРП не удовлетворяет пользователя, то необходимо заново сформировать представительные выборки (в частности, увеличив их размер, уменьшив количество заданных классов, объединив сильно пересекающиеся тематики). После этого необходимо повторить процедуру классификации указанными методами.

Таким образом, использование программного комплекса позволяет обеспечить более высокую точность классификации по сравнению с известными решениями за счет использования методов классификации, предложенных автором. Кроме того, ПК *TextCat* автоматизирует процесс получения и анализа информации в интересах пользователя (специалиста-предметника) и повышает тем самым эффективность научной деятельности.

Проверка работоспособности предложенной методики была проведена на трех группах выборок. Из документов по Информатике было сформировано 11 выборок, из документов по Праву и судебной психиатрии – 5, из документов по Нейронаукам – 4 выборки. Известные профильные методы (РО-, НМИ-, С-С- J-)

показали более высокую точность (на всех выборках) по сравнению с наиболее точным классическим методом к-ближайших соседей. UNI5 и UNI6 продемонстрировали более высокую точность по сравнению с известными профильными методами. Наилучшую точность удалось получить путем синтеза КРП, состоящего из пяти членов, обученных как на одноязычных, так и на двуязычных выборках. Результаты приведены в таблице 1:

Таблица 1. Средние ошибки классификации для всех методов по трем группам выборок, %

Метод	Информатика	Право и судебная психиатрия	Нейронауки
КБС	15,41	16,38	15,71
РО	11,16	11,23	13,57
НМИ	10,56	13,51	11,90
С-С	12,64	12,95	15,47
J	11,87	12,19	12,64
UNI5	10,05	10,85	10,95
UNI6	9,96	9,52	11,42
КРП	7,18	9,32	8,09

ПК *TextCat* внедрен в Институте проблем химической физики РАН, где данный комплекс успешно применяется для обработки базы данных статей, написанных сотрудниками института. ПК *TextCat* прошел также апробацию на кафедре Управления и информатики НИУ «МЭИ» и используется в настоящее время для автоматизированного отслеживания научных публикаций в электронных журналах по тематикам, указанным специалистами кафедры.

Литература

1. Мохов А.С. Толчеев В.О. Разработка методов высокоточной классификации двуязычных текстовых библиографических документов // Информационные технологии, №5, 2014, с. 8-13.
2. Мохов А.С., Толчеев В.О. Разработка новых модификаций профильных методов классификации и построение коллективов решающих правил // Заводская лаборатория. №3 (81), 2015. с 70-75.