Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей (05.13.11)

УДК 004.89
DOI: 10.24160/1993-6982-2020-5-132-139

Решение задачи классификации сообщений в системах голосового взаимодействия

И.Е. Куриленко, И.Е. Никонов

Рассмотрен метод решения задачи классификации коротко-текстовых сообщений, представляющих собой фразы клиентов при разговоре на телефонной линии организаций. Для решения поставленной задачи разработан классификатор, основанный на комбинации двух методов — описания предметной области в виде иерархии сущностей и правдоподобных рассуждений на основе активно используемого в искусственном интеллекте метода прецедентов (Case-Based Reasoning). При решении различных задач интеллектуального анализа данных укзанные методы показали высокую степень эффективности, масштабируемости и независимости от структуры данных.

В рамках применения в классификаторе прецедентого подхода предложена модификация меры TF-IDF (Term Frequency — Inverse Document Frequency) оценки содержания текста с учетом известной информации о распределении документов по тематикам. Данная модификация повышает качество классификации по сравнению с классическими мерами, поскольку учитывает информацию о распределении слова не только в отдельном документе или тематике, но во всей базе прецедентов. Приведены экспериментальные результаты, подтверждающие эффективность предложенной метрики и разработанного классификатора применительно к задаче классификации фраз клиентов и предоставлению им необходимой информации в зависимости от результата классификации. Разработанный прототип сервиса текстовой классификации использован в рамках модуля голосового взаимодействия с пользователем в задаче роботизации системы маршрутизации телефонных звонков и перехода от кнопочного взаимодействия между пользователем и системой к голосовому.

Ключевые слова: интеллектуальные системы, рассуждения на основе прецедентов, маршрутизация вызовов, программные средства классификация, анализ данных.

Для цитирования: Куриленко И.Е., Никонов И.Е. Решение задачи классификации сообщений в системах голосового взаимодействия // Вестник МЭИ. 2020. № 5. С. 132—139. DOI: 10.24160/1993-6982-2020-5-132-139.

Solving the Message Classification Problem in Voice Interaction Systems

I.E. Kurilenko, I.E. Nikonov

A method for solving the problem of classifying short-text messages in the form of sentences of customers uttered in talking via the telephone line of organizations is considered. To solve this problem, a classifier was developed, which is based on using a combination of two methods: a description of the subject area in the form of a hierarchy of entities and plausible reasoning based on the case-based reasoning approach, which is actively used in artificial intelligence systems. In solving various problems of artificial intelligence-based analysis of data, these methods have shown a high degree of efficiency, scalability, and independence from data structure.

As part of using the case-based reasoning approach in the classifier, it is proposed to modify the TF-IDF (Term Frequency - Inverse Document Frequency) measure of assessing the text content taking into account known information about the distribution of documents by topics. The proposed modification makes it possible to improve the classification quality in comparison with classical measures, since it takes into account the information about the distribution of words not only in a separate document or topic, but in the entire database of cases. Experimental results are presented that confirm the effectiveness of the proposed metric and the developed classifier as applied to classification of customer sentences and providing them with the necessary information depending on the classification result. The developed text classification service prototype is used as part of the voice interaction module with the user in the objective of robotizing the telephone call routing system and making a shift from interaction between the user and system by means of buttons to their interaction through voice.

Key words: intelligent systems, case-based reasoning, call routing, classification software, data analysis.

For citation: Kurilenko I.E., Nikonov I.E. Solving the Message Classification Problem in Voice Interaction Systems. Bulletin of MPEI. 2020;5:132—139. (in Russian). DOI: 10.24160/1993-6982-2020-5-132-139.

Введение

Одна из актуальных задач развития искусственного интеллекта (ИИ) — разработка интеллектуальных систем поддержки принятия решения реального времени (ИСППР РВ), выполняющих важные задачи мониторинга и управления сложными техническими объектами и системами, построения систем обучения, поиска и обработки больших массивов данных и др. [1]. При моделировании правдоподобных рассуждений указанные системы используют методы неклассической логики, эвристические алгоритмы и накопленные знания [2]. Одним из направлений задачи обработки данных является задача текстовой классификации [3].

Средние и крупные компании в своем большинстве имеют отделы специалистов, предоставляющих информацию по продуктам и услугам компании, решению возникших при их использовании проблем и ошибок по телефону (call-центры). В некоторых из них звонки напрямую транслируются на операторов, в других есть кнопочные меню — программные системы голосового взаимодействия (Interactive Voice Response, IVR). В последние годы такие системы все больше роботизируются, позволяя предоставлять информацию и взаимодействовать с клиентами без участия оператора. Громоздкие голосовые меню с предварительно записанными звуками заменяются голосовыми роботами–помощниками — системами, использующими при взаимодействии с человеком распознавание речи на естественном языке (ЕЯ), классификацию распознанных фраз, синтез речи на ЕЯ по некоторому тексту. Переход к подобным системам дает положительный экономический эффект, поскольку робот–помощник берет на себя часть обязанностей операторов, а также позволяет уменьшить время ожидания клиентов, выполняя поиск информации в автоматическом режиме, что намного быстрее ручного. Данная задача актуальна, а в последние годы заметна высокая тенденция появления роботов-помощников на телефонных линиях, чат-ботах, в виде виртуальных сотрудников [4].

Рассмотрены методы решения задачи текстовой классификации распознанных фраз на ЕЯ по тематикам для построения диалогового общения с клиентом. Задача распознавания речи в данной работе не затрагивается. С некоторыми методами ее решения можно ознакомиться в работах [5, 6].

Наиболее популярны следующие методы классификации: на основе прецедентов (Case-Based Reasoning, CBR) [7], Байесовский классификатор [8], метод опорных векторов (Support Vector Machine, SVM) [9], деревья решений [10]. В [11] показано, что CBR-метод классификации показывает результаты текстовой классификации, не уступающие остальным методам. При этом он достаточно прост для реализации и независим к большим расхождениям данных в обучающем множестве. В работе [12] доказано, что наибольшую точность классификации можно достичь путем использования ансамблей (комбинаций) классификаторов. В [13] представлены результаты классификации фраз при комбинировании различных метрик оценки весов слов. В данной работе представлен подход к решению задачи классификации на основе комбинирования CBR-метода классификации и представления предметной области (ПО) в виде таксономии ее сущностей.

Описание данных для формирования прецедентов

Рассмотрена задача классификации фраз клиентов на основе звонков в call-центр банка. Она сводится к соотнесению фразы, произнесенной клиентом и преобразованной в текстовый формат, к определенной тематике. При этом длина фразы редко бывает больше 10...15 слов, а в большинстве случаев не превышает и 5. Примеры входящих фраз: «я хочу закрыть карту», «мне нужно узнать, когда я могу забрать карту», «изменить ПИН-код для карты» и т. д.

Множество записей для формирования обучающих прецедентов состоит из 1,1 тыс. фраз, распределенных по следующим тематикам: «Баланс», «Дата и сумма платежа», «Претензии», «Операции по счетам», «График работы», «Активация», «Готовность карт», «Кредитные заявки».

Классификация на основе описания предметной области

Для описания ПО разработана модель, включающая ключевые слова, сущности, правила и тематики. Ключевые слова описывают некоторую сущность (например, слова «карта» и «кредитка» описывают сущность «карта»), набор сущностей формирует правило (сущности «карта» и «отделение» формируют правило «карта — отделение»), а совокупность правил составляет тематику (пример для двух тематик представлен на рис. 1).

При формировании модели для 8 тематик создано 160 правил, включающих в себя 113 сущностей, описанных 398 ключевыми словами. При поступлении фразы в классификатор слова с целью корректного поиска среди ключевых слов нормализуются с помощью алгоритма лемматизации — приведения слова к нормальной форме [14].

Тематику текстовых сообщений определяют по алгоритму рис. 2.

Классификация на основе прецедентов

Другой классификатор разработан с использованием механизма рассуждений на основе прецедентов. В общем случае прецедент описывает некоторую ситуацию в виде набора:

$$Case = (x_1, x_2, ..., x_n, R),$$

где $x_1, \ldots, x_n$ — характеристики ситуации $x_1 \in X_1, \ldots, x_n \in X_n$; $X_1, \ldots, X_n$ — области допустимых значений соответствующих характеристик; R — предлагаемое решение.

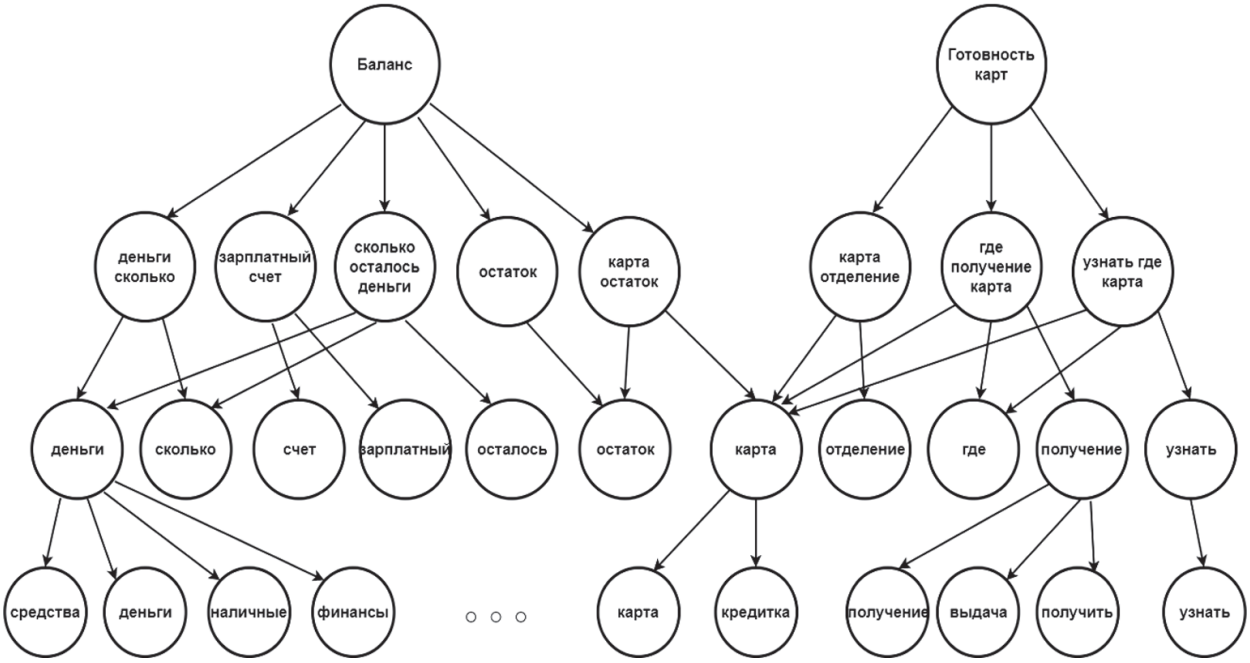

Рис. 1. Пример иерархии модели ПО для двух тематик

Применительно к задаче текстовой классификации, характеристики прецедента представляют из себя набор слов некоторого текста, а решение — сопоставление этого набора определенной тематике.

При возникновении новой проблемы CBR-система обращается к базе прецедентов (БП) для поиска и извлечения соответствующего прецедента. Преимуществом CBR-методов является, во-первых, то, что они применимы для плохо формализуемых ПО, а во-вторых, что в отличии от универсальных систем обучения они извлекают только релевантные прецеденты для решения проблемы и позволяют достаточно просто описывать ПО с отсутствием однородности данных в прецедентах.

Типовой CBR-цикл состоит из четырех этапов [15]:

- извлечения наиболее подходящих прецедентов;
- повторного использования извлеченных прецедентов для решения новой проблемы;
- проверки прецедента на корректность решения применительно к новой задаче и адаптации решения в случае необходимости;
- сохранения нового прецедента в БП.

Выбор правильного множества характеристик и ключевых слов для представления произвольного текстового документа в виде набора атрибутов (векторной модели) — сложная проблема, связанная с решением задач автоматического реферирования и аннотирования текста [16].

Векторная модель текста

При использовании векторной модели каждое слово взвешивается некоторым числом — весом, определяющим его информативность. Наиболее распространена метрика TF-IDF, используемая при текстовой классификации и состоящая из двух множителей [17]: Term Frequency (TF) — «говорит» о том, что чем выше частота употребления слова t в документе d, тем больше его значимость:

$$TF_{t,d} = \frac{k_t}{N_d},$$

где k_t — количество вхождений слова t в документ d; N_d — общее количество слов в документе.

Множитель IDF (Inverse Term Frequency) — инверсия частоты, с которой слово t встречается во множестве всех документов:

$$IDF_{t,N} = \log \frac{N}{df_t},$$

где N — общее число документов в БП; df_t — количество документов, в которых присутствует слово t.

При использовании прецедентов в системе имеется доступная база промаркированных темами текстов. Однако, в метрике TF-IDF данные о распределении слов относительно тематик никак не учитываются. Выделим из БП отличительные слова, характеризующие конкретные классы (тематики). Пусть t_1, t_2 — два слова, встречающиеся k раз в БП. При этом t_1 встречается равномерно во всех классах, а t_2 — появляется в классе c_i гораздо чаще, чем в остальных. Тогда значение $IDF_{t,N}$ для обоих слов одинаково, однако t_2 обладает большей идентифицирующей мощностью, и его присутствие в новых текстовых данных ясно говорит о принадлежно-

Входные данные: P = {w1,…,wn} – рассматриваемая фраза, состоящая из совокупности слов; D = <T, R, E, K> – описание предметной области, где T – множество тематик, R – множество правил, E – множество сущностей, K – множество ключевых слов
Выходные данные: t ∈ T – тематика, соответствующая фразе P или Ø

```
Set matchKeyWords ← {} //множество ключевых слов, содержащихся во фразе P
Set matchEntities ← {} //множество сущностей, соответствующих найденным ключевым словам
Set matchRules ← {} //множество правил, соответствующих найденным сущностям
Set resRules ← {} //результирующее множество правил для определения выходной тематики
foreach (w ∈ P) {
 foreach (k ∈ K) {
  if ( w == k) {
   matchKeyWords.add(k) //формирование множества ключевых слов по фразе
   break
  }
 }
}
foreach (mkw ∈ matchKeyWords) {
 foreach (e ∈ E) {
  if (e.contains(mkw)) {
   matchEntities.add(e) //поиск сущностей по ключевым словам
   break
  }
 }
}
foreach (r ∈ R) { //поиск правил по полученному списку сущностей
 if (matchEntities.containsAll(r.getEntities())) matchRules.add(r)
}
if (matchRules.isEmpty()) return Ø //если правил нет – тематика не определена
else {
 int weight = 1
 foreach (mr ∈ matchRules) {
  if (mr.getEntitiesCount() > weight) { //выбирается правило с максимальным весом – количеством описывающих его сущностей
   resRules.clear()
   resRules.add(mr)
   weight = mr.getEntitiesCount()
  } //если единственное правило с максимальным весом не найдено, возвращается список правил с одинаковым весом
  else if (mr.getEntitiesCount() == weight) {
   resRules.add(mr)
  }
 }
} //если найдено единственное правило, тематика, которую оно описывает, возвращается в качестве ответа
if (resRules.size() == 1) return T.getThemeByRule(resRules.get(0))
else return T.getThemeByMaxRules(resRules) //иначе возвращается тематика, содержащая наибольшее число правил
```

Рис. 2. Алгоритм определения тематики фраз

сти к классу c_i. Создадим метрику, использующая эту информацию для более точной классификации.

Вычисление метрики слов с учетом распределения по тематикам

Для адекватного описания тематик по прецедентам рассмотрим документы, принадлежащие к одной тематике, как единое целое и рассчитаем информативность слов относительно тематик, а не документов.

Для информативных в заданной тематике слов должны выполняться следующие условия:

- слово обязано иметь высокую частоту появления, что говорит о его информативности в пределах тематики;
- слову следует обладать низкой частотой появления в других тематиках;
- у него должна быть высокая степень распространения среди всех документов в пределах тематики, т. е. чем в большем количестве документов появляется слово, тем больше его информативность. Слова, встречающиеся с большой частотой в пределах небольшого числа документов тематики, рассматриваются как зашумленные данные.

Таким образом, получим следующую формулу для вычисления веса слова *t* для тематики *c*:

$$Weight_t(c) = \frac{p_{t,c}}{IDF_{t,c}} IDF_{t,rest},$$

где $IDF_{t,c}$ — степень распределения слова *t* по документам в пределах тематики *c* (чем больше слово *t* распределено по документам, составляющим тематику *c*, тем меньше значение $IDF_{t,c}$, и, тем самым, выше вес слова в тематике); $IDF_{t,rest}$ — степень распределения слова *t* по документам для других тематик *rest*; $p_{t,c}$ — вероятность того, что если слово *t* присутствует в некотором документе тематики *c*, то оно принадлежит ей,

$$p_{t,c} = \frac{k_{t,c}}{k_{t,N}},$$

где $k_{t,c}$ — количество вхождений слова t в тематику c; $k_{t,N}$ — число вхождений слова во все тематики.

При классификации нового документа d с неизвестным содержимым для каждой описанной по прецедентам тематике вычисляется метрическое значение близости и выбирается максимальное. Найдем это значение по формуле:

$$Sim(d,c) = \sum_t k_{t,d} Weight_t(c),$$

где $k_{t,d}$ — количество появлений слова t во входном документе d.

Реализация сервиса классификации

В рамках данной работы на языке Java с использованием фреймворка Spring Boot разработан rest-сервис классификации, вызываемый из IVR-приложения. Базовым считается классификатор на основе описательной модели, а результат работы CBR-классификатора нужен в случае, если фразу не удалось классифицировать базовым классификатором. Подобные случаи сохраняются для дальнейшего анализа и обогащения модели.

Общее описание процесса.

1. Клиент звонит на входящую линию организации.

2. Звонок попадает на MPP-сервер, который «общается» с приложением IVR по HTTP-протоколу, передавая фразы и действия (нажатия кнопок) клиента и отправляя ему звуковые ответы от приложения.

3. IVR при получении голосовой фразы отдает ее в модуль распознавания речи.

4. После получения распознанной фразы в текстовом формате, IVR пересылает фразу сервису классификации и строит логику обработки звонка согласно распознанной (или нет) тематике.

Общая архитектура процесса изображена на рис. 3. Пример работы системы дан на рис. 4.

Клиент звонит на линию организации и говорит: «Здравствуйте, я хочу активировать кредитную карту». Эта фраза передается в приложение IVR, распознается в модуле распознавания и отправляется в текстовом формате сервису классификации. Классификатор разбирает фразу, нормализует слова и ищет наиболее подходящие тематики: тематике «приветствие» соответствует слово «здравствуйте», тематике «приобретение кредитной карты» — слова «хотеть, кредитный, карта», а тематике «активация» – «хотеть, активация, кредитный, карта». Таким образом, наибольшее число слов соответствует тематике «активация», передаваемой в IVR в качестве результата классификации для дальнейшего построения логики приложения (переход в модуль активации карт).

Результаты компьютерного моделирования

Для количественной оценки качества классификации взята классическая для данного класса задач F_1 мера [18], вычисляемая на основе точности (Precision) и полноты (Recall):

$$F_1 = 2\frac{precision \cdot recall}{precision + recall},$$

где под точностью системы в пределах класса понимается доля документов, действительно принадлежащих данному классу относительно всех документов, которые система отнесла к этому классу, а под полнотой системы — доля найденных классификатором документов, принадлежащих классу относительно всех документов данного класса.

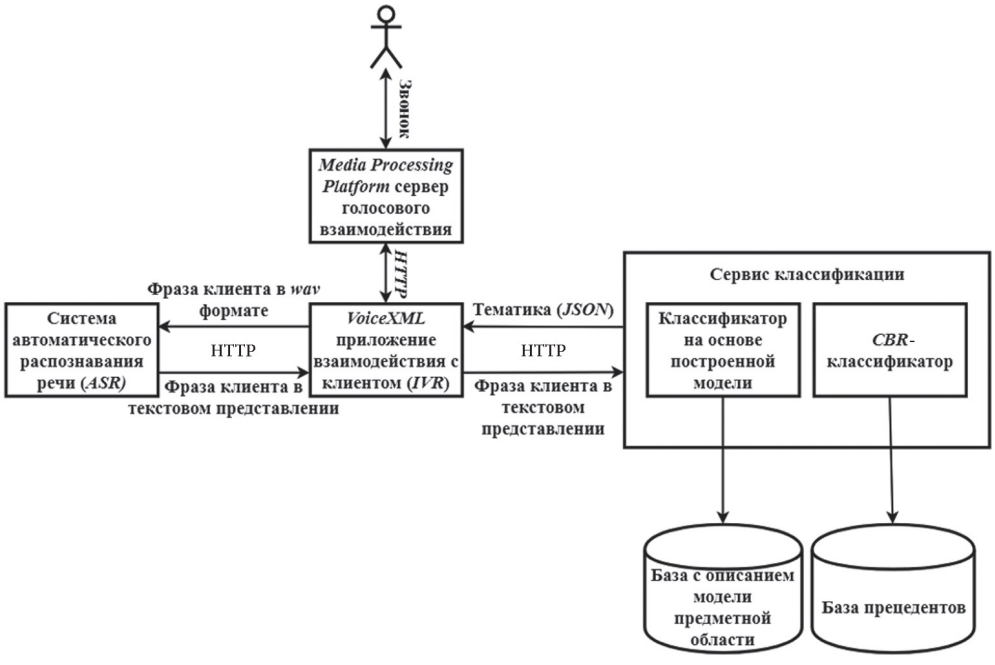

Рис. 3. Архитектура приложения

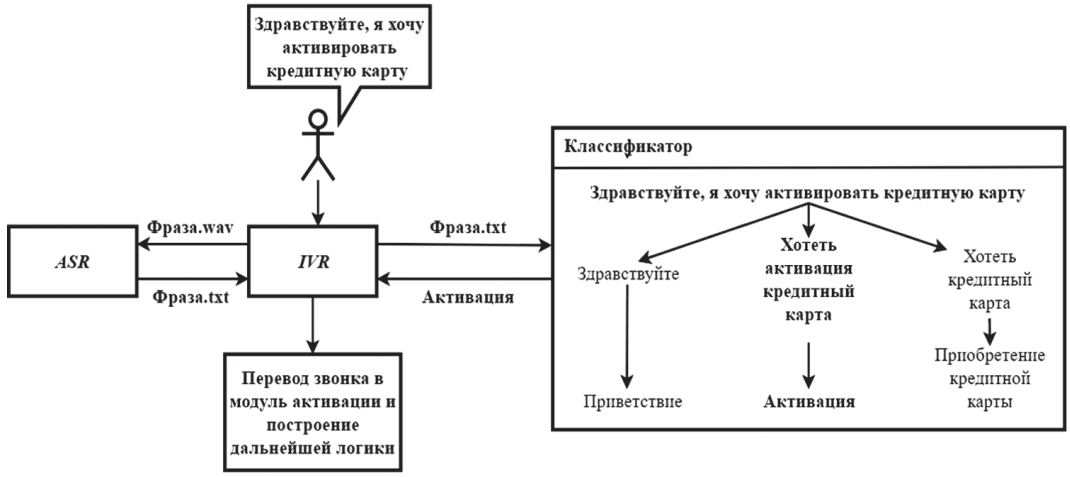

Рис. 4. Пример работы системы

В таблице 1 представлены результаты для различных метрик оценки весов — TF-IDF, Relevance Frequency (RF) [19] и разработанной метрики, примененных в CBR-классификаторе.

Таблица 1

Сравнение качества классификации по метрикам

Метрика	Мера F_1
TF-IDF	0,83
RF	0,89
Модификация TF-IDF	0,90

В таблице 2 даны результаты классификации при использовании классификаторов по отдельности и в комбинации.

Таблица 2

Сравнение качества классификации по классификаторам

Классификатор	Мера F_1
CBR	0,90
На основе построенной модели	0,93
Комбинация классификаторов	0,94

Из таблицы 1 следует, что предложенная метрика показала наилучший результат, поскольку учитывала информацию о распределении слов по тематикам. Из таблицы 2 очевидно, что классификатор, использующий построенную модель предметной области, дал наилучший результат, что говорит об адекватности и корректности построенной модели. Это объясняется минимальной длиной классифицируемых фраз и ограниченностью словаря предметной области. Комбинация методов классификации позволила улучшить классификацию и доказала, что не все возможные случаи учтены в разработанной модели.

Заключение

Проанализирована задача классификации фраз клиентов на звонках в call-центр в виде текстовых сообщений.

Создана модифицированная TF-IDF-метрика для оценки весов слов в документах и представлены вычислительные результаты, доказывающие ее эффективность.

Разработаны два классификатора: с применением CBR-технологии и использованием модели предметной области. Полученный веб-сервис показал высокую степень классификации для рассматриваемой задачи. Спроектированный CBR-классификатор также необходим для расширения каналов взаимодействия с виртуальным сотрудником [20].

В дальнейшем планируются развитие модели в плане повышения классификации путем учета семантической взаимосвязи слов внутри фраз [21], а также автоматизация процесса формирования новых правил для модели из новых классифицируемых в CBR-классификаторе случаев.

Литература

1. **Башлыков А.А., Еремеев А.П.** Основы конструирования интеллектуальных систем поддержки принятия решений в атомной энергетике. М.: ИНФРА-М, 2018.

2. **Еремеев А.П., Варшавский П.Р., Куриленко И.Е.** Моделирование временных зависимостей в интел-

References

1. **Bashlykov A.A., Eremeev A.P.** Osnovy Konstruirovaniya Intellektual'nykh Sistem Podderzhki Prinyatiya Resheniy v Atomnoy Energetike. M.: INFRA-M, 2018. (in Russian).

2. **Eremeev A.P., Varshavskiy P.R., Kurilenko I.E.** Modelirovanie Vremennykh Zavisimostey v Intellektual'-

лектуальных системах поддержки принятия решений на основе прецедентов // Information Technol. and Knowledge. 2012. V. 6. No. 3. Pp. 227—239.

3. **Kobayashi V.B. e. a.** Text Classification for Organizational Researchers: a Tutorial // Organizational Research Methods. 2018. V. 21(3). Pp. 766—799.

4. **Куриленко И.Е.** Применение рассуждений на основе прецедентов для реализации виртуального сотрудника отдела сопровождения программного обеспечения // Труды 16 Национальной конф. по искусственному интеллекту с междунар. участием. М.: РКП, 2018. Т. 2. С. 238—244.

5. **Tiken M.** Automatic Speech Recognition System: a Survey Report // Sci. & Technol. J. 2016. V. 4. Pp. 152—155.

6. **Arora J., Rishi Sh.** Automatic Speech Recognition: a Review. Intern. J. Computer Appl. 2012. V. 60. Pp. 34—44.

7. **Еремеев А.П., Варшавский П.Р.** Моделирование рассуждений на основе прецедентов в интеллектуальных системах поддержки принятия решений // Искусственный интеллект и принятие решений. 2009. № 2. С. 45—57.

8. **Lewis D.D.** Naive (Bayes) at Forty: the Independence Assumption in Information Retrieval // Proc. 10 European Conf. Machine Learning. Heidelberg: Springer, 1998. V. 1398. Pp. 4—15.

9. **Dumais S., Platt J, Heckerman D., Sahami M.** Inductive learning Algoritms and Representations for Text Categorization // Proc. 7 Intern. Conf. Information and Knowledge Management. N.-Y.: ACM Press, 1998. Pp. 148—155.

10. **Noormanshah W., Nohuddin P., Zainol Z.** Document Categorization Using Decision Tree: Preliminary Study // Intern. J. Eng. and Technol. 2018. V. 7. Pp. 437—440.

11. **Healy M.** Investigating Text Message Classification Using Case-based Reasoning. Dublin: Dublin Institute of Technology, 2007.

12. **Епрев А.С.** Автоматическая классификация текстовых документов // Математические структуры и моделирование. 2010. № 1. С. 65—81.

13. **Sergienko R., Shany M., Minkerz W., Semenkin E.** Topic Categorization Based on Collectives of Term Weighting Methods for Natural Language Call Routing // J. Siberian Federal University. Mathematics and Physics. 2016. V. 9. Pp. 235—245.

14. **Korobov M.** Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. Communications in Computer and Information Sci. Springer Intern. Publ., 2015. V. 542. Pp. 320—332.

15. **Aamodt A., Plaza E.** Case-based Reasoning: Foundational Isuues, Methodological Variations, and System Approaches // AI Communications. 1994. V. 7 (1). Pp. 39—59.

16. **Mazyad A., Teytaud F., Fonlupt C.** A Comparative Study on Term Weighting Schemes for Text Classification // Proc. Third Intern. Conf. Machine Learning, Optimization and Big Data. Tuscany, 2017.

nykh Sistemakh Podderzhki Prinyatiya Resheniy na Osnove Pretsedentov. Information Technol. and Knowledge. 2012;6;3:227—239. (in Russian).

3. **Kobayashi V.B. e. a.** Text Classification for Organizational Researchers: a Tutorial. Organizational Research Methods. 2018;21(3):766—799.

4. **Kurilenko I.E.** Primenenie Rassuzhdeniy na Osnove Pretsedentov dlya realizatsii Virtual'nogo Sotrudnika Otdela Soprovozhdeniya Programmnogo Obespecheniya. Trudy 16 Natsional'noy Konf. po Iskusstvennomu Intellektu s Mezhdunar. Uchastiem. M.: RKP, 2018;2:238—244. (in Russian).

5. **Tiken M.** Automatic Speech Recognition System: a Survey Report. Sci. & Technol. J. 2016;4:152—155.

6. **Arora J., Rishi Sh.** Automatic Speech Recognition: a Review. Intern. J. Computer Appl. 2012;60:34—44.

7. **Eremeev A.P., Varshavskiy P.R.** Modelirovanie Rassuzhdeniy na Osnove Pretsedentov v Intellektual'nykh Sistemakh Podderzhki Prinyatiya Resheniy. Iskusstvennyy Intellekt i Prinyatie Resheniy. 2009;2:45—57. (in Russian).

8. **Lewis D.D.** Naive (Bayes) at Forty: the Independence Assumption in Information Retrieval. Proc. 10 European Conf. Machine Learning. Heidelberg: Springer, 1998;1398: 4—15.

9. **Dumais S., Platt J, Heckerman D., Sahami M.** Inductive learning Algoritms and Representations for Text Categorization. Proc. 7 Intern. Conf. Information and Knowledge Management. N.-Y.: ACM Press, 1998:148—155.

10. **Noormanshah W., Nohuddin P., Zainol Z.** Document Categorization Using Decision Tree: Preliminary Study. Intern. J. Eng. and Technol. 2018;7:437—440.

11. **Healy M.** Investigating Text Message Classification Using Case-based Reasoning. Dublin: Dublin Institute of Technology, 2007.

12. **Eprev A.S.** Avtomaticheskaya Klassifikatsiya Tekstovykh Dokumentov. Matematicheskie Struktury i Modelirovanie. 2010;1:65—81. (in Russian).

13. **Sergienko R., Shany M., Minkerz W., Semenkin E.** Topic Categorization Based on Collectives of Term Weighting Methods for Natural Language Call Routing. J. Siberian Federal University. Mathematics and Physics. 2016;9:235—245.

14. **Korobov M.** Morphological Analyzer and Generator for Russian and Ukrainian Languages. Analysis of Images, Social Networks and Texts. Communications in Computer and Information Sci. Springer Intern. Publ., 2015;542:320—332.

15. **Aamodt A., Plaza E.** Case-based Reasoning: Foundational Isuues, Methodological Variations, and System Approaches. AI Communications. 1994;7 (1):39—59.

16. **Mazyad A., Teytaud F., Fonlupt C.** A Comparative Study on Term Weighting Schemes for Text Classification. Proc. Third Intern. Conf. Machine Learning, Optimization and Big Data. Tuscany, 2017.

17. **Sparck K.J.** A Statistical Interpretation of Term Specificity and its Application in Retrieval // J. Documentation. 1972. V. 28. No. 1. Pp. 11—21.

18. **Goutte C., Gaussier E.** A Probabilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation // Proc. 27 European Conf. Advances in Information Retrieval Research. Berlin: Springer-Verlag, 2005. Pp. 345—359.

19. **Lan M., Tan Ch., Su J., Lu Y.** Supervised and Traditional Term Weighting Methods for Automatic Text Categorization // IEEE Trans. Pattern Analysis and Machine Intelligence. 2009. V. 31. No. 4. Pp. 721—735.

20. **Kurilenko I.E., Nikonov I.E.** Virtual Employee Implementation Using Temporal Case-based Reasoning // Enterprise Engineering and Knowledge Managemen: Selected Papers XXII Intern. Conf. Moscow, 2019. V. 2413. Pp. 77—85.

21. **Фомин В.В., Флегонтов А.В., Осочкин А.А.** Метод частотно-морфологической классификации текстов // Программные продукты и системы. 2017. № 3. С. 478—486.

17. **Sparck K.J.** A Statistical Interpretation of Term Specificity and its Application in Retrieval. J. Documentation. 1972;28;1:11—21.

18. **Goutte C., Gaussier E.** A Probabilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation. Proc. 27 European Sonf. Advances in Information Retrieval Research. Berlin: Springer-Verlag, 2005:345—359.

19. **Lan M., Tan Ch., Su J., Lu Y.** Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. IEEE Trans. Pattern Analysis and Machine Intelligence. 2009;31;4:721—735.

20. **Kurilenko I.E., Nikonov I.E.** Virtual Employee Implementation Using Temporal Case-based Reasoning. Enterprise Engineering and Knowledge Managemen: Selected Papers XXII Intern. Conf. Moscow, 2019;2413:77—85.

21. **Fomin V.V., Flegontov A.V., Osochkin A.A.** Metod Chastotno-morfologicheskoy Klassifikatsii Tekstov. Programmnye Produkty i Sistemy. 2017;3:478—486. (in Russian).

Сведения об авторах:

Куриленко Иван Евгеньевич — кандидат технических наук, доцент кафедры прикладной математики и искусственного интеллекта НИУ «МЭИ», e-mail: ivan@appmat.ru
Никонов Игорь Евгеньевич — аспирант кафедры прикладной математики и искусственного интеллекта НИУ «МЭИ», e-mail: nikonovie@gmail.com

Information about authors:

Kurilenko Ivan E. — Ph.D. (Techn.), Assistant Professor of Applied Mathematics and Artificial Intelligence Dept., NRU MPEI, e-mail: ivan@appmat.ru
Nikonov Igor E. — Ph.D-student of кафедры Applied Mathematics and Artificial Intelligence Dept., NRU MPEI, e-mail: nikonovie@gmail.com

Работа выполнена при поддержке: РФФИ (проекты № 18-01-00459 а, № 20-07-00498 а, № 18-29-03088 мк, № 20-57-00015 Бел_а)
The work is executed at support: RFBR (Projects No. 18-01-00459 a, No. 20-07-00498 a, No. 18-29-03088 мк, No. 20-57-00015 Бел_а)

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов
Conflict of interests: the authors declare no conflict of interest

Статья поступила в редакцию: *27.12.2019*
The article received to the editor: *27.12.2019*