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. **№ 5** **Москва 2020**

ИНФОРМАЦИОННЫЕ СИСТЕМЫ

УДК 004.91

В.С. Егоров, Е.С. Козлова, К.Е. Ломотин, О.В. Федорец, А.В. Филимонов, А.В. Шапкин

Система автоматической классификации текстов для обработки потока научных публикаций в ВИНИТИ РАН*

Представлены результаты разработки и тестирования системы автоматической классификации научных текстов, позволяющей определять тематику текстов по трём классификационным схемам в пакетном и диалоговом режимах. Описаны структурно-функциональные компоненты, используемые методы оценки качества классификации, методика обучения, выбор оптимальной модели классификации, основные направления внедрения автоматического классификатора в технологию обработки электронного документального потока в ВИНИТИ РАН.

Ключевые слова: *автоматическая классификация текста, Word2Vec, машинное обучение, персептрон, логистическая регрессия, обработка естественного языка, производственная технология информационного центра*

DOI: 10.36535/0548-0027-2020-05-1

* Работа выполнена в рамках Госзадания № 0003-2019-0002 "Разработка, лексикографическое и понятийно-терминологическое развитие системы взаимосвязанных классификаций научно-технической информации (ГРНТИ, УДК, ВАК, Scopus, WoS, MSC, МКС). Развитие научно-методической и программно-технологической компонент информационной технологии формирования и анализа качества реферативной базы данных ВИНИТИ. Разработка и экспертиза стандартов системы СИБИД".

ВВЕДЕНИЕ

Технологической основой любого информационного комплекса (СМИ, сайты Интернета, библиотеки, центры научно-информационного обслуживания и т. д.) является производственная система, обеспечивающая для пользователей удобную и оперативную навигацию доступа к создаваемым информационным массивам. Традиционно построение подобной системы базируется на реализации тематической классификации обрабатываемых объектов. Для ВИНИТИ объекты обработки – это научные публикации, т. е. текстовая информация.

Определимся с терминологией, которую будем использовать. Под классификацией текста мы подразумеваем его индексирование по рубрикатору. Здесь термин "рубрикатор" – это синоним термина "классификационная схема"; соответственно, термин "рубрика" – синоним термина "класс". Индексирование по рубрикатору (рубрицирование) – процедура присвоения объекту индексов рубрик (одной или нескольких), взятых из заранее определённого списка тематических рубрик. Типичные примеры – индексирование литературы в библиотеках, или документов информационных центрах. Наиболее известные классификаторы – Универсальная десятичная классификация, Десятичная классификация Дьюи, Библиотечно-библиографическая классификация, Государственный рубрикатор научно-технической информации, Международная патентная классификация.

Присвоение документу классификационного индекса – это весьма дорогостоящая операция, поскольку требует привлечения квалифицированных специалистов в предметной области.

Успехи развития современных компьютерных технологий в области интеллектуального анализа данных позволяют создавать программы автоматической классификации, "обученные" анализировать тексты с целью отнесения их к тем или иным классам конкретного рубрикатора − с той или иной степенью вероятности.

Разработке подобных систем посвящается значительное число теоретических и прикладных исследований в области информационных технологий [1–4]. Примеры практического использования систем автоматической классификации текстов в области обработки научно-технической информации можно найти в литературе последних лет [5–9], где выделяются общие обсуждаемые проблемы:

- выбор моделей классификации и программных инструментов, позволяющих выполнять "интеллектуальную" обработку текстов;
- выбор методов предварительной обработки текста (нормализация, векторизация и пр.);
- подбор гиперпараметров модели классификации, наилучшим образом соответствующих характеру обрабатываемых текстов и используемых классификационных схем;
- обучение автоматического классификатора на коллекциях текстов в заданной области знания;
- оценка качества классификации и достоверности результатов.

В настоящей статье представлена система автоматической тематической классификации текстов (АТКТ), разработанная в ВИНИТИ РАН с использованием известных на настоящее время методов обработки документов, доступных инструментов и программных продуктов, пригодных для индексации [10-12]. Мы постарались дать общие сведения о построении системы АТКТ и особенностях ее реализации в аспекте вышеуказанных проблем, а также показать начальный опыт ее использования для обработки потока поступающих в ВИНИТИ документов.

Отметим некоторые черты технологического процесса обработки научно-технической литературы в ВИНИТИ, знание которых потребуется для понимания целей и задач рассматриваемой разработки.

ТЕХНОЛОГИЯ ИНДЕКСАЦИИ ДОКУМЕНТОВ В ВИНИТИ РАН

Сегодня в ВИНИТИ РАН работают 16 профилированных по тематике отделов – научно-редакционных подразделений: Автоматика и радиоэлектроника, Астрономия, Биология, География, Геология, Информатика, Математика, Машиностроение, Металлургия, Механика, Охрана окружающей среды, Транспорт, Физика, Химия, Экономика промышленности, Электротехника.

Общий объем входного потока документов в настоящее время составляет более 1 млн в год. При этом доля русскоязычных документов составляет 29%, англоязычных − 63%. Ежемесячно формируются и издаются около 190 тематических выпусков Реферативного журнала (РЖ), при этом в разных отделах число выпусков варьируется от 2-3-х до нескольких десятков. В процессе обработки документы проходят три стадии тематического классифицирования.

На первой стадии технологического процесса выполняется разметка входного потока по областям науки и техники − для каждого документа определяется научно-тематический отдел, специалисты которого будут его обрабатывать.

На второй стадии (в отделах) документы в соответствии с тематикой распределяются по выпускам Реферативного журнала. Направление документа в соответствующий выпуск РЖ, как правило, означает определение научного сотрудника, ответственного за его обработку.

На третьей стадии обработки документов научные редакторы − специалисты в своих предметных областях − осуществляют глубокое индексирование документов по Рубрикатору ВИНИТИ, который представляет собой иерархический классификатор, включающий более 50 тыс. рубрик.

Таким образом, задача тематического классифицирования решается в ВИНИТИ последовательно на трех уровнях: *отдел → выпуск РЖ → индекс Рубрикатора*. Применительно к системе АТКТ это означает, что она должна "уметь" определять тематику документа в соответствии с тремя классификационными схемами, которые будем называть "вариантами классификации" и для краткости обозначать:

1) *КС-ОНТ* − области науки и техники − список отделов (16 классов);
2) *КС-РЖ* − выпуски РЖ − список из 194 классов;

3) *КС-ГРНТИ* − тематические классы Рубрикатора; при этом на данном этапе решается упрощенная задача – грубое определение классов на 2-м уровне ГРНТИ[1], включающем около 800 индексов (из них тематике ВИНИТИ соответствуют лишь 460). Это обусловлено практической невозможностью обучить систему работе на всём 50-тысячном множестве классов Рубрикатора ВИНИТИ.

Конкретные классы будем называть "рубриками". Каждая рубрика имеет два атрибута − уникальный индекс и название. Примеры рубрик:

Рубрикатор	*Индекс*	*Название*
КС-ОНТ	e3	Биология
	f7	Химия
КС-РЖ	04Р1	Биотехнология. Бионанотехнологии. Бионаноматериалы
	04Т4	Токсикология
	04Т6	Фармакология
КС-ГРНТИ	62.09	Сырье и продуценты для биотехнологического производства
	62.13	Биотехнологические процессы и аппараты
	62.33	Клеточная инженерия
	62.35	Технологическая биоэнергетика
	62.37	Прикладная генетическая инженерия
	62.39	Инженерная энзимология
	62.41	Иммунобиотехнологические методы анализа

Этот пример, в частности, демонстрирует различную глубину определения тематики в рассматриваемых рубрикаторах.

Материалом для работы системы АТКТ являются русско- и англоязычные научные тексты, извлекаемые из метаданных научно-технической литературы, поступающей в ВИНИТИ во входном потоке. Существенными для классификации элементами данных являются оригинальное заглавие, авторская аннотация и ключевые слова. Результат работы АТКТ − это оценка вероятности соответствия заданного текста рубрикам в различных вариантах классификации.

Построение АТКТ основано на технологии машинного обучения, для чего используются коллекции документов из политематической базы данных ВИНИТИ. Обучающие выборки содержат результаты ручной обработки документов, выполненной профильными специалистами при формировании информационных продуктов. Для обучения используются заглавия, рефераты и ключевые слова документов, снабжённых кодами отделов, кодами выпусков РЖ, индексами Рубрикатора ГРНТИ.

В оперативном доступе имеется несколько миллионов документов, обработанных за последние годы, что вполне достаточно для обучения системы и исследований. Причем эти документы в наибольшей мере соответствуют решаемой задаче, так как позволяют обучать систему автоклассификации в условиях, максимально приближенных к области её планируемого применения. Немаловажно и то, что большой объем имеющихся материалов позволяет добиваться достаточной представительности рубрик в обучающих выборках, варьировать в широком диапазоне структуру выборок для обеспечения нужной равномерности представления рубрик.

Методы машинного обучения предполагают многократность их выполнения с последующим анализом результатов, внесением изменений в управляющие параметры системы и корректировкой исходной выборки. Мониторинг соответствия результатов ручного и автоматического индексирования является важным средством оценки качества АТКТ и выработки рекомендаций по её совершенствованию. Применительно к ВИНИТИ правильность автоматически получаемых рубрик (отдел, выпуск РЖ, индекс ГРНТИ) может быть проверена по результатам ручной обработки документов в процессе формирования Реферативного журнала.

ПОСТРОЕНИЕ АВТОКЛАССИФИКАТОРА. ОБУЧЕНИЕ И НАСТРОЙКА

Основными структурно-функциональными компонентами программного комплекса АТКТ являются текстовый предобработчик, векторайзер и классификатор.

Модуль предобработчика текста решает задачу очистки входного документа от элементов разметки, служебных символов, а также выполняет лемматизацию, т. е. приведение всех слов к начальной форме. В результате текст переводится в универсальное внутреннее представление, независимое от входного формата. Для учета специфики русско- и англоязычных текстов в АТКТ предусмотрены соответствующие средства.

Векторайзер преобразует текст в его векторную модель заданной размерности. Для этого в АТКТ используется технология *Word2Vec* [13], что позволяет представить текст как числовой вектор, сохранив при этом его семантику в виде наиболее весомых или усредненных тематических признаков слов, составляющих текст. Основной параметр *Word2Vec* − это размерность получаемого вектора; она определяет детальность выражения смысла в тематическом пространстве.

Классификатор на основе векторного представления текста принимает решение о том, к каким рубрикам относится текст, и оценивает его вероятность для каждой рубрики. Классификатор формирует тематический профиль текста − по всем вариантам классификационных схем, работе с которыми он обучен. В этом модуле реализована возможность применения нескольких наиболее известных в настоящее время алгоритмов машинного обучения и моделей принятия решений, которые используют методы обработки естественного языка на основе компьютерного анализа с применением средств искусственного интеллекта и компьютерной лингвистики.

[1] ГРНТИ – Государственный рубрикатор научно-технической информации России. Имеет три уровня иерархии. Рубрикатор ВИНИТИ фактически является его развитием до 9-го уровня.

Программное приложение АТКТ реализовано на платформонезависимом интерпретируемом языке программирования *Python* версии 3.6 в среде 64-разрядной ОС *Windows* версии не ниже 7. При разработке нами использованы следующие компоненты:

1) *Scikit-Learn* – модуль Python, который содержит модели машинного обучения и средства обработки данных;
2) *Gensim* – библиотека, реализующая модель *Word2Vec*;
3) *Pandas* – библиотека, предоставляющая контейнер, построенный по реляционному принципу и используемый для операций с данными;
4) *PyQt5* – интерфейс фреймворка *Qt* для *Python*; применяется для реализации событийно-ориентированной архитектуры и межпотокового взаимодействия, а также для создания пользовательского интерфейса;
5) *PyMystem3* – интерфейс библиотеки *Mystem* для *Python*; служит для лемматизации текста;
6) *Textblob lemmatizer* – модуль библиотеки обработки естественного языка *Textblob*, предназначенный для лемматизации;
7) *NLTK* – набор инструментов для обработки естественного языка; включает, в частности, стеммер Портера, стеммер Ланкастера, лемматизатор на основе базы лексики *WordNet*.

Работоспособность системы АТКТ определяется, в частности, правильным выбором настроек ее функционирования. В нашем случае определяющими среди них можно считать такие факторы:

- выбор алгоритмов предобработки текстов (очистки, лемматизации и пр. – с учётом особенностей естественных языков);
- параметры векторизации текста для модели *Word2Vec*;
- оценка эффективности различных моделей машинного обучения, доступных для классификации; оптимизация их настроек (также называемых гиперпараметрами) в соответствии с применяемыми классификаторами;
- формирование "правильных" обучающих выборок − достаточного объёма, сбалансированных по классам, достаточными размерами текстов и пр.

Точно оценить вклад каждого фактора в конечный результат не представляется возможным, но можно его оценить, т. е. измерить качество результатов автоматического индексирования. Решение таких задач обеспечивает специально построенная система обучения АТКТ, которая позволяет в автоматизированном режиме проводить сравнительные исследования алгоритмов, заложенных в АТКТ, с использованием различных обучающих выборок и различных параметров настройки.

Для оценки качества автоматической классификации используются изобретённые в прошлом веке для оценки результатов информационного поиска [14–16] традиционные меры:

- доля правильных ответов (*accuracy*);
- точность (*precision*)
- полнота (*recall*);
- *F*-мера (F_1-мера, *F-measure*, f_1 *score*).

Эти меры вычисляются по отдельности для каждой рубрики во всех вариантах классификации. Для обобщённой оценки качества индексирования используется макро-усреднение (*macro-averaging*) и микро-усреднение (*micro-averaging*) перечисленных мер. Оценка качества базируется на сравнении результатов автоматического и экспертного индексирования.

Прежде чем перейти к математическим формулировкам, необходимо подчеркнуть принципиальную разницу между множеством рубрик, присвоенных документу компьютерной программой (алгоритмом), и множеством рубрик, присвоенных документу экспертом (человеком). Алгоритм классифицирует согласно заложенной в него математической модели, поэтому выдаёт нечёткое множество упорядоченных пар (код рубрики, релевантность рубрики). По сути, релевантность является функцией принадлежности рубрики нечёткому множеству, так как указывает, в какой степени элемент принадлежит множеству. Как и функция принадлежности, релевантность принимает значения в числовом отрезке [0, 1].

В отличие от алгоритма, человек назначает документу обычное (чёткое) множество рубрик, в котором функция принадлежности каждого элемента равна единице. Теоретически экспертизу можно усложнить: попросить экспертов упорядочить рубрики по убыванию их соответствия тематике документа, т. е. ранжировать. Эту информацию можно было бы использовать для более точного сравнения результатов автоматического и экспертного индексирования. Однако, не имея возможности постфактум получить от экспертов ранги, считаем все рубрики эксперта равнозначными.

Таким образом, необходимо сравнивать четкие и нечёткие множества рубрик. Самый очевидный подход – избавиться от нечёткости, получив чёткое множество рубрик из нечёткого множества, присвоенного документу алгоритмом. С этой целью применяются два метода отбора рубрик чёткого множества – пороговый и ранговый.

Пороговый метод использует пороговое значение функции принадлежности, в нашем случае это пороговое значение релевантности. Пусть A нечёткое множество с функцией принадлежности $\mu_A(x)$=$relevant_A(x)$:

$$A = \left\{ \left(x, \mu_A(x)\right) \mid x \in X \right\}, \text{ где } \mu_A(x) \in [0,1],$$

тогда чёткое множество A_α, которое называется альфа-срезом (α-срезом) нечёткого множества A, определяется следующим образом:

$$A_\alpha = \left\{ x \in X \mid \mu_A(x) \geq \alpha \right\},$$

где α – пороговое значение функции принадлежности.

Ранговый метод использует пороговое значение ранга. Ранг элемента x в множестве B – $rank_B(x)$ – это порядковый номер элемента x в нечётком множест-

ве B, упорядоченном по убыванию функции принадлежности. Тогда, если B упорядоченное нечёткое множество:

$$B = \left\{ \left(x, \mu_B(x), rank_B(x) \right) \mid x \in X \right\},$$

где $rank_B(x) > 0$,

то чёткое множество B_r ранга r (ранговый r-срез нечёткого множества B) получаем следующим образом:

$$B_r = \left\{ x \in X \mid rank_B(x) \le r \right\},$$

где r – пороговое значение ранга.

Мы используем оба метода отбора элементов из нечёткого множества рубрик. Ранговый метод применяется для вычисления точности, полноты и F_1-меры; пороговый метод – для вычисления меры сходства двух альтернативных множеств рубрик, присвоенных одному документу, если одно из них нечёткое, а другое чёткое.

Приведем математические формулировки точности, полноты, F_1-меры и меры сходства.

Пусть D – множество документов, отнесённых алгоритмом к рубрике; E – множество документов, отнесённых экспертами к рубрике, а величины tp, fp, fn, tn определены как мощности конечных множеств:

$$tp = \left| \left\{ x \mid x \in D \,\&\, x \in E \right\} \right|$$
$$fp = \left| \left\{ x \mid x \in D \,\&\, x \notin E \right\} \right|$$
$$fn = \left| \left\{ x \mid x \notin D \,\&\, x \in E \right\} \right|$$
$$tn = \left| \left\{ x \mid x \notin D \,\&\, x \notin E \right\} \right|,$$

тогда доля правильных ответов (*accuracy*), точность (*precision*) и полнота (*recall*) вычисляются по формулам:

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$
$$precision = \frac{tp}{tp + fp}$$
$$recall = \frac{tp}{tp + fn}.$$

Обобщённой мерой качества индексирования будем считать F_1-меру (f_1 *score*), которая равна среднему гармоническому точности и полноты:

$$f_1score = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 \cdot precision \cdot recall}{precision + recall}.$$

Такую меру иногда называют сбалансированной F-мерой, так как значения *precision* и *recall* имеют одинаковый вес.

Описанные меры (*accuracy*, *precision*, *recall*, f_1*score*) рассчитываются по каждой рубрике и после завершения обучения выдаются в отчёт системой АТКТ, что позволяет сразу оценить его успешность.

Обычно обучение длится долго и запускается редко, поэтому дополнительно, для оперативного контроля качества индексирования между обучениями, нужны другие меры, способные вычисляться на любом подмножестве документов, т. е. не только по рубрикам. Например, можно оценивать качество индексирования по отраслям науки, по видам документов (статьи, книги, патенты и т.д.), по издательствам, по статьям из заданного списка научных журналов и т. п. С этой целью удобно измерить точность, полноту и F_1-меру для каждого документа в подмножестве, а затем использовать средние значения этих мер для оценки всего подмножества.

Введём понятие точности и полноты индексирования для документа с идентификатором *ID*. Пусть D_{ID} – множество рубрик, присвоенных алгоритмом документу *ID*, E_{ID} – множество рубрик, присвоенных экспертами документу *ID*. Тогда точность ($precision_{ID}$), полнота ($recall_{ID}$) и F_1-мера (f_1score_{ID}) для документа *ID* имеют следующий вид:

$$precision_{ID} = \frac{\left| D_{ID} \cap E_{ID} \right|}{\left| D_{ID} \right|}$$

$$recall_{ID} = \frac{\left| D_{ID} \cap E_{ID} \right|}{\left| E_{ID} \right|}$$

$$f_1score_{ID} = \frac{2 \cdot precision_{ID} \cdot recall_{ID}}{precision_{ID} + recall_{ID}}. \qquad (1)$$

Нетрудно убедиться, что формула (1) преобразуется следующим образом:

$$f_1score_{ID} = \frac{2 \cdot \left| D_{ID} \cap E_{ID} \right|}{\left| D_{ID} \right| + \left| E_{ID} \right|}. \qquad (2)$$

Эта бинарная мера сходства предложена датским биологом Торвальдом Сёренсеном (Thorvald Sørensen) в 1948 г. В русскоязычных источниках она обычно называется мерой или коэффициентом Сёренсена [17, 18].

На сегодняшний день нами проведены исследования семи алгоритмов машинного обучения и влияния их гиперпараметров на качество индексирования. Это логистическая регрессия (*logistic regression*), метод опорных векторов (*svm*), персептрон (*perceptron*), метод *k*-ближайших соседей (*knn*), дерево решений (*decision tree*), случайный лес (*random forest*), адаптивный бустинг (*AdaBoost*).

Эксперименты по выбору моделей и подбору значений гиперпараметров проводились на обучающих выборках небольшого объёма (менее 100 тыс. документов). Такой объём выборки позволяет обучить несколько моделей в течение 1-2 недель на сервере с 8 Гбайт ОЗУ и двумя 4-ядерными процессорами *Intel Xeon* с тактовой частотой 2,5 ГГц.

Скорость обучения зависит не только от объёма обучающей выборки, но и от выбранной размерности векторного пространства модели *Word2Vec*, а также от параметров настройки моделей классификации. При слишком больших значениях размерности или гиперпараметров процесс обучения выполняется слишком долго, поэтому его приходится останавливать, уменьшать значения и запускать заново. Именно по причине низкой скорости обучения пришлось отказаться от метода опорных векторов (*svm*), несмотря на его высокие показатели качества при малых размерностях *Word2Vec* (50-100). Логистическая регрессия и персептрон демонстрировали близкие к *svm* показатели качества индексирования наряду с высокой скоростью обучения. Поэтому было решено сосредоточиться на этих двух моделях.

По завершении обучения мы проводили тестирование полученной модели – выполнялось индексирование документов тестовой выборки.

В обучающие и тестовые выборки отбираются документы, опубликованные в тематических выпусках Реферативного журнала ВИНИТИ. Для обучения используются названия, рефераты и ключевые слова документов, снабженные кодами отделов, выпусков РЖ, индексами Рубрикатора ВИНИТИ и ГРНТИ. В среднем объем документа составляет 850 символов; число тематических кодов – от 1 до 3 для каждой классификационной схемы.

При формировании выборок соблюдались правила: обучающая и тестовая выборки не пересекаются; объём тестовой выборки не менее 1/3 объёма обучающей выборки; год издания документов обеих выборок относится примерно к одному хронологическому периоду длительностью 2-3 года (допускается сдвиг хронологического отрезка тестовой выборки вперёд на 1 год).

По окончании цикла обучения-тестирования формируются отчёты − по каждой модели и для каждого из трех вариантов классификации (ОНТ, РЖ, ГРНТИ). Отчеты включают меры *Accuracy*, *Precision*, *Recall*, f_1 *score*, исходные значения переменных *tp*, *fp*, *fn*, *tn*, средние значения мер с использованием макро- и микро-усреднения (*macro-average* и *micro-average*). Статистика выдаётся по пяти ранговым срезам: $r \in \{1, 2, 3, 4, n\}$, где n означает "все ответы классификатора". Для примера в табл. 1 показана статистика по двум ранговым срезам для варианта *КС-ОНТ + percepron*[200] – при размерности *Word2Vec,* равной 400; объёмы обучающей и тестовой выборок равны 90 тыс. и 45 тыс. документов, соответственно.

Из табл. 1 видно, как возрастает полнота (*recall*) и убывает точность (*precision*) при увеличении порогового значения ранга с 1-го до 2-х, а также, что мера *accuracy* бесполезна на тестовых выборках, в которых количество документов по рубрикам существенно различается. Обучающие и тестовые выборки должны быть репрезентативными, поэтому они, хотя и в сглаженном (сбалансированном) виде, должны отражать неравномерность частотного распределения научных публикаций по тематикам науки. Для *КС-ОНТ* мы считаем выборку сбалансированной, если количество документов в рубриках различается не более чем в 10 раз. На практике количество документов по рубрикам различается более значительно. Мы видим, что у двух рубрик, наиболее слабо представленных в потоке документов (это ″информатика" и "астрономия"), оказались самые высокие значения *accuracy,* что является завышенной оценкой качества классификации. На самом деле, согласно более объективной оценке f_1*score*, ″информатика" стоит только на пятом месте, а "астрономия" вообще имеет f_1*score* ниже среднего.

Таблица 1

Статистика тестирования модели персептрон[200] для *КС-ОНТ*

Отдел	Один ответ классификатора (r = 1)				Два ответа классификатора (r = 2)				*tp*+*fn*
	accuracy	*precision*	*recall*	f_1*score*	*accuracy*	*precision*	*recall*	f_1*score*	
Авт. и радиоэлектр.	0,963	0,756	0,794	0,774	0,881	0,399	0,941	0,560	4558
Астрономия	0,991	0,788	0,760	0,774	0,979	0,478	0,889	0,622	1097
Биология	0,989	0,970	0,947	0,958	0,925	0,638	0,985	0,774	7338
География	0,983	0,870	0,778	0,822	0,938	0,449	0,934	0,607	2883
Геология	0,978	0,775	0,844	0,808	0,931	0,438	0,936	0,597	3104
Информатика	0,995	0,820	0,886	0,851	0,979	0,418	0,946	0,580	883
Математика	0,992	0,915	0,817	0,863	0,968	0,504	0,910	0,649	1852
Машиностроение	0,968	0,789	0,753	0,771	0,905	0,422	0,918	0,579	4031
Металлургия	0,975	0,733	0,742	0,737	0,941	0,437	0,906	0,590	2653
Механика	0,984	0,724	0,662	0,692	0,958	0,373	0,834	0,516	1514
Охрана окр. среды	0,986	0,775	0,881	0,825	0,912	0,290	0,964	0,446	2080
Транспорт	0,978	0,743	0,871	0,802	0,942	0,471	0,949	0,630	2945
Физика	0,972	0,850	0,828	0,839	0,909	0,493	0,946	0,648	4992
Химия	0,954	0,827	0,883	0,854	0,820	0,458	0,966	0,622	8642
Экономика пром-сти	0,983	0,894	0,848	0,871	0,965	0,665	0,943	0,780	3766
Электротехника	0,969	0,846	0,715	0,775	0,927	0,501	0,877	0,638	4155
Среднее									
macro	0,979	0,817	0,813	0,813	0,930	0,465	0,928	0,615	3530,8
micro	0,979	0,830	0,830	0,830	0,930	0,469	0,939	0,626	-

Примечание: *percepron*[200] – это модель персептрон с 200-ми элементами в одном скрытом слое

Таблица 2

Влияние размерности Word2Vec на качество классификации

Модель классификатора	Классификационная схема	кол-во рубрик	max(*macro* f_1 *score*) / max(*micro* f_1 *score*)		
			***vector_dim*=50**	***vector_dim*=100**	***vector_dim*=400**
Perceptron[100]	*КС-ОНТ*	16	0,63 / 0,67	0,69 / 0,72	0,75 / 0,77
Logistic regression	*КС-ОНТ*	16	0,60 / 0,64	0,61 / 0,66	0,67 / 0,70
Perceptron[100]	*КС-РЖ*	196	0,31 / 0,39	0,55 / 0,55	0,56 / 0,58
Logistic regression	*КС-РЖ*	196	0,40 / 0,42	0,54 / 0,54	0,81 / 0,78
Perceptron[100]	*КС-ГРНТИ*	436	0,12 / 0,38	0,62 / 0,62	0,67 / 0,69
Logistic regression	*КС-ГРНТИ*	436	0,25 / 0,36	0,68 / 0,62	0,95 / 0,86

Примечание: В этой таблице используются оба метода усреднения *F*-меры (макро- и микро-).

В результате серии экспериментов мы наблюдали интересный эффект. Если количество рубрик классификационной схемы невелико, то увеличение размерности *Word2Vec* не оказывает существенного влияния на качество индексирования. Но если количество рубрик превышает размерность *Word2Vec*, то это существенно улучшает качество индексирования.

Приведём конкретный пример. Список отделов ВИНИТИ содержит только 16 позиций, поэтому размерность 50-100 оказывается вполне достаточной. Варианты классификации *КС-РЖ* и *КС ГРНТИ* содержат более 100 рубрик – 196 и 436 соответственно. Видимо поэтому для них размерность *Word2Vec*, равная 50, оказалась явно недостаточной. Табл. 2 демонстрирует, какое влияние на *F*-меру оказывает увеличение размерности векторного представления слов (*vector_dim*) на примере обучения двух моделей: персептрона с одним скрытым слоем (100 элементов), и логистической регрессии.

Из табл. 2 следует простое эмпирическое правило: чем больше классов в классификационной схеме, тем больше должна быть размерность модели *Word2Vec*. При переходе от размерности 50 к размерности 100 качество индексирования по выпускам РЖ и по ГРНТИ существенно повысилось у обеих моделей. При переходе от размерности 100 к размерности 400 качество индексирования существенно повысилось только у логистической регрессии. Возможно, персептрону недостаточно 100 ассоциативных элементов в скрытом слое, чтобы улучшить результат при увеличении количества входных элементов до 400.

Последующие эксперименты с *КС-ОНТ* показали, что увеличение количества ассоциативных элементов со 100 до 200 улучшает показатели max(*macro* f_1 *score*) и max(*micro* f_1 *score*) с 0,75 и 0,77 до 0,81 и 0,83, соответственно. С использованием *КС-РЖ* и *КС-ГРНТИ* подобные эксперименты пока не проводились. Настройка гиперпараметров нейронной сети для автоматической индексации текстов при изменении размерности векторного представления слов и количества классов – это отдельная тема, которая находится за рамками настоящей статьи.

В общем виде технология выбора модели индексирования и оптимизации значений гиперпараметров похожа на конкурс, который проходит в несколько туров.

В первом туре эксперименты проводятся с небольшими обучающими выборками (менее 100 тыс. документов) и небольшой размерностью *Word2Vec* (50-100 координат), так как на этом этапе важно иметь высокую скорость обучения для перебора большого количества моделей и комбинаций гиперпараметров.

В следующие туры проходят наиболее перспективные модели, с которыми проводятся эксперименты по увеличению объёма обучающей выборки и размерности *Word2Vec*, а также по дальнейшей настройке гиперпараметров. Такой подход базируется на предположении, что модели, показавшие относительно низкое качество индексации на маленьких выборках и размерностях векторов, покажут относительно низкое качество и при их увеличении. Не для всех моделей и комбинаций гиперпараметров это соотношение верно, но полная проверка этого предположения требует огромных вычислительных ресурсов. На наших выборках наиболее перспективными оказались логистическая регрессия и персептрон. От метода опорных векторов пришлось отказаться из-за низкой скорости обучения, остальные модели показали невысокое качество индексирования на небольших выборках и размерностях *Word2Vec*.

В итоге для каждого варианта индексирования выбиралась своя модель принятия решений с набором гиперпараметров, оптимизирующим целевую метрику. В нашем случае точность и полнота индексирования одинаково важны, поэтому максимизируется F_1-мера. Аналогичным образом можно максимизировать точность, полноту, несбалансированную F_β-меру – выбор целевой метрики зависит от приоритетов внедрения автоклассификатора в конкретную информационную систему.

Вначале опишем алгоритм выбора модели в общем виде, затем приведём конкретный пример. Обозначим $F(i, A_j)$ значение целевой функции, вычисленное на срезе A_j нечёткого множества рубрик, присвоенных документам при использовании *i*-й модели. Соберём все значения целевой функции в матрицу $(a_{i,j})$ размера $m \times n$, где $a_{i,j} = F(i, A_j)$, $i = \overline{1, m}$, $j = \overline{1, n}$, m – количество исследуемых моделей, n – количество ранговых *r*-срезов нечёткого множества рубрик. Оптимальной назовём модель, которой соответствует строка с номером s, если $a_{s,j} = \max\limits_{i,j} a_{i,j}$.

Таблица 3

Матрица выбора оптимальной модели для *КС-ОНТ*

Модель классификатора	*macro* f_1 *score* на трёх ранговых срезах		
	***r*=1**	***r*=2**	***r*=3**
Perceptron[100]	0,752	0,595	0,462
Perceptron[200]	**0,813**	0,615	0,471
Logistic regression	0,665	0,568	0,465

Таблица 4

Средний коэффициент Сёренсена при различных пороговых значениях *α*

Классификационная схема	α=0,0	α=0,1	**α=0,2**	α=0,3	α=0,4	α=0,5	α=0,6	α=0,7	α=0,8	α=0,9
КС-ОНТ	0,461	0,649	**0,651**	0,640	0,620	0,594	0,560	0,522	0,475	0,407
КС-РЖ	0,295	0,481	**0,505**	0,486	0,437	0,364	0,288	0,217	0,145	0,072
КС-ГРНТИ	0,150	0,410	**0,427**	0,415	0,383	0,331	0,276	0,223	0,164	0,097

Описанный алгоритм продемонстрируем на примере варианта классификации *КС-ОНТ*. Обучающая выборка 90 тыс. документов, размерность *Word2Vec* равна 400. Исследуются три модели на трёх ранговых срезах. В качестве целевой функции используется *macro* f_1 *score*. Полученные значения целевой метрики сведены в матрицу размера 3×3 (табл. 3), где видно, что максимальное значение целевой функции (0,813) дает модель *perceptron*[200]. Таким образом, именно эту модель следует использовать в варианте *КС-ОНТ*.

Для изучения сходства результатов автоматического и экспертного индексирования системой АТКТ мы проиндексировали более 1 млн документов, опубликованных в РЖ ВИНИТИ в 2017-2018 гг. При этом для *КС-РЖ* и *КС-ГРНТИ* применялась логистическая регрессия, для *КС-ОНТ* – персептрон. В качестве меры сходства был использован коэффициент Сёренсена, вычисленный для каждого документа при пороговых значениях релевантности от 0,0 до 0,9 с шагом 0,1.

Коэффициент Сёренсена для документа вычисляется по формуле (2) и совпадает с F_1-мерой, вычисляемой по формуле (1). Средние значения коэффициентов собраны в табл. 4, позволяющую выбрать оптимальное пороговое значение релевантности − в отдельности для каждого варианта. Для этого выбирается значение α, при котором средний коэффициент Сёренсена достигает максимального значения.

Согласно табл. 4, оптимальное пороговое значение релевантности равно 0,2, причём для всех трёх вариантов классификации.

ОПЫТ ПРАКТИЧЕСКОГО ПРИМЕНЕНИЯ АВТОКЛАССИФИКАТОРА В ВИНИТИ РАН

Система АТКТ работает в ВИНИТИ с начала 2019 г. Она установлена на специально выделенном сервере, который предоставляет внешним приложениям услугу: в ответ на предъявленный текст выдаёт взвешенный список тематических рубрик по любой из трех классификационных схем − отделы, выпуски РЖ, индексы ГРНТИ. Это позволяет использовать систему в производственном процессе и рассматривать перспективные направления её применения.

Реализуемую в ВИНИТИ систему АТКТ удобно представить в виде многоярусной схемы, при которой на каждом ярусе используются результаты работы предыдущего яруса:

- на первом ярусе входной поток документов (естественно, его часть в электронной форме) подвергается автоматической индексации по всем предусмотренным классификационным схемам; результаты складываются в таблицы Единой технологической базы данных (ЕТБД), доступные для всех приложений;
- на втором ярусе результаты автоиндексации могут использоваться как рекомендации при тематическом индексировании документов − на разных стадиях обработки;
- на третьем ярусе можно проводить сравнение результатов автоиндексации с результатами ручной индексирования, выполненного научными сотрудниками при формировании информационных продуктов ВИНИТИ;
- на четвертом ярусе можно определять недостатки автоиндексации и выдвигать предложения по необходимости переобучения системы АТКТ или по совершенствованию моделей и алгоритмов АТКТ.

Функции 1-го яруса выполняет программное приложение "Электронный эксперт", которое управляет процессом автоиндексации: формирует потоки данных для АТКТ, осуществляет запуск системы АТКТ, загрузку и обработку результатов ее работы в ЕТБД. Программа работает ежедневно по расписанию и обрабатывает документы входного потока, поступившие в ВИНИТИ за последние сутки. В результате формируется массив, в котором отображаются показатели вероятности отнесения документа к той или иной рубрике классификационной схемы.

Программа оптимизирована для обработки больших массивов документов. Оптимизация состоит в разделении входного потока документов для АТКТ на порции и реализации изолированной параллельной обработки порций. Производительность Электронного эксперта оценивалась на выборке порядка миллиона документов при различных значениях количества документов в порции и количества одновременно запущенных процессов АТКТ, на различных конфигурациях компьютеров. Объемы порции варьировались от 100 до 10000 документов, а количество потоков - от 1 до количества ядер компьютера. Эксперименты показали, что на 64-разрядной рабочей станции с частотой 2,5 ГГц с 4 ядрами и 8 Гб памяти среднее время классификации (равное 30 мс) достигается при использовании параллельно двух процессов АТКТ и размере порции 2000 документов. При сокращении размера порции до 1000 документов и увеличении до шести количества параллельных процессов АТКТ среднее время индексирования будет равно 10 мс. При такой скорости можно классифицировать 2,9–8,6 млн документов в сутки по одной классификационной схеме. Поскольку в ВИНИТИ каждый документ индексируется по трём схемам, производительность Электронного эксперта на практике составляет 1–3 млн документов в сутки, что значительно превышает потребности Института.

Первое, наиболее очевидное, применение результаты работы Электронного эксперта нашли в Отделении обработки входного потока научно-технической литературы, где осуществляется научная систематизация (тематическая разметка) документов для направления в отделы научной информации ВИНИТИ. Здесь результаты классификации в варианте *КС-ОНТ* используются в качестве "подсказок" для разметчика. Соответствующие функции внедрены в автоматизированное рабочее место для поддержки принятия решений. В дальнейшем предполагается продвинуться в сторону автоматической разметки, когда некоторые (пока ограниченные, специально выделенные) материалы входного потока можно направлять в тематические отделы, целиком полагаясь на результат автоиндексации, т. е. без участия оператора-разметчика. Автоматический режим можно предположить осуществимым, например, для отдельных научных журналов, литературы некоторых издательств и определённых тематик и т. п.

Другое направление использования результатов АТКТ связано с повышением уровня представления в продуктах ВИНИТИ документов входного потока. Как отмечалось выше, традиционные продукты Института охватывают не более 60% поступающих документов, а 40% документов не проходят никакой обработки и не отражаются в РЖ и БД. Между тем пользователям могут быть полезны просто библиографические сведения о новых публикациях. И здесь весьма уместно использовать результаты автоиндексации по ГРНТИ и предоставить пользователям услугу просмотра новых поступлений литературы по тематическим рубрикам – с существенно более широким охватом, чем в традиционных продуктах. Дополнительным бонусом для пользователей может быть и то, что автоиндексатор дает более широкий спектр тематических принадлежностей публикаций, чем это делаются при ручной обработке узкопрофильные специалисты (это особенно важно в связи с возрастанием доли многодисциплинарных публикаций, на что не успевают должным образом реагировать устоявшиеся технологии). Работа над такими продуктами ведётся, и мы рассчитываем в ближайшее время представить их на суд потребителей.

Третье направление применения АТКТ – это внедрение системы в практику работы научных сотрудников, осуществляющих аналитико-синтетическую переработку документов в отраслевых отделах Института. В настоящее время проводятся эксперименты по подключению данных Электронного эксперта в качестве рекомендаций при индексировании документов по Рубрикатору ВИНИТИ. Кроме того, представляется перспективным дать научным сотрудникам возможность поиска во входном потоке документов, которые в действительности соответствуют тематике отраслевых отделов или выпусков РЖ, но по каким-либо причинам (возможно ошибочно) не были направлены в отдел в результате ручной разметки.

Еще раз подчеркнем, что результаты работы Электронного эксперта носят консультационный характер; в конечном счёте, ответственность за правильность научной систематизации несет человек. Понятно, что уровень доверия к результатам автоиндексации зависит как от субъективных предпочтений оценивающего, так и от объективных факторов, определяемых степенью совершенства алгоритмов системы АТКТ.

Качество работы системы АТКТ подвергается непрерывному мониторингу посредством анализа результатов Электронного эксперта. С этой целью в технологический контур внедрён комплекс процедур на языке Transact SQL, которые в автоматическом режиме обеспечивают предоставление информации о работе Электронного эксперта: количество обработанных документов, оперативные оценки полноты и точности классификации, оповещения о выходе этих показателей за допустимые пределы. Для анализа используются результаты ручного индексирования документов, выполненного научными сотрудниками ВИНИТИ при формировании выпусков РЖ. По данным мониторинга принимаются решения о переобучении системы АТКТ.

В частности, анализ показывает явную неравномерность качества автоиндексации при обработке документов из различных областей науки и техники. Если для публикаций по биологии, экономике промышленности, математике, физике, химии система АТКТ дает хорошие чётко выраженные ответы, то для других областей результаты получаются более размытыми, с большим количеством альтернатив, без явных приоритетов. Это можно объяснить двумя причинами.

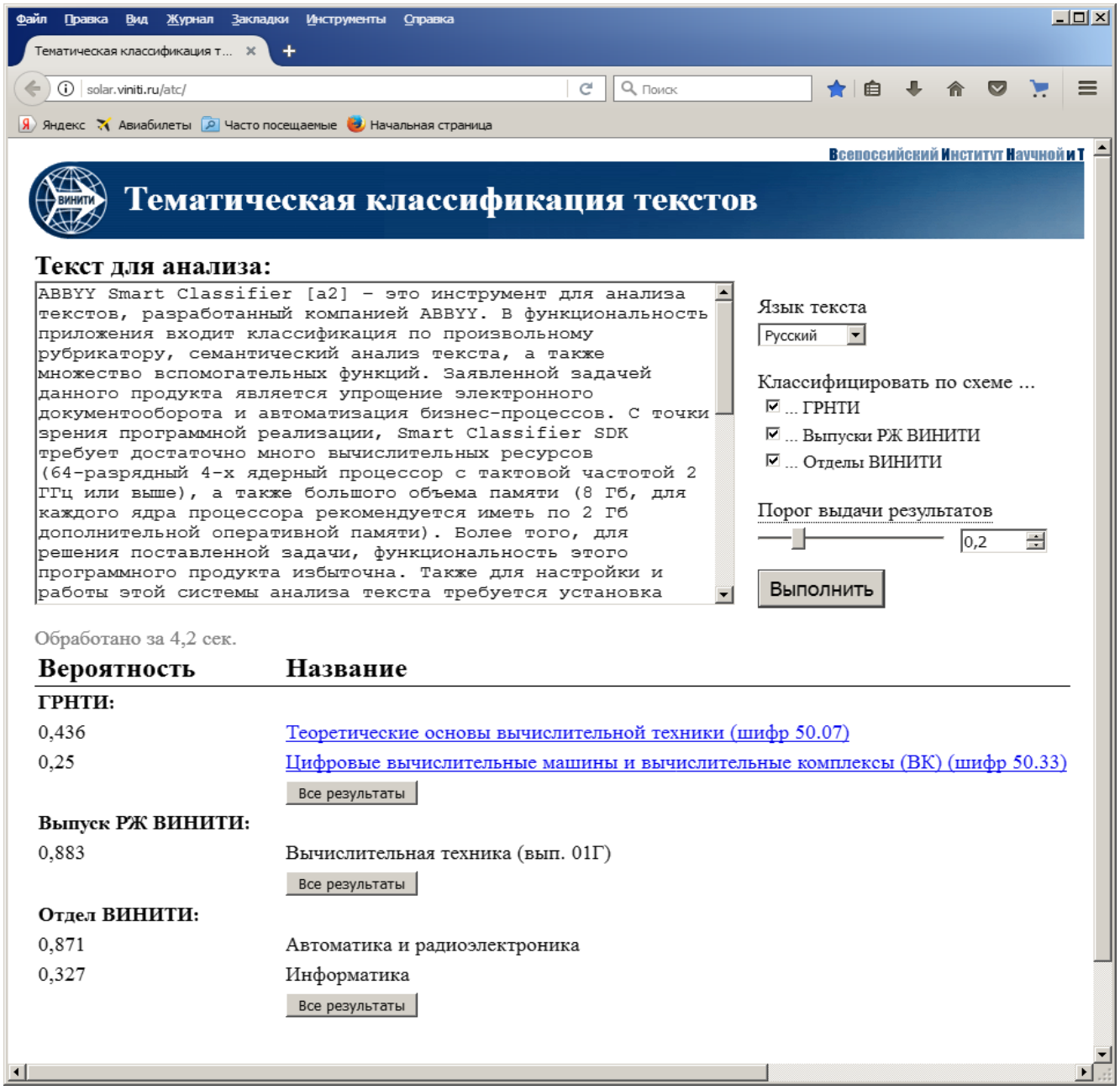
Во-первых, – свойствами текстов. Например, публикации по биологии или химии содержат существенно больше характерных терминов, чем, скажем, публикации по машиностроению или транспорту.

Во-вторых, – недостатками классификационных схем, связанными с условностями деления на отрасли

в политематических и междисциплинарных областях. Например, классификационные схемы некоторых выпусков РЖ по транспорту, машиностроению, автоматике и радиоэлектронике в значительной степени пересекаются.

Из этих наблюдений очевидно, что степень практического применения методов автоиндексации может существенно различаться в подразделениях ВИНИТИ, осуществляющих обработку документов. По одним тематическим направлениям можно в большой мере доверять результатам автоиндексации и активно использовать их при формировании традиционных информационных продуктов и даже для создания новых продуктов и услуг. По другим тематическим направлениям результаты автоиндексации менее достоверны, поэтому не могут конкурировать с ручным индексированием (хотя могут использоваться в качестве чернового материала). Здесь для повышения достоверности автоиндексации необходимы серьёзные усилия по совершенствованию моделей принятия решений, самих классификационных схем и технологий обучения.

Интересным средством демонстрации и ручного контроля возможностей системы АТКТ является *web*-услуга "Тематическая классификация текстов"[2] Она позволяет сотрудникам Института самостоятельно оценивать работу автоклассификатора. Пример показан на рисунке.



Интерфейс Тематического классификатора текстов

[2] В создании этого продукта принимают участие Б.В. Крутиков − ведущий программист Управления информационных систем, и А.В. Ефимов − младший научный сотрудник Отделения обработки входного потока ВИНИТИ РАН.

Введя любой текст, пользователь может получить результаты его классификации − по отраслям науки и техники (отделы), по номенклатуре выпусков Реферативного журнала ВИНИТИ, по рубрикам ГРНТИ 2-го уровня. В качестве результатов анализа текста выдаются рубрики заданных классификационных схем со значениями их веса.

Пока эта услуга доступна только в Интранете ВИНИТИ. В будущем, при положительной оценке результатов тестирования и наличии достаточной вычислительной мощности, возможен доступ к этому сервису внешних пользователей через Интернет.

ВЫВОДЫ

В заключение отметим, что в ближайшем будущем вряд ли можно планировать отказ от ручного индексирования научных публикаций силами квалифицированных научных сотрудников, и автоматическое индексирование не заменит полностью интеллектуальную работу специалистов. Однако применение методов автоиндексации на определенных стадиях обработки документов представляется весьма перспективным – если не сказать неизбежным – в условиях нарастающей лавины документального потока и существующего запроса пользователей в оперативном информировании о поступлении документов по их тематическим профилям. Задача в том, чтобы при формировании информационных услуг находить адекватные механизмы сосуществования и взаимного дополнения людей и программ-роботов, выполняющих интеллектуальный анализ текстов.

Авторы настоящей статьи, не являясь профессиональными разработчиками собственно средств искусственного интеллекта, рассчитывают на совершенствование своей системы за счет применения новых инструментов обработки текстов, которые постоянно появляются в мире компьютерных технологий и становятся доступными для использования.

При любых подходах технология машинного обучения не может достичь высоких результатов без хороших целенаправленно сформированных обучающих выборок (подобно тому, как ученики не могут добиться успеха без учителей по своим предметам). Участие человека в обработке текстов необходимо, так как именно результаты ручной индексирования позволяют обучать и регулярно переобучать программу-робот. ВИНИТИ РАН обладает поистине огромным запасом "хорошо индексированных" документов, и этот запас постоянно пополняется и может использоваться для совершенствования средств автоиндексации.

СПИСОК ЛИТЕРАТУРЫ

1. Паттерсон Дж., Гибсон А. Глубокое обучение с точки зрения практика / пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2018. – 418 с.
2. Aghaebrahimian A., Cieliebak M. Hyperparameter Tuning for Deep Learning in Natural Language. – URL: CEUR-WS.org/Vol-2458/paper5.pdf (дата обращения: 05.05.2020)
3. A Framework for Evaluating Automatic Indexing or Classification in the Context of Retrieval [Основа для оценки автоматической индексации или классификации в контексте поиска] // Journal of the association for information science and technology – 2015. DOI: 10.1002/asi
4. Altınel B., Can Ganiz M. Semantic text classification: A survey of past and recent advances // Information Processing & Management. – 2018 November. – Vol. 54(6). – P. 1129-1153. DOI: 10.1016/j.ipm.2018.08.001
5. Golub K., Hagelbäck J., Ardö A. Automatic Classification Using DDC on the Swedish Union Catalogue. – URL: https://www.semanticscholar.org/, (дата обращения: 05.05.2020).
6. Arash Joorabchi, Abdulhussain E. Mahdi. Classification of scientific publications according to library controlled vocabularies: A new concept matching-based approach. – URL: https://www.semanticscholar.org/ (дата обращения: 05.05.2020). DOI: 10.1108/LHT-03-2013-0030
7. Some Thoughts on Preserving Functions of Library Catalogs [Размышления о сохранении функций библиотечных каталогов в сетевых средах] // Bulletin of the Association for Information Science and Technology – October/November 2016 – Vol. 43, Number 1
8. 12 years on - Is the NLM medical text indexer still useful and relevant? // Journal Biomed Semantics. – 2017 Feb 23/ – Vol. 8(1). – P. 8. – URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5324252/. DOI: 10.1186/s13326-017-0113-5.,
9. Создание базы данных ресурсов AgNIC с использованием полуавтоматической индексации материала // Journal of Agricultural & Food Information. – 2014. – № 15. – P. 159–179.
10. Romanov A., Lomotin K., Kozlova E. Application of Natural Language Processing Algorithms to the Task of Automatic Classification of Russian Scientific Texts // Data Science Journal. – 2019. – Vol. 18. – №. 1.
11. Козлова Е.С., Ломотин К.Е., Романов А.Ю. Применение алгоритмов обработки естественного языка: инструмент для автоматической классификации текста // Цифровые Трансформации и Глобальное Общество (DTGS-2018): материалы международной конференции. – Хам, Швейцария: Springer, 2018. – С. 310-323. DOI: 10.1007/978-3-030-02846-6_25.
12. Ломотин К.Е., Козлова Е.С., Романов А.Ю. Сравнительный анализ методов автоматической классификации для генерации кода УДК для научных статей // Информационные Инновационные Технологии: материалы международной научно-практической конференции. – М.: Ассоциация выпускников и сотрудников ВВИА имени профессора Н.Е. Жуковского, 2017. – С. 359-363.

13. Mikolov T. et al. Distributed representations of words and phrases and their compositionality // Advances in neural information processing systems. – 2013. – P. 3111-3119.
14. Информационный поиск // Википедия. [2018—2018]. Дата обновления: 29.06.2018. – URL: https://ru.wikipedia.org/?oldid=93657750 (дата обращения: 26.07.2019).
15. Precision and recall // Wikipedia, The Free Encyclopedia. – URL: https://en.wikipedia.org/w/index.php?title=Precision_and_recall&oldid=900147451 (дата обращения: 26.07.2019).
16. Documentation of scikit-learn 0.21.2. User Guide. 3.3. Model evaluation: quantifying the quality of predictions. – URL: https://scikit-learn.org/stable/modules/model_evaluation.html (дата обращения: 26.07.2019).
17. Коэффициент сходства // Википедия. [2019—2019]. Дата обновления: 08.01.2019. –URL: https://ru.wikipedia.org/?oldid=97352771 (дата обращения: 26.07.2019).
18. Коэффициент Сёренсена // Википедия. [2017—2017]. Дата обновления: 22.03.2017. –URL: https://ru.wikipedia.org/?oldid=84432232 (дата обращения: 26.07.2019).

Материал поступил в редакцию 27.04.20.

Сведения об авторах

ЕГОРОВ Виктор Серафимович – заместитель начальника Управления информационных систем ВИНИТИ РАН
e-mail: vs-egorov@viniti.ru

КОЗЛОВА Екатерина Сергеевна – программист Отдела программных систем, ВИНИТИ РАН
e-mail: hse.kozlovaes@gmail.com

ЛОМОТИН Константин Евгеньевич – программист Отдела программных систем ВИНИТИ РАН,
e-mail: ke.lomotin@gmail.com

ФЕДОРЕЦ Олег Владимирович – кандидат технических наук, начальник Отдела программных систем ВИНИТИ РАН
e-mail: ovf@viniti.ru

ФИЛИМОНОВ Алексей Викторович – главный специалист Отдела программных систем ВИНИТИ РАН
e-mail: afil@viniti.ru

ШАПКИН Александр Владимирович – кандидат технических наук начальник Управления информационных систем ВИНИТИ РАН
e-mail: ss@viniti.ru