УДК 004.942

Б. Я. Советов, Т. М. Татарникова, А. И. Яшин
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина)

Использование технологии TextMining для выявления скрытых угроз в сообщениях, распространяемых по социальным сетям

Предложено решение задачи анализа текста с применением технологии TextMining для обнаружения угроз, скрытых в сообщениях, которыми обмениваются пользователи в социальных сетях. Обсуждаются возможности технологии TextMining в задачах выявления знаний в неструктурированных информационных массивах. Последовательность анализа текста представлена в виде методики. Раскрыто содержание этапов методики и рассмотрены основные приемы, используемые на каждом этапе. Выделены способы вычисления функций взвешивания, на основе которых формируется список ключевых слов и словосочетаний. Рассмотрены способы построения семантических сетей на основе множества ключевых слов. Для автоматизации анализа текста разработан программный комплекс, реализующий технологию TextMining. В функции программного комплекса входит выявление ключевых слов, связей, эмоционального портрета пользователя, что позволяет перейти от данных к их смыслу и сделать выводы об информационной безопасности текста.

TextMining, семантическая сеть, ключевые слова, частота появления слов, смысл текста, тема, классификация текста, словарь, тональность текста, программный комплекс

Интеллектуальные информационные технологии обработки текстовой информации больших объемов разнородных данных призваны значительно повысить эффективность решения задач в этих сферах. Одной из таких технологий является TextMining.

TextMining – это технология анализа текстов, позволяющая обрабатывать большие объемы неструктурированной информации, извлекать знания и высококачественную информацию из текстовых массивов. Инструменты TextMining позволяют автоматически анализировать большие объемы информации с целью поиска тенденций, шаблонов и взаимосвязей, способных помочь в принятии стратегических решений [1].

На данный момент можно выделить 4 основных вида приложений технологий TextMining [2]:

1. Задача классификации, которая по своей сути является задачей распознавания, т. е. отнесения текстового документа к той или иной заранее предопределенной категории. Интеллектуальные механизмы TextMining позволяют оптимизировать процесс классификации при обработке больших объемов разнородных текстовых документов. Классификация применяется, например, в таких задачах, как группировка документов в intranet-сетях и на web-сайтах, размещение документов в определенные папки, сортировка сообщений электронной почты, избирательное распространение новостей подписчикам.

2. Задача кластеризации, заключающаяся в выделении групп документов с близкими свойствами. Технология TextMining позволяет найти признаки и разделить документы по группам в соответствии с этими признаками. Кластеризация, как правило, предшествует задаче классификации. Различают 2 основных типа кластеризации – иерархическую и бинарную, первая заключается в построении дерева кластеров, в каждом из которых размещается небольшая группа документов, а вторая обеспечивает группировку и просмотр документальных кластеров по ссылкам подобия. Кластеризация применяется при реферировании больших документальных массивов, определении взаимосвязанных групп документов, для упрощения процесса просмотра при поиске необходимой информации, нахождении уникальных документов из коллекции, выявлении дубликатов или очень близких по содержанию документов.

3. Прогнозирование, которое состоит в том, чтобы предсказать по значениям одних признаков документа значения остальных. Одним из примеров прогнозирования в TextMining может служить задача автоматической генерации текста. Также следует отметить такое актуальное направление TextMining, как анализ тональности текста, что позволяет решать задачи прогнозирования в различных прикладных областях. Например, в маркетинге, анализируя Твиттер можно прогнозировать спрос на рынке товаров и услуг; в политологи, собирая данные из блогов можно предотвратить развитие нежелательных проявлений, дестабилизирующих нормальную обстановку; в социологии, собирая данные из социальных сетей можно прогнозировать отношение пользователей к той или иной ситуации.

4. Нахождение исключений, т. е. поиск документов, которые своими характеристиками сильно выделяются из общей массы. Для этого сначала выясняются средние параметры документов, а потом исследуются те из них, параметры которых наиболее сильно отличаются от средних значений. Поиск исключений широко применяется, например, в работе спецслужб. Подобный анализ часто проводится после классификации, для того чтобы выяснить, насколько последняя была точна.

Этапы решения задачи средствами TextMining. Основные стадии решения задач средствами TextMining приведены на рис. 1.

На этапе поиска информации и определения исходных данных необходимо определить, какие документы должны быть подвергнуты анализу, и обеспечить их доступность.

Предварительная обработка документов подразумевает необходимые преобразования документов для представления их в виде, с которым работают методы TextMining. Целью таких преобразований является удаление лишних слов и придание тексту более строгой формы.

Извлечение информации предполагает выделение в анализируемом документе ключевых понятий. Ключевые понятия – это наиболее часто встречающиеся слова и словосочетания в тексте, которые, собственно, и определяют его тематику [3].

На этапе применения методов TextMining извлекаются шаблоны и отношения, скрытые в тексте. Шаблон может быть построен в виде графа (семантического дерева или семантической сети), корреляционной матрицы ключевых слов, фреймовой модели или в другом визуализированном виде. Визуализация предполагает графическое представление извлеченных ключевых понятий *a*, *b*, *c* и т. д. и связей между ними, что помогает быстро идентифицировать тематику текста. Веса связей показывают расстояние между ключевыми понятиями.

Последний шаг в процессе анализа текста предполагает интерпретацию полученных результатов. На этом этапе к работе подключается лицо, принимающее решение, – аналитик по работе с текстом.

Как видно из содержания этапов TextMining, в результате применения комплекса техник к анализируемому тексту может быть получен шаблон, способствующий извлечению семантических связей между отдельными словами, причем этот процесс должен выполняться автоматически без участия человека, а понимание смысла текста происходит с участием человека.

Семантические сети определяются как граф общего вида, в котором можно выделить множество вершин и ребер. Каждая вершина графа представляет некоторый объект, а дуга – отношение между парой объектов. В качестве таких объ-

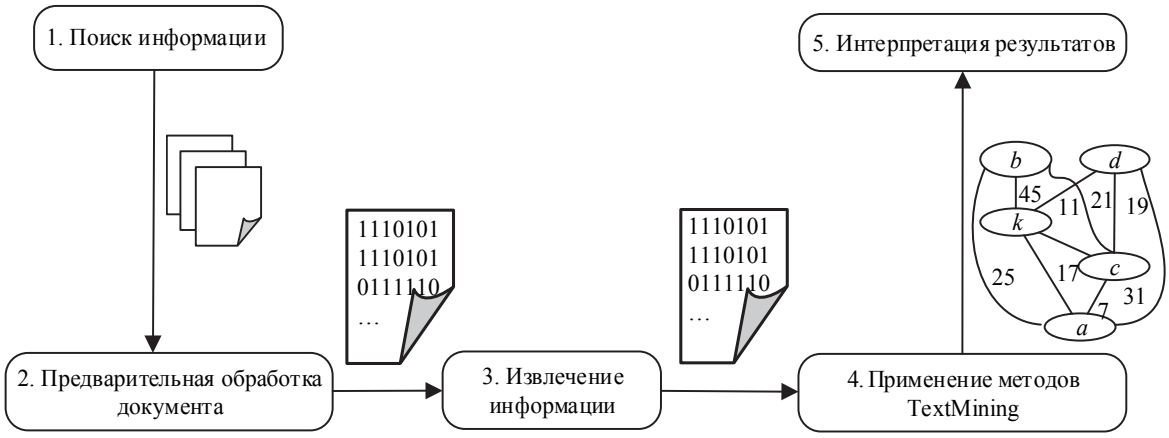

Рис. 1

ектов могут выступать понятия, события, процессы и т. д.; свойства объектов также представляются вершинами сети и служат для описания классов объектов. Имена вершин и дуг совпадают с именами соответствующих объектов и их отношений, используемых в профессиональном языке проблемной области. Метка и направление дуги конкретизируют семантику. Метки вершин не несут семантической нагрузки, а используются как справочная информация [4].

Дуга и связываемые ею вершины образуют подграф, являющийся минимальной информационной единицей в системе анализа текста. Более сложные подграфы сети отражают и более сложные факты (утверждения).

Объекты могут быть трех основных типов: обобщенные, конкретные и агрегатные. Обобщенный объект на самом деле представляет собой целый класс объектов (более низкого уровня) предметной области, а конкретный – некоторым образом выделенную сущность из класса. Под агрегатным понимается объект предметной области, составленный из других объектов. В качестве агрегатного может выступать как обобщенный, так и конкретный объект.

Между двумя объектами могут существовать различного типа отношения. В качестве наиболее распространенных (базовых) можно отметить следующие отношения между объектами:

– принадлежит (объект принадлежит данному классу);

– обладает (объект обладает некоторым свойством);

– значение (определяет значения свойств объекта);

– следствие (отражает причинно-следственные связи: свойство является следствием некоторой причины).

В последние 5 лет открылся целый спектр отраслей, в которых возможности TextMining только начинают использоваться. К их числу относятся корпоративная бизнес-аналитика, мониторинг социальных медиа и деловая разведка [5], [6].

В статье рассматривается возможность применения технологии TextMining в вопросах предотвращения потенциальных угроз, которые могут быть скрыты в сообщениях. Проблема безопасности в мессенджерах, социальных сетях, форумах и электронных письмах является актуальной, анализ этих данных позволит пресечь нежелательную утечку информации или планирование враждебных действий по отношению к кому-либо или чему-либо [7], [8].

Содержание этапов методики анализа текста. Методика анализа текста включает следующую последовательность операций:

– формирование файла;

– токенизация по словам;

– фильтрация от стоп-слов;

– нормализация;

– преобразование регистра слов;

– построение семантической сети;

– классификация;

– оценка тональности текста.

Рассмотрим приемы, используемые на этапе предварительной обработки, доступные в языке программирования Python 3.6, на котором реализован программный комплекс.

Одним из основных приемов является токенизация текста, т. е. разбиение текстового документа на отдельные слова, получение так называемого мешка слов. Результаты данного разбиения называются токенами. Также можно получить биграммы и триграммы слов, если их необходимо использовать для дальнейшего анализа текста.

При подаче на вход программного комплекса следующего сообщения: «*Социальные сети могут быть использованы и для организации утечек важной для компании информации, а также для подрыва ее репутации*. …» токенизация выдает следующий список слов: ['Социальные', 'сети', 'могут', 'быть', 'использованы', 'и', 'для', 'организации', 'утечек', 'важной', 'для', 'компании', 'а', 'также', 'для', 'подрыва', 'ее', 'репутации', …].

После токенизации, как правило, следует фильтрация стоп-слов. К стоп-словам относятся такие слова, которые не содержат в себе никакого смысла, например союзы, предлоги, артикли, междометия, частицы и т. п. Список стоп-слов составляется заранее в зависимости от языка обрабатываемого текста. Также возможно пользоваться библиотеками типа Nltk, Numpy, однако проблема заключается в том, что они могут не поддерживать обработку текста на русском языке.

После фильтрации стоп-слов получим следующий список: ['Социальные', 'сети', 'использованы', 'организации', 'утечек', 'важной', 'компании', 'подрыва', 'репутации', …].

Следующим шагом является нормализация слов. Все слова в текстовом документе приводятся к нормальной форме – именительный падеж,

единственное число, без особенностей устной речи. Естественно, что семантика предложений и словосочетаний (биграмм и триграмм слов) будет нарушена, но это позволит точнее определить частоту появления однокоренных слов. Наиболее известным алгоритмом нормализации слов русского языка является Snowball, основная идея которого заключается в нахождении однокоренных слов и отсечении у них окончаний, суффиксов и т. п.

После нормализации получим следующий список слов: ['Социальный', 'сеть', 'использовать', 'организация', 'утечка', 'важный', 'компания', 'информация', 'подрыв', 'репутация', …].

Далее происходит преобразование регистра слов, т. е. преобразование символов слов к одному регистру (верхнему или нижнему), что также позволит точнее определить частоту появления слов в анализируемом тексте. Например, слова «document», «Document», «DOCUMENT» без преобразования их к одному регистру будут считаться разными словами, т. е. несущими разный смысл.

Выполнение преобразования регистра дает следующий список слов: ['социальный', 'сеть', 'использовать', 'организация', 'утечка', 'важный', 'компания', 'информация', 'подрыв', 'репутация', …].

Для выполнения дальнейшего этапа – классификации – необходимо найти ключевые слова, т. е. термины, имеющие наибольший вес по частоте их появления в анализируемом тексте [9]. Именно ключевые слова определяют принадлежность документа к той или иной категории.

Выделяют следующие способы вычисления функций взвешивания:

– бинарная частота (0/1 бит-вектор: 1 – термин встречается в тексте, 0 – термин не встречается);

– частота слова TF (term frequency), которая может принимать абсолютное значение (рис. 2): TF = N, где N – число появлений слова в документе или относительное значение: TF = N/Len. Здесь Len – длина документа (количество слов в документе). Для нелинейного преобразования TF существуют и другие эвристики, например TF = log N;

– мера TF–IDF (term frequency – inverse document frequency), где численное значение IDF для каждого термина определяется как IDF = = log [DN/TN]. Здесь DN – общее количество документов; TN – количество документов, в которых содержится термин (рис. 3).

Для терминов, встречающихся в большом числе документов, IDF стремится к нулю, что позволяет выделить наиболее значимые термины.

Значение коэффициента TF–IDF определяется как произведение (TF·IDF), где мера TF выступает в качестве повышающего сомножителя, а мера IDF – в качестве понижающего.

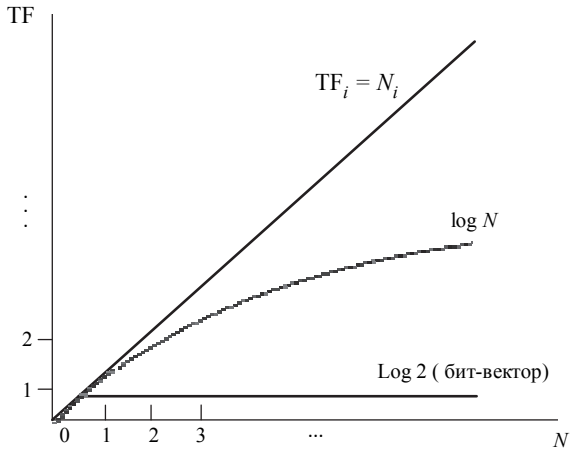

Рис. 2

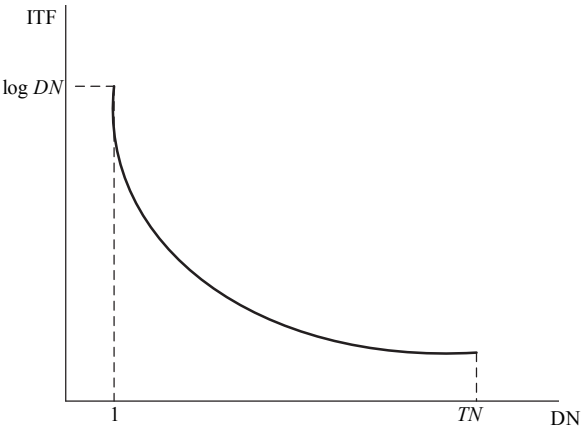

Рис. 3

Результат функции взвешивания в программном комплексе визуализируется в виде графика (рис. 4) и выводится в виде списка относительных значений TF.

Связи между ключевыми словами визуализируются в виде семантической сети. Эти оценки позволят сравнить относительный вклад различных понятий и их связей в семантику текста, выявить более или менее подробно проработанную в тексте тематику, задать способ сортировки информации, и наконец, взглянуть на весь текстовый материал по пластам – смысловым срезам различной глубины.

Примеры семантических сетей, построенных по ключевым словам анализируемого сообщения, приведены на рис. 5.

Таким образом, выделяя в наборе текстовых документов значения (TF·IDF) для наиболее значимых терминов в рамках заданных категорий

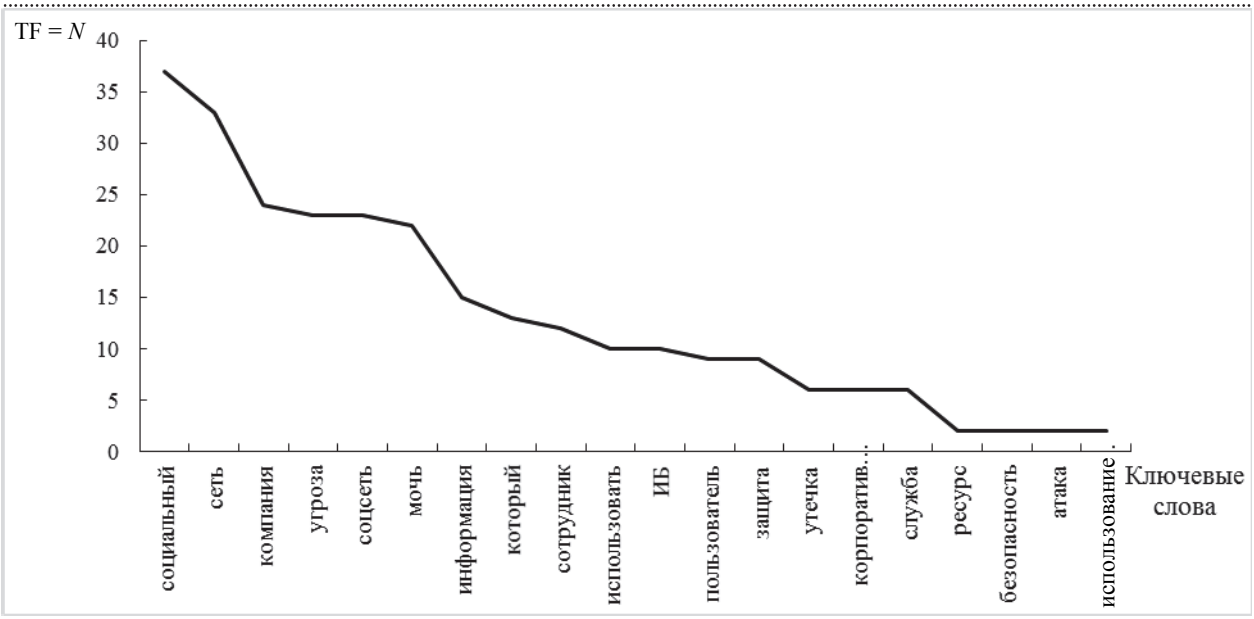

Рис. 4

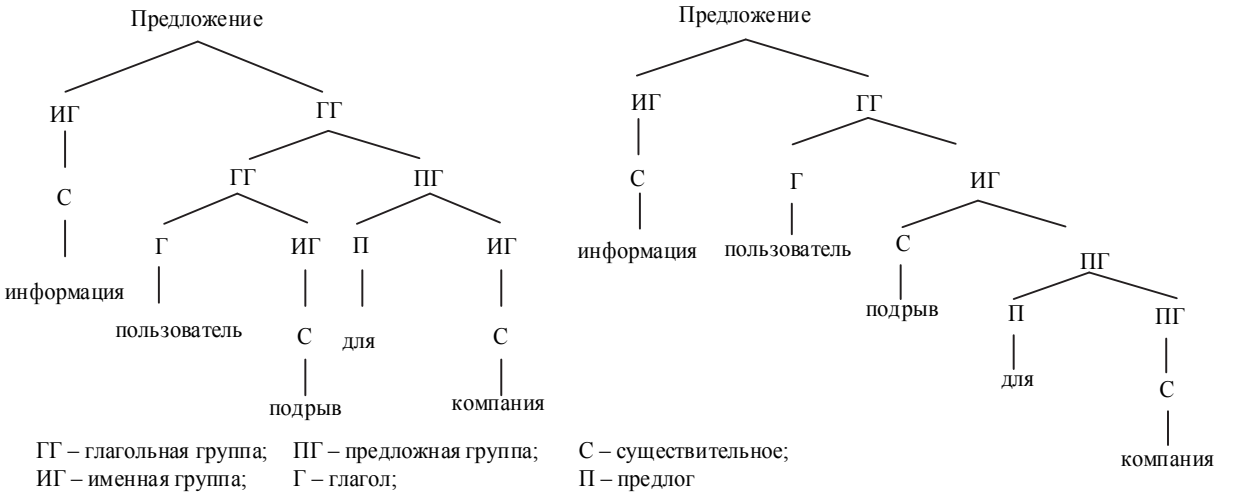

Рис. 5

документов, можно сформировать обучающую выборку для построения классификатора.

Классификация текстов может осуществляться двумя способами [10]: автоматически на основе заданного экспертом набора правил и автоматически на основе методов машинного обучения. В последнем случае набор правил принятия решений по классификации текстовых документов формируется автоматически на основе обучения классификатора на данных из обучающей выборки, представляющих собой набор эталонных образов из всех классов текстовых документов.

Поскольку в постановке задачи рассматривается бинарная классификация: класс 1 – сообщение содержит информацию о каких-либо угрозах; класс 2 – все остальные сообщения, то для определения класса сообщения были использованы такие средства языка Python, как регулярные выражения, в частности сравнение строк с шаблоном и определение класса для скомпилированного регулярного выражения в результате сравнения.

Анализируемое сообщение отнесено к классу 1.

При оценке тональности текста необходимо иметь словари слов, составленных специалистами-психологами, наиболее адекватно отражающих настроение автора сообщения, его физическое и когнитивное состояние, особенность самого сообщения. В таблице приведены примеры слов, соответствующие некоторым эмоциональным меткам.

Эмоциональная метка	Пример
Эмоция	Бешенство, бояться
Настроение	Враждебность, любезный
Состояние (физическое)	Уставший, бодрость
Состояние (когнитивное)	Воодушевленный, потрясение
Чувство	Тепло, чувствовать
Отношение	Толерантность, защита
Особенность	Агрессивность, конкурирующий

Сообщение, которое рассматривается в качестве примера, было определено как сообщение, не содержащее эмоциональной метки, что свидетельствует о том, что этот текст, возможно, не несет потенциальной угрозы. Для полной уверенности необходимо мнение эксперта.

Описание предлагаемого программного комплекса. Предлагается следующий программный комплекс, реализующий технологию TextMining для вскрытия потенциальных угроз в анализируемом тексте, представляющий собой прототип экспертной системы.

Программный комплекс состоит из следующих модулей (рис. 6):

– решателя (интерпретатора);

– рабочей памяти, называемой также базой данных (БД);

– базы знаний (БЗ);

– компонентов приобретения знаний;

– объяснительного компонента;

– компонента визуализации результата.

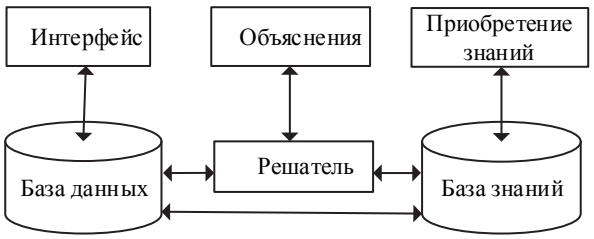

Рис. 6

База данных (рабочая память) предназначена для хранения исходных и промежуточных данных задачи, которая решается в текущий момент времени. В решаемой задаче рабочая память хранит анализируемый текст, массивы ключевых слов и словосочетаний, различные метрики текста, например расстояния между словами, словарь синонимов и выявленные шаблоны.

База знаний предназначена для хранения долгосрочных данных, описывающих свойства проблемной области и правил, описывающих преобразования данных. В рамках решаемой задачи база знаний содержит все процедуры предварительной обработки текста и оценки смыслового веса ключевых понятий.

Решатель, используя исходные данные из рабочей памяти и знания, формирует правила формирования семантической сети. Происходит этот процесс следующим образом. Каждый элемент сети (ключевое слово) характеризуется числовой оценкой – так называемым смысловым весом. Смысловой вес определяется исходя из частотного анализа текста. Пороговое значение повторяемости слова в общем объеме текста, позволяющее отнести или не отнести слово к ключевому понятию, выбирается экспертом исходя из тематики и объема анализируемого текста. От количества ключевых понятий будет зависеть ширина построенной семантической сети.

Компонент приобретения знаний необходим для наполнения экспертной системы новыми знаниями. Это могут быть новые словарные слова, новые алгоритмы обработки текста и др.

Объяснительный компонент позволяет увидеть промежуточные результаты, что объясняет, как система получила решение задачи и какие знания она при этом использовала. Наличие этого компонента повышает доверие к полученному результату со стороны лица, принимающего решение.

Интерфейс позволяет реализовать удобный режим работы с экспертной системой за счет удобного ввода исходных данных, визуализации результатов и т. п.

В статье описаны возможности технологии TextMining для автоматического анализа текста с целью выявления потенциальных угроз, которые могут быть скрыты в сообщениях, передаваемых в мессенджерах, социальных сетях, форумах и электронных письмах. Анализ таких сообщений позволит пресечь нежелательную утечку информации или планирование враждебных действий по отношению к кому-либо или чему-либо.

Предложен программный комплекс, функциональные возможности которого позволяют выделить ключевые моменты текста, сущности, связи, составить эмоциональный портрет пользователя, перейдя таким образом от данных к их смыслу и сделав выводы об информационной безопасности текста.

Достоверность выявления скрытых шаблонов в тексте проверена на открытых источниках в Интернете – сайтах и страницах пользователей социальных сетей.

СПИСОК ЛИТЕРАТУРЫ

1. Кутукова Е. С. Технология TEXTMINING // Науч. труды SWorld. 2013. № 4. С. 33–36.

2. Анализ данных и процессов / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. СПб.: БХВ-Петербург, 2009. 512 с.

3. Шереметьева С. О., Осминин П. Г. Методы и модели автоматического извлечения ключевых слов // Вестн. ЮУрГУ. Сер. «Лингвистика». 2015. Т. 12, № 1. С. 76–81.

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

5. Татарникова Т. М. Защищенные корпоративные сети: задачи по защите информации / РГГМУ. СПб., 2012. 113 с.

6. Aggarwal C. C., Zhai C. Mining Text Data. Springer, 2012. 527 p.

7. Almeida T. A., Yamakami A. Advances in spam filtering techniques // Computational Intelligence for Privacy and Security Studies in Computational Intelligence. 2012. Vol. 394. P.199–214.

8. Berry M. W., Browne M. E-mail surveillance using nonnegative matrix factorization // Computational & Mathematical Organization Theory. 2005. Vol. 11, № 3. P. 249–264.

9. Татарникова Т. М. Анализ данных / СПбЭУ. СПб., 2018. 82 с.

10. Горковенко Д. К. Применение методов text mining для классификации информации, распространяемой в социальных сетях // Молодой ученый. 2016. № 4. С. 66–72.

B. Ya. Sovetov, T. M. Tatarnikova, A. I. Yashin
Saint Petersburg Electrotechnical University «LETI»

USE OF TECHNOLOGY TEXTMINING FOR IDENTIFYING HIDDEN THREATS IN COMMUNICATIONS DISTRIBUTED BY SOCIAL NETWORKS

A solution to the problem of text analysis using TextMining technology to detect threats hidden in messages exchanged between users in social networks has been proposed. The possibilities of TextMining technology in the tasks of knowledge detection in unstructured information arrays are discussed. The sequence of text analysis is presented in the form of a methodology. The content of the stages of the methodology is disclosed and the main techniques used at each stage are considered. The methods of calculating the weighting functions are selected, on the basis of which a list of keywords and phrases is formed. Ways of building semantic networks based on a set of keywords are considered. To automate text analysis, a software package that implements TextMining technology has been developed. The functions of the software package include the identification of keywords, relationships, and the emotional portrait of the user, which allows you to move from data to their meaning and draw conclusions about the information security of the text.

TextMining, semantic network, keywords, word occurrence frequency, text meaning, theme, text classification, dictionary text tonality, software package

УДК 004.942

Е. Е. Котова, А. С. Писарев
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина)

Задача классификации учащихся с использованием методов интеллектуального анализа данных

Технологически поддерживаемые учебные среды генерируют большое количество данных, которые могут быть собраны и проанализированы с помощью релевантных алгоритмов. Функции аналитики обучения необходимы для планирования и внесения изменений в организацию процессов обучения, обеспечения адаптивных рекомендаций и персонализированного анализа учебной деятельности. Существует несколько различных методов классификации, используемых в обнаружении знаний и добыче данных (Knowledge Discovery and Data Mining). Каждый метод или методика имеет свои преимущества. В статье применяются методы анализа данных к задаче классификации обучающихся. Один из вопросов – определение признаков дифференциации обучающихся. Предложены признаки дифференциации, характеризующие индивидуальные параметры познавательной сферы и составляющие модель когнитивно-стилевого потенциала обучающихся. Методы интегрированы в веб-среду интеллектуальной поддержки процессов обучения. Анализируются результаты, полученные при помощи нескольких классификаторов. Классификация по методу, предложенному в статье, с применением признаков когнитивно-стилевого потенциала дает более точные результаты по сравнению с полученными в других исследованиях и опубликованными в доступных источниках.

Анализ данных, классификация обучающихся, когнитивно-стилевой потенциал, веб-среда интеллектуальной поддержки

Разработка методов использования данных, получаемых из образовательного контекста, входит в активно развивающуюся область междисциплинарных исследований Educational Data Mining (EDM).