УДК 004.056

В.А. Лохвицкий[1]
А.Г. Борозенец[2]

ПРОГРАММНЫЙ КОМПЛЕКС КЛАССИФИКАЦИИ НЕСТРУКТУРИРОВАННЫХ ДАННЫХ НА ОСНОВЕ МЕТОДА ОПОРНЫХ ВЕКТОРОВ

V.A. Lokhvitsky
A.G. Borozenets

SOFTWARE SYSTEM CLASSIFICATION OF UNSTRUCTURED DATA BASED ON SUPPORT VECTOR METHOD

Представленная работа посвящена разработке модели поиска информации, которая позволяет наилучшим образом производить процесс автоматизированной классификации неструктурированных данных.

Актуальность разработки модели поиска информации обусловлена необходимостью своевременного анализа больших потоков информации в различных управленческих структурах.

Значимость полученных результатов заключается в том, что разработанная модель поиска информации позволяет повысить оперативность поиска неструктурированных данных в открытых источниках.

Результаты могут быть использованы в организациях и предприятиях, имеющих дело с большими потоками информации.

Ключевые слова: *классификация, веб-ресурсы, векторная модель данных, неструктурированные данные, веб-контент, метод опорных векторов.*

This work is devoted to the development of an information retrieval model, which allows the best way to make the process automatic classification of unstructured data.

The urgency of developing a model of information retrieval is cansed due to the need for timely analysis of the large data flows at different administrative structures.

The significance of these results is that the developed model of information retrieval improves search efficiency of unstructured data in the public domain.

The results can be used in organizations and enterprises dealing with large flows of information.

Keywords: *classification, web-resource, vector data model, unstructured data, web-content, support vector method.*

Введение

В современных условиях многократно увеличились объемы информации, цикл ее изменения сократился с недель и суток до часов и минут. Возникла острая необходимость в постоянном комплексном мониторинге процессов, происходящих в мире.

Проблемная ситуация заключается в анализе больших объемов данных, когда аналитик или управленец не в состоянии вручную обработать большие массивы данных и принять решение. При этом необходимо каким-то образом представить исходную информацию в более компактном виде, с которой может справиться человеческий мозг за приемлемое время.

1. Цель работы

Цель работы заключается в повышении оперативности поиска информации заданной тематической направленности в открытых источниках.

2. Анализ путей достижения цели

На сегодняшний день интенсивно развивается направление, связанное с интеллектуализацией методов обработки и анализа данных. Интеллектуальные системы анализа данных призваны минимизировать усилия лица, принимающего решения, в процессе анализа данных, а также в настройке алгоритмов анализа.

[1] Кандидат технических наук, доцент, Военно-космическая академия им. А.Ф. Можайского.

[2] Курсант, Военно-космическая академия им. А.Ф. Можайского.

Одной из технологий данного направления является Data Mining. Основная особенность Data Mining – это сочетание широкого математического инструментария (от классического статистического анализа до новых кибернетических методов) и последних достижений в сфере информационных технологий.

Акцентируя внимание на работу с открытыми источниками, которыми, в основном, являются веб-ресурсы, целесообразно будет ограничиться использованием лишь определенным направлением технологии Data Mining.

Рассмотрим все направления, касающиеся анализа веб-ресурсов.

Извлечение веб-контента включает в себя методы извлечения полезной информации из веб-ресурсов, таких, как содержание, данные, документы и др. Актуальность данного направления возрастает в связи с тем, что в настоящее время прослеживается тенденция предоставления компаниями (организациям) доступа к своим ресурсам. Это относится не только к статической информации, представленной в виде HTML-страниц, но также к данным, хранящимся в БД компаний, и другим ресурсам.

Извлечение веб-структур предполагает построение модели, отображающей взаимосвязи между веб-страницами. Модель основывается на топологии гиперссылок с описанием или без описания этих ссылок. Такая модель может использовать категоризацию веб-страниц и быть полезна для генерации информации об отношении и подобности между веб-сайтами. Данное направление может быть использовано для распознавания авторских сайтов и обзорных сайтов по темам.

Исследование использования веб-ресурсов. Здесь анализируется информация, генерируемая в процессе пользовательских сессий (взаимодействия пользователя с веб-ресурсами) и поведения пользователей. В отличие от первых двух направлений, которые работают с первичной информацией (веб-ресурсами), исследование использования веб работает со вторичной информацией, порождаемой как результат взаимодействия пользователей с веб-ресурсами. К таким источникам информации относятся протоколы доступа веб-серверов, протоколы прокси-серверов, протоколы браузеров, пользовательские профили, регистрационные данные, пользовательские запросы.

Из перечисленных направлений в большей мере поиску информации соответствует первое (извлечение веб-контента). Для достижения поставленной цели на основе вышеописанной технологии будет разработан программный комплекс, который позволит автоматизировать процесс поиска информации, тем самым повысив оперативность ее поиска.

3. Постановка задачи

3.1. Содержательная постановка задачи

В целом процесс решения поставленной в работе задачи можно представить как последовательность нескольких шагов.

1. Поиск информации. На первом шаге необходимо идентифицировать, какие документы должны быть подвергнуты анализу, и обеспечить их доступность. Здесь задается перечень анализируемых веб-ресурсов и период публикаций.

2. Предварительная обработка документов. На этом шаге выполняются преобразования документов для представления их в более удобном для обработки виде. Целью таких преобразований является удаление лишних слов и придание тексту более строгой формы.

3. Извлечение информации. Извлечение информации из выбранных документов предполагает выделение в них ключевых понятий, над которыми в дальнейшем будет выполняться анализ. На данном этапе осуществляется построение векторной модели текста.

4. Применение метода классификации. На данном шаге решаются две задачи:

– задача обучения системы (построение модели классификации);

– задача классификации.

5. Интерпретация результатов. Последний шаг в процессе анализа информации предполагает интерпретацию полученных результатов.

3.2. Математическая постановка задачи классификации

Дано:

Множество категорий $C = \{c_j\},\ j = 1, \overline{1, |C|}$ и обучающее множество документов $\Omega \subset D$, где $D = \{d_i\},\ i = 1, |D|$ – полное множество документов – формируют эксперты. Применяется алгоритм классификации с учителем – алгоритм категоризации – использует обучающее множество Ω, чтобы построить классификатор $\Phi: D \times C \to \{\text{истина, ложь}\}$, обеспечивающий высокую точность на всем множестве документов D, используя предположение, что обучающие и новые данные похожи. Обычно множество документов Ω делят на две части: одна часть – данные для обучения алгоритма, вторая – тестовые данные для оценки качества полученного классификатора.

4. Представление неструктурированных (веб-) данных

4.1. Предварительная обработка

Особенностью веб-ресурсов является разно-

родность представленной информации: текстовые файлы, изображение, звук, видео, метаданные, а также гиперссылки. Получаемые из веб данные относят к следующим типам:

1. Неструктурированные данные (текст, медиаинформация).

2. Слабоструктурированные данные (HTML, XML и др.).

В данной работе предполагается анализ только неструктурированных данных, а именно – текстовых.

Одной из главных проблем анализа текстов является большое количество слов в документе. Если каждое из этих слов подвергать анализу, то время поиска новых знаний резко возрастет и вряд ли будет удовлетворять требованиям пользователей. Таким образом, удаление неинформативных слов, а также приведение близких по смыслу слов к единой форме значительно сокращают время анализа текстов. В реализации программного комплекса применены следующие методы:

– удаление стоп-слов. Стоп-словами называются слова, которые являются вспомогательными и несут мало информации о содержании документа;

– стемминг – морфологический поиск. Он заключается в преобразовании каждого слова в его нормальную форму. Нормальная форма исключает склонение слова, множественную форму, особенности устной речи.

4.2. Векторное представление текста

Тексты не могут напрямую интерпретироваться алгоритмами классификации, так как они оперируют числовыми данными, а текст – это всего лишь последовательность символов. Поэтому требуется процесс индексации, в результате которого получается компактное представление документа d_j, удобное для дальнейшей обработки. Выбор представления текста зависит от того, что считается несущими смысл частями текста и какие правила обработки естественного текста допустимо применять к этим частям.

Все алгоритмы классификации машинного обучения используют представление документа d_j в виде вектора весов термов $\vec{d}_j = \langle w_{j,1}, \ldots, w_{j,|T|} \rangle$, где T – множество всех термов, которые учитываются в тексте; $w_{j,k,}$ – вес k-го терма в документе d_j показывает, насколько большую смысловую нагрузку несёт k-й терм в документе.

5. Метод опорных векторов

Метод опорных векторов строит классифицирующую функцию:

$$f(\vec{x}) = sign\left(\overrightarrow{w^T}\vec{x} + b\right), \qquad (1)$$

где $\vec{w}$ – вектор нормали к разделяющей поверхности, или вектор весов; b – параметр сдвига.

Те объекты, для которых $f(x) = 1$, попадают в один класс, а объекты с $f(x) = -1$ – в другой. Выбор именно такой функции не случаен: любая гиперплоскость может быть задана в виде $\langle w, x_i \rangle + b = 0$ для некоторых w и b.

Далее, мы хотим выбрать такие w и b, которые максимизируют расстояние до каждого класса. Можно подсчитать, что данное расстояние равно $\frac{1}{\|w\|}$. Проблема нахождения максимума $\frac{1}{\|w\|}$ эквивалентна проблеме нахождения минимума $\|w\|^2$. Запишем всё это в виде задачи оптимизации:

$$\begin{cases} \arg\ \min_{w,b} \|w\|^2 \\ y_i\left(\langle w, x_i \rangle + b\right) \geq 1,\ i = 1,\ \ldots,\ m, \end{cases} \qquad (2)$$

которая является стандартной задачей квадратичного программирования и решается с помощью множителей Лагранжа.

6. Программная реализация

Разработанный программный комплекс позволяет пользователю классифицировать текстовые документы по заданной тематике. Последовательность работы комплекса такова:

1. Загружается текстовый документ с необходимой тематикой.

2. Документ проходит предварительную обработку.

3. Выводятся результаты в виде векторной модели текста.

4. Загружается любой другой документ.

5. Производятся аналогичные предыдущим операции.

6. Алгоритм опорных векторов определяет принадлежность второго документа к тематике.

На рис. 1 представлен этап загрузки текстового файла для обучения классификатора.

Дальнейшие этапы работы комплекса показаны на рис. 2. В левой части формы отображена векторная модель классификатора (обучающей выборки). В середине отображены результаты предварительной обработки текста. И в правой части представлена векторная модель тестируемого документа.

В данном случае документ был определен к соответствующей заданной тематике. Точность классификации зависит от качества подобранной обучающей выборки.

Заключение

В настоящей работе представлены результаты разработки программного комплекса классификации неструктурированных данных на основе метода опорных векторов.

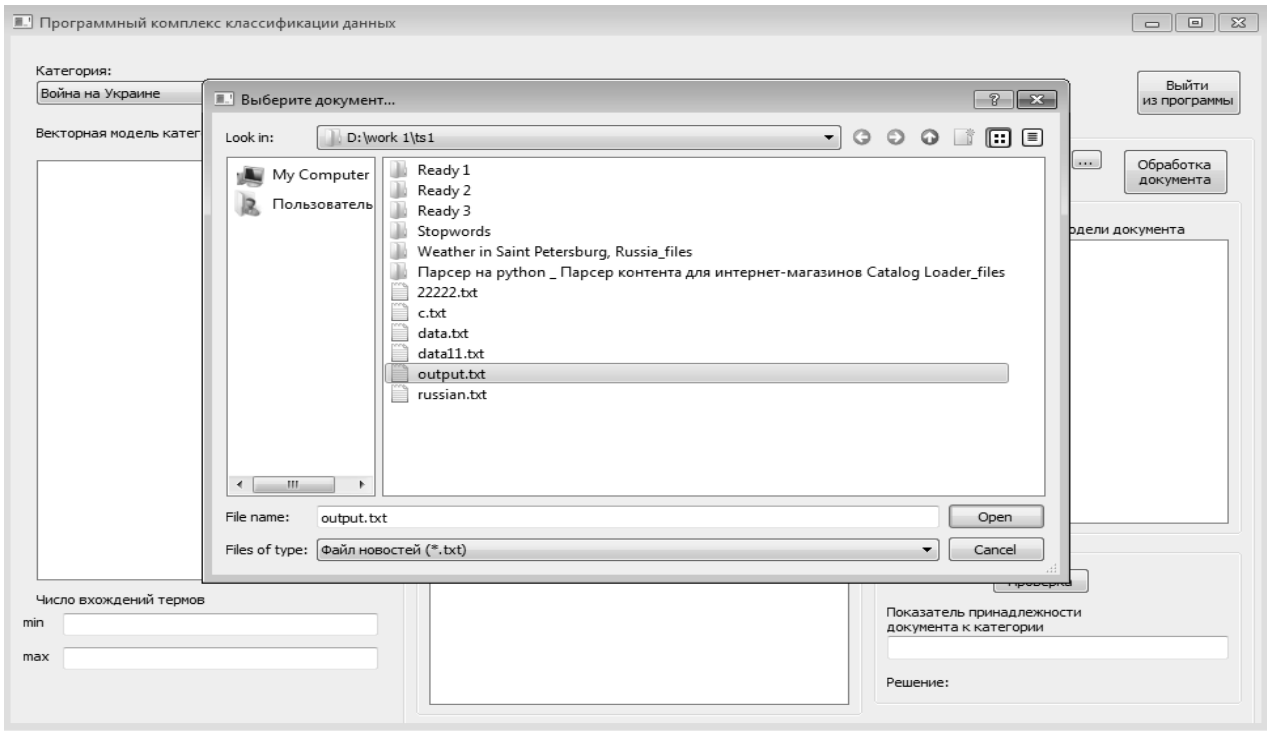

Рис. 1. Загрузка текстового файла

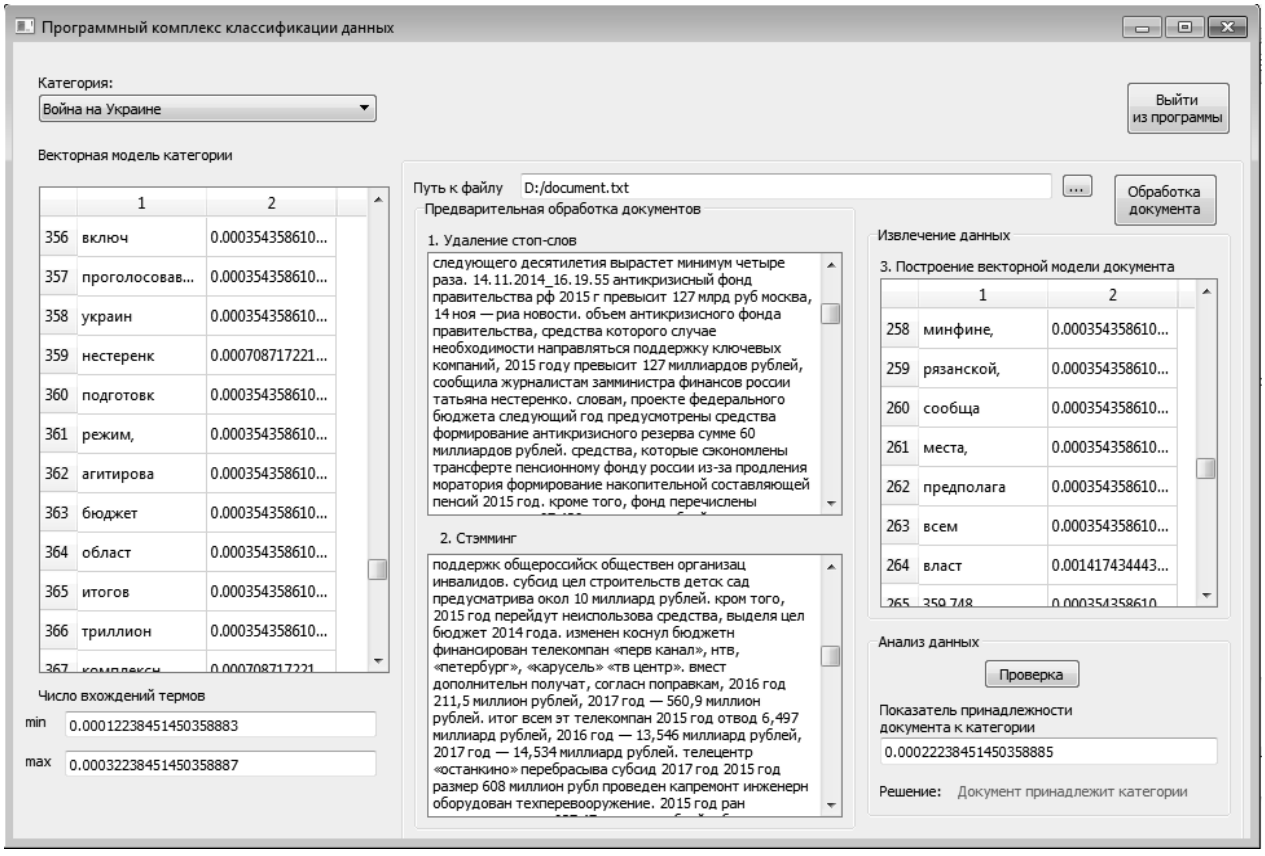

Рис. 2. Отображение результатов классификации

Разработанный программный комплекс позволяет автоматизировать процесс классификации текстовых данных по тематикам. Применение метода опорных векторов с векторной моделью текста обеспечивает точность отнесения документа к определенной тематике.

Использование данного комплекса позволит повысить оперативность поиска информации из открытых источников.

Технологии, используемые при разработке программного модуля визуализации, могут быть использованы при создании новых и совершенствовании существующих систем классификации неструктурированных данных.

Литература

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика / В.И. Большакова, Э.С. Клышинский, Д.В. Ландэ, А.А. Носков, О.В. Пескова, Е.В. Ягунова : учебное пособие. – М. : МИЭМ, 2011. – 272 с.

2. Анализ данных и процессов / А.А. Барсегян, М.С. Куприянов, И.И. Холод, М.Д. Тесс, С.И. Елизаров : учебное пособие, 3-е изд., перераб. и доп. – СПб., БХВ-Петербург, 2009. – 512 с.

3. Лохвицкий В.А. Подход к построению системы автоматизированной интеграции информации в базу данных для её своевременной актуализации / В.А. Лохвицкий, С.В. Калиниченко, А.А. Нечай // Мир современной науки. – 2014. – № 2 (24). – С. 8–12.

4. Нечай А.А. Выявление недекларированных возможностей аппаратно-программного обеспечения / А.А. Нечай // Экономика и социум. – 2014. –№ 1–2 (10). – С. 457–460.

5. Нечай А.А. Специфика проявления уязвимостей в автоматизированных системах управления критически важными объектами / А.А. Нечай, П.Е. Котиков // Современные тенденции в образовании и науке : сборник научных трудов по материалам Международной научно-практической конференции : в 14 ч. – Тамбов, 2014. – С. 96–97.

6. Нечай А.А. Выбор и обоснование показателей эффективности решения задачи распределения объектов по средствам поражения / А.А. Нечай, С.В. Матвеев, В.М. Сафонов // Мир современной науки. – 2014. – № 2 (24). – С. 13–16.

7. Вепрев, С.Б. Скрытый метод выявления утечек инсайдерской информации / С.Б. Вепрев, П.И. Гончаров // Вестник Российского нового университета. – 2014. – № 4. – С. 152–155.