Ссылка на статью:
// Ученые записки УлГУ. Сер. Математика и информационные технологии. УлГУ. Электрон. журн. 2020, № 1, с. 138-145.

Поступила: 29.04.2020
Окончательный вариант: 27.05.2020

УДК 004.912

Алгоритм обработки запросов пользователей государственных информационных систем

Шлеменкова Е.О.*[*], *Ципес Г. Л.

[*]xjatr00@mail.ru
НИТУ МИСИС, Москва, Россия

В данной статье предложен алгоритм, позволяющий повысить качество и скорость предоставления методических материалов в государственных информационных системах, основанный на анализе данных учетной записи пользователя и его действий в системе. В работе приводится подробное описание этапов построение алгоритма, а также его процесс работы в нотации BPMN. Описаны методы поиска и ранжирования текстовой информации, которые использовались при реализации алгоритма. Данные для анализа качества алгоритма основаны на реальных обращениях пользователей в рассматриваемую государственную информационную систему. Результаты эффективности представлены в виде сравнительного анализа вывода запрашиваемой информации с использованием предложенного алгоритма и без него. За основу принята методика оценки качества и информативности поиска. Предложенный алгоритм может быть также использован в других аналогичных государственных системах.

***Ключевые слова**: методические материалы, обработка текстов, алгоритм, действия пользователей, государственная система.*

Введение

Последнее время все более явно наблюдается тенденция к автоматизации процесса поддержки пользователей в различных государственных системах. В известных государственных систем, таких как «ГОСУСЛУГИ» или «ФНС», предоставления методической поддержки пользователю осуществляется в виде централизованного вывода методических материалов, которые представляют собой официальные текстовые документы следующего вида: инструкции по взаимодействию с системой, руководства пользователей, правила организации доступа и другие. При этом как в этих, так и в аналогичных системах в процессе формирования предоставляемых материалов пользователю не учитываются его информационные потребности и сведения о его учетных данных в системе.

Предлагается алгоритм, который, анализируя запрос пользователя и его действия в системе, выводит ему соответствующие конкретные рекомендации.

Описание алгоритма

Алгоритм обработки запросов пользователей состоит из следующих основных этапов:

1. Анализ пользовательского запроса
2. Определение функциональной области.
3. Определение формы пользовательского запроса.
4. Анализ данных учетной записи пользователя.
5. Обработка методических материалов
6. Вывод соответствующих конкретных рекомендаций.

Рассмотрим данные этапы подробнее.

На первом этапе при анализе запроса определяется его сложность. Запрос считается сложным, если текст содержит более одного предложения. Для подобных запросов по каждому предложению осуществляется проверка на наличие в нем ключевых слов. Данный набор слов необходимо предварительно составить экспертным путем, а также определить для каждого слова его весовую значимость. При условии отсутствия ключевых слов в предложении, оно не будет учитываться в дальнейшей обработке.

На втором этапе определяется функциональная область запроса, представляющая собой отдельный модуль государственной системы. Предварительно необходимо обозначить данный ряд областей, в соответствии с системой. Например, функциями могут выступать следующие категории услуг: транспорт, образование, здоровье и другие, по которым пользователю необходимо получить методические рекомендации.

На данном этапе принято использовать метод тематической классификации текстовой информации [1] для соотнесение запроса пользователя с выделенными экспертным путем функциональными областями с возможностью обучения классификатора для повышения эффективности его работы. В результате чего определится основная область системы, на которую направлен запрос пользователя.

Важно получить наиболее высокий процент эффективности работы классификатора при определении функциональной области, так как ее выбор является основой для дальнейшей обработки запроса пользователя.

В качестве критерия, оценивающего работу классификатора, принято использовать F-меру (1) – метрика, которая объединяет в себе информацию о точности и полноте клас-

сификатора. Данное значение представляет собой гармоническое среднее между точностью и полнотой. Она стремится к нулю, если точность или полнота стремится к нулю [2].

$$F=2*\frac{(\text{Точность}*\text{Полнота})}{(\text{Точность}+\text{Полнота})} \quad (1)$$

Для выбранной метрики экспертным путем необходимо определить процентный диапазон значений, относительного которого произведется выбор функциональной области. Для этого в работе проведено тестирование работы тематического классификатора на реальных обращениях пользователей в службу поддержки. При начальном использовании классификатора среднее значение F равнялось 29,5%, после обучения данное значение увеличилось до 68,7%. В соответствии с полученными результатами экспертным путем принято выделить в качестве максимального F значение 70%, которое дает положительный результат. А для отрицательного результата значение F равное 30%.

В случае получения значения F менее 30%, функциональная область, которая определена классификатором, не используется и на данном шаге анализируется учетная запись пользователя, в частности наличие связи с модулями системы.

Если область, выбранная классификатором, совпадает с данными у пользователя, тогда на последующих этапах учитывается данная функциональная область.

В противном случае модуль определяется случайно из всех имеющихся в системе областей.

Для значения F, составляющего более 70%, в дальнейшей работе применяется выбранная классификатором функциональная область. При получении значения F в диапазоне от 30% до 70% необходимо обратиться к предварительно составленному экспертным путем набору последовательных функциональных областей.

В данной работе ряд функциональных областей запроса пользователя определен исходя из анализа обращений пользователей в службы поддержки системы в разрезе модулей системы, времени поступления запроса пользователя и уровня управление, к которой привязана учетная запись пользователя. В качестве уровней управление рассматриваются следующие органы власти: ФОИВ (Федеральный орган исполнительной власти), РОИВ (Региональный орган исполнительной власти) и ОМСУ (Орган местного самоуправления). Порядок определен согласно количеству обращений по каждому из модулей системы в разрезе месяцев и уровня управление от большего значения к меньшему.

При определении данного набора функциональных областей также имеет место использование группировки обращений пользователей по следующему виду: юридиче-

ские и физические лица, иностранные граждане, индивидуальные предприниматели - данный ряд необходимо выделить в соответствии с особенностями системы.

На третьем этапе определяется форма запроса, которая представляет собой тип пользовательского запроса, например: необходимы рекомендации для получения доступа к функционалу системы, инструкции по заполнению полей форм, вопросы, связанные с возможностью интеграции данных и т.п. На данном этапе также применяется метод тематической классификации. Оценка эффективности работы классификатора не производится.

На четвертом этапе анализируются данные о пользователе, который имеет учетную запись в системе. В работе рассматриваются следующие пользовательские данные:

1. Наличие прав доступа пользователя к модулям системы.
2. Данные о программном обеспечении ПК пользователя, которое он использует при работе с системой.
3. Уровень управления, к которой привязана учетная запись пользователя.
4. Действия пользователя в системе, в частности, данные о посещении им разделов.
5. Имеющиеся в системе запросы от пользователей в методическую или техническую службы поддержки.

В случае нахождения в данных пользователя ключевой информации, которая может быть использована при ответе на поступивший запрос, например, аналогичный имеющийся в системе пользовательский запрос и ответ на него, решение выводится автоматически. Иначе, для ответа на запрос необходима обработка методической информации, которая содержится в системе.

На пятом этапе для формирования адаптированных под пользователя методических материалов необходимо предварительно экспертным путем определить источники информации, которые являются информативными для выбранных функциональных областей и форм запроса. В качестве источников выступают документы, хранимые в системе в структурированном виде.

Для выделения ***отдельных смысловых фрагментов*** из источников предлагается провести процесс автоматической разметки документов:

- токенизацию – разделение текста на слова, удаление знаков препинания, вводных слов, частиц, стоп слов и других различных символов;

− определение частей речи при помощи регулярных выражений [3], и методов по их использованию;

− нормализацию – нахождение основы каждого слова, которая необязательно должна совпадать с морфологическим корнем слова. В качестве способа решения данной задачи может использоваться стеммер Портера [4].

Полученные слова анализируются, и составляется список часто встречающихся слов и соответствующее им весовое значение. Для выявления совпадений нормализованных слов запроса с текстом источника методических материалов используется алгоритм нечеткого поиска [5].

В случае наличия в тексте ключевых слов, определенных на первом этапе, вес увеличивается в соответствии со значимостью определенных слов.

После определения весовых значений, для каждой группы (запрос – смысловой фрагмент источника данных) выводятся слова с весом более 2 – данный вес служит рангом, после слова упорядочиваются по степени возрастания значений рангов.

На данном этапе в работе используется ранговый критерий Спирмена [6] для определения наиболее значимого смыслового фрагмента источника, который будет использован в качестве ответа на запрос пользователя. Для корректной работы рангового критерия Спирмена необходимо, чтобы рассматриваемые наборы были одинаковой длины, поэтому проводится усечение наиболее длинного набора.

Статистическая значимость полученного коэффициента оценивается при помощи t-критерия Стьюдента.

На шестом этапе при получении коэффициента корелляции не превосходящего 0,3 предлагается в качестве ответа на запрос пользователя стандартное решение, подобранное экспертным путем.

При значениях коэффициента корелляции в диапазоне от 0,3 до 0,7 выводятся контакты специалистов службы поддержки.

При значениях более 0,7 выводится методическая рекомендация, полученная в результате работы алгоритма.

Процесс описанного алгоритма в нотации BPMN приводится на Рис. 1.

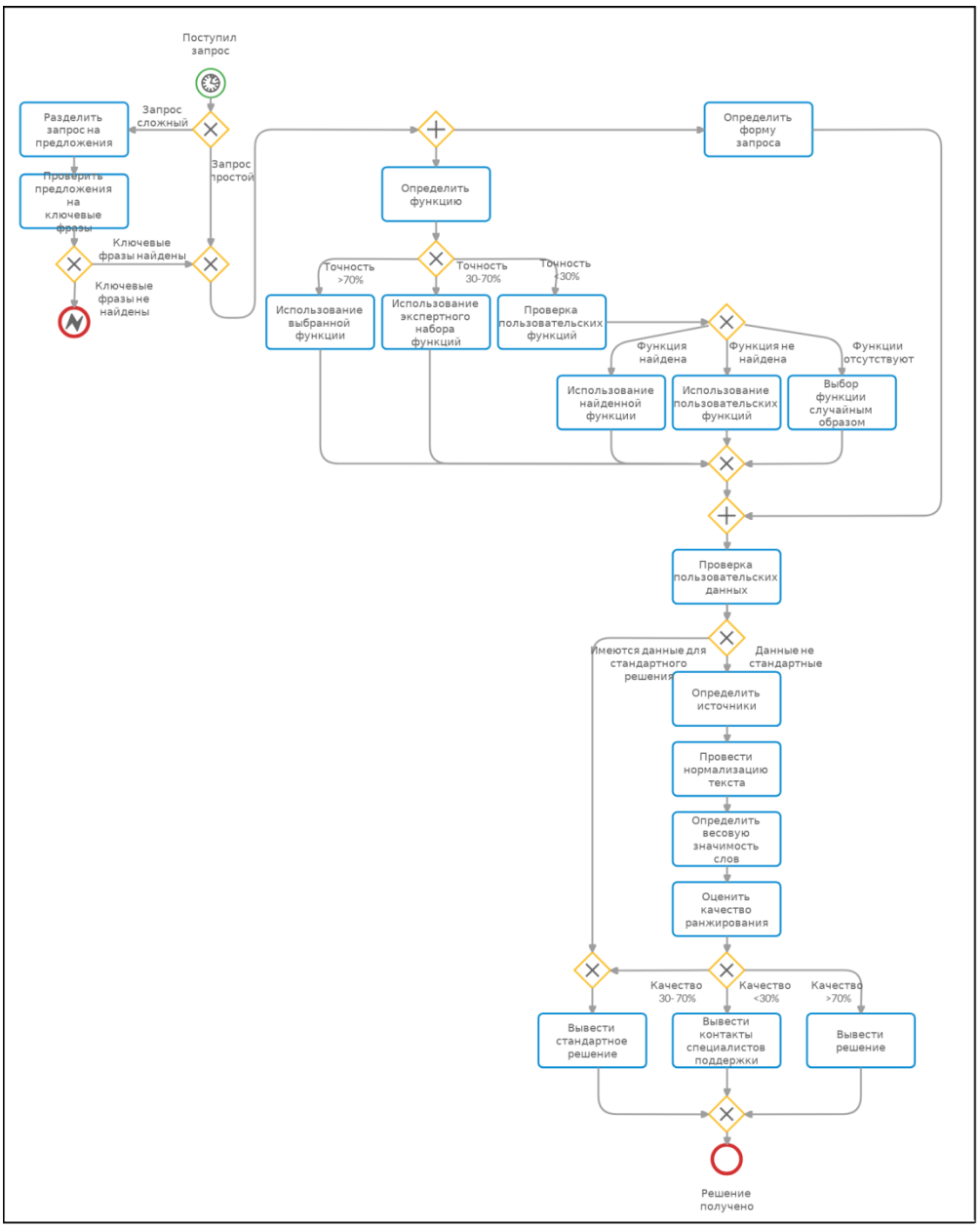

Рис. 1. Алгоритм обработки запросов пользователей

Тестирования алгоритма проводилось на 72 обращениях, среди которых были выделены шесть функциональных областей и двенадцать форм запроса, которые наиболее часто поступают в службу поддержки рассматриваемой системы.

Для оценки качества алгоритма используется методика оценки качества и информативности поиска[7]. Полученные результаты работы алгоритма сравниваются с реальными ответами сотрудников службы поддержки. На Рис. 2 представлены результаты данного сравнения.

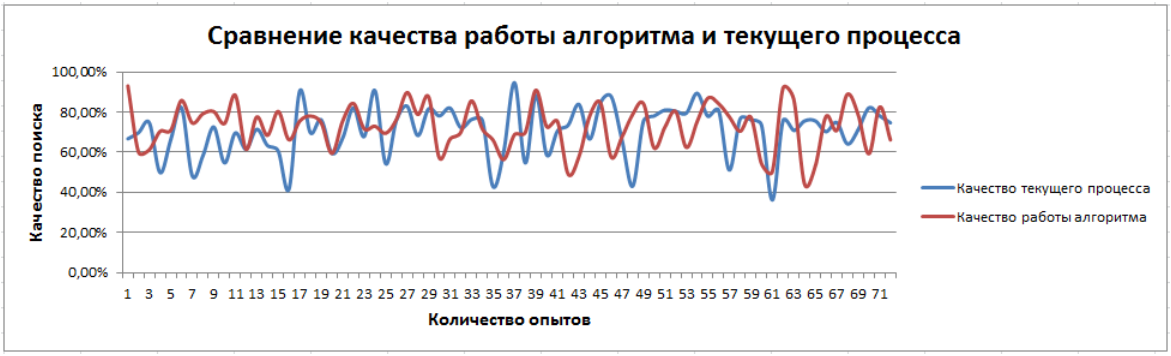

Рис. 2. Сравнение качества работы алгоритма с ответами службы поддержки

Исходя из полученных данных, можно сделать вывод, что в 59,72% случаев с использованием алгоритма качество предоставляемых материалов выше. Средняя оценка качество работы алгоритма – 72,89%, а предоставления ответов службой поддержки – 70,93%.

Заключение

В работе представлен алгоритм обработки запросов пользователей государственных информационных систем, который предоставляет ответы согласно информационным потребностям каждого пользователя. Особенностью данного подхода является адаптация выдаваемых методических материалов пользователю, основанная на поэтапном учете профиля и действий пользователя в системе.

На реальных запросах пользователей алгоритм показал высокое качество выводимой методической информации – 72,89% %, что выше значения 70,93%, которое показывает текущий процесс по оказанию методической поддержки.

Алгоритм можно использовать для аналогичных государственных систем с доработкой экспертных данных по функциональным областям, формам запроса, источникам данных и ключевым словам.

Анализируемые данные пользователей могут также быть дополнены для получения результата в соответствии с используемой системой. Методы, которые используются в данной работе, могут быть изменены на наиболее подходящие для системы, в которой применяется алгоритм.

Решение, полученное в данной работе, позволит повысить качество обслуживания пользователей и снизить затраты на службу поддержки.

Список литературы

1. Иванов В.К., Иванов К.В. *Введение в информационно-поисковые системы. Часть 2. Методические указания по изучению дисциплины "Мировые информационные ресурсы" для студентов специальностей «Прикладная информатика (в экономике)» и «Информационные системы и технологии»*. Тверь, 2005.

2. Дудченко, П. В. Метрики оценки классификаторов в задачах медицинской диагностики // *Сборник трудов XVI Международной научно-практической конференции студентов, аспирантов и молодых ученых.* Томский политехнический университет, 2019. С. 165.
3. Фридл Дж. *Регулярные выражения*, 3-е издание. Пер. с англ. СПб.: Символ-Плюс, 2008. 608 с.
4. Willett P. The Porter stemming algorithm: then and now. *Program: electronic library and information systems*, 2006, v.40 (3), p. 219-223.
5. Мосалев П.М. Обзор методов нечеткого поиска текстовой информации // *Вестник Московского государственного университета печати*, 2013. С.87
6. Гланц. С. *Медико-биологическая статистика*. Пер. с англ. М., Практика, 1998. 459 с.
7. Шлеменкова Е.О. Методика оценки качества и информативности поиска // *75-е Дни науки НИТУ МИСИС*, 2020.