УДК 004.91:336.7

Г. Г. КУЛИКОВ, К. А. ТИМОФЕЕВ

КЛАССИФИКАЦИЯ ДОКУМЕНТОВ БАНКА ПО ТЕХНОЛОГИЧЕСКИМ ПРОЦЕССАМ

Рассматривается подход к построению системы автоматизированной классификации документов по технологическим процессам. Производится построение модели документа процесса, приведено описание алгоритма классификации. *Документ; процесс; модель; алгоритм; классификация; обучение*

Сегодня высокие показатели деятельности банка на рынке розничных услуг достижимы лишь при условии активного освоения и внедрения новых технологий управления кредитными организациями. Процессный подход к управлению организацией является одним из основных способов существенного улучшения основных показателей предприятия, повышения его конкурентоспособности и снижения издержек [1].

Специфической чертой банковской деятельности (как для кредитной организации, так и для территориального учреждения Банка России) является то, что практически все виды работ и операции документируются. Документ является основанием для принятия управленческих решений и фиксирует их, является свидетельством об исполнении, обеспечивает закрепление информации для передачи ее во времени и пространстве, материалом для справочной работы. Учитывая большое количество видов документов и важность функций, выполняемых документом в процессе управления организацией, необходимо максимально эффективно наладить работу с документами. В Банке России информационно-технологическую основу корпоративной системы документооборота составляет Система автоматизации документооборота и делопроизводства (САДД). В настоящее время осуществляется эксплуатация системы в Центральном аппарате и в 78 главных управлениях и национальных банках Банка России, а также в 9 организациях при ЦБ РФ. Не меньшую значимость для улучшения функционирования всей структуры Банка России имеет внедряемое процессное управление, характеризующееся наличием описания всех технологических процессов организации и документов, появляющихся во время их исполнения. В настоящее время деятельность Национального банка Республики Башкортостан Банка России (НБ РБ) разделена на 18 направлений деятельности, каждое из которых состоит из процессов и подпроцессов, и описывается 850 процессами, а число видов документов, используемых при их выполнении, достигает количества в 100 видов [2].

В ходе выполнения своих обязанностей сотрудники НБ РБ создают документы и размещают их в системе автоматизации документооборота, а затем вручную осуществлять привязку этих документов к исполняемым ими процессам в системе поддержки процессного управления. Необходимость привязки созданного документа к технологическому процессу обусловлена принципами технологии workflow (потоки работ), согласно которым исполнение процесса не продолжится, пока в контрольной точке не будет зарегистрирован необходимый документ. Также необходимость регистрации документов в контрольных точках связана с задачами оценки своевременности исполнения должностных обязанностей сотрудниками, оценки фактической стоимости процессов, сбора статистики.

В этих условиях актуальна задача создания автоматизированной системы, целью которой является снижение трудоемкости предобработки документов и автоматическое определение, к какому именно технологическому процессу относится поступивший входящий документ или созданный внутренний документ. Система должна прикреплять документ из системы автоматизации документооборота к соответствующему запуску процесса в системе поддержки процессного управления или инициировать новый запуск в случае необходимости.

Традиционно в документе выделяются следующие зоны: заголовок, содержательная часть, задание, уведомление. В зоне заголовка содержится служебная информация, необходимая для правильной передачи и интерпретации всего документа в целом. Содержательная часть документа содержит сам документ вместе с приложениями; зона задания – информацию о выданных заданиях на исполнение и обработку документа (в виде резолюций, поручений, сопроводительных писем, напоминаний и др.), а зона уведомление – ответную информацию о доставке сообщения, об ошибках приема и интерпретации сообщения, о регистрации полученного документа и др [3]. В предлагаемом подходе к классификации документов учитывается информация как формализованной части документа (заголовок), так и неформализованной (содержание); их совместная обработка позволяет повысить достоверность классификации документов по технологическим процессам. Поскольку в тексте документа достаточно информации для отнесения его специалистом к нужному процессу, возможно создать систему автоматизированной классификации, которая будет повторять ассоциации, возникающие у человека при чтении документа, с процессом, к которому документ относится. В данной работе предлагается метод классификации, не требующий явного построения модели, описывающей действия человека, вместо этого предлагается обработать уже классифицированные вручную документы, выявить признаки, позволяющие однозначно классифицировать документы по технологическим процессам и построить систему, автоматизирующую данные операции.

Суть предлагаемого подхода к построению модели заключается в выделении для документов каждого технологического процесса такого набора терминов, наличие произвольного количества которых в классифицируемом документе позволит сделать вывод о принадлежности документа к соответствую-

щему процессу. Такой набор терминов предложено называть моделью документов технологического процесса и представлять в виде объекта, описываемого следующим кортежем:

$Mi = \{T_i\}$,

где M_i – модель документов i-го процесса, T_i – множество уникальных терминов документов i-го процесса. Совокупность моделей документов всех технологических процессов (репозиторий) обозначается как набор моделей отдельных процессов:

$$M = \{M_i, M_n, M_u, i = \overline{1, N}\},$$

где M – репозиторий, M_i, $M_п$, $M_и$ – соответственно модель документа i-го процесса, методы получения и использования данных терминов, N – количество технологических процессов .

Множество уникальных терминов документов i-го процесса T_i является матрицей терминов, выявляемых для каждого процесса с помощью метода получения по следующему алгоритму:

Формируется матрица A терминов, содержащихся в во всех документах, принадлежащих процессу i:

$A = \{A_i\}$,

где A_i – единичный термин, приведенный к нижнему регистру.

В процессе выделения терминов их документов проверяется наличие для аббревиатур соответствующих терминов в словаре аббревиатур, описываемого кортежем:

Аб = {Аб$_j$, Термин$_i$},

где Аб – j-ая аббревиатура, Термин$_j$ – соответствующий ей термин.

Из данного набора терминов исключаются термины, присутствующие в словаре стоп-слов SW, являющимся вектором слов:

$SW = \{SW_k\}$,

где S_k – элемент словаря стоп-слов, таким образом:

$$B_i = \{A_j\}, \forall j \in N : A_j \notin SW$$

Из полученного множества B исключаются термины, частота встречаемости которых менее заданной экспертом:

$$C_i = \{B_j\}, \forall j \in N : F(B_j) \geq F_{min},$$

где $F(Bj)$ – функция, определяющая частоту вхождения термина в документы i-го процесса, F_{min} – задаваемое пользователем пороговое значение минимальной частоты.

Производится формирование матрицы C общих терминов:

$C = \{C_i\}$.

Из полученной матрицы C выбираются те элементы, которые входят не более чем в P матриц C_i:

$$T_i = \{C_{ij}, i, j \in N : K(C_{ij}) \leq P\},$$

где $K(C_{i,j})$ – функция, определяющая частоту вхождения термина $C_{i,j}$ в вектора $C_{k,l}$, $k<>j$, $k, l \in N$.

Таким образом, получается набор T моделей документов процессов T_i, представляющий собой матрицу, по строкам которой расположены процессы, по столбцам – уникальные термины документов этих процессов, на основании которых можно сделать вывод о принадлежности документа, содержащего любой из этих терминов, к соответствующему технологическому процессу:

$$T = \{T_i\}, T_i = \begin{Bmatrix} C_{11} & C_{12} & .. & C_{1M} \\ C_{21} & C_{22} & .. & C_{2M} \\ .. & .. & .. & .. \\ C_{N1} & C_{N2} & .. & C_{NM} \end{Bmatrix},$$

где N – количество технологических процессов, M – максимальное количество уникальных терминов документов процессов.

Мощностью вектора уникальных терминов T_i называется количество элементов этого множества:

$S_i = |T_i|$.

Вполне вероятно, что мощности векторов разных строк будут различаться в связи с разным количеством документов по разным процессам и разным количеством уникальных терминов, которые можно в них выделить. Для нормирования результатов классификации документов по процессам с различающимися мощностями моделей предлагается учитывать мощность модели, описание алгоритма приведено ниже.

Метод использования, упоминающийся выше как неотъемлемая часть объекта «модель документа процесса», подразумевает собой определение степени соответствия классифицируемого документа всем моделям репозитория. Задачу классификации документа по процессу можно представить в виде задачи определения весов дуг графа, приведенного на рисунке 1, корневой вершиной которого является множество терминов классифицируемого документа, а остальными вершинами – модели документов процессов, а дуги – степень близости соответствующего узла к корневой вершине.

Определение данных весов предлагается осуществлять по следующему алгоритму:

- формирование вектора терминов классифицируемого документа;
- вычисление степени близости данного вектора к каждой модели репозитория;
- выбор модели-решения на основании значения весов вершин.

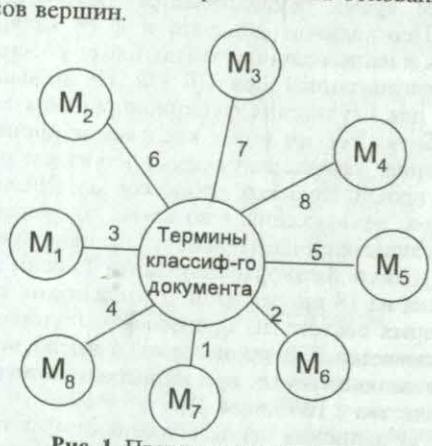

Рис. 1. Представление задачи классификации в виде графа

На первом шаге формируется вектор терминов классифицируемого документа по правилам, аналогичным для выделения терминов при формировании моделей документов. Вектор описывается следующим образом:

$$B_i = \{A_j\}, \forall j \in N : A_j \notin SW$$

На втором шаге производится вычисление весов дуг, соединяющих вершины графа с корневой вершиной. Вычисление производится по следующему алгоритму:

- производится перебор всех терминов каждой модели;
- j-й термин текущей модели ищется в векторе терминов классифицируемого документа и подсчитывается количество его вхождений;
- результат поиска делится на количество терминов в модели для нормирования результатов поиска в моделях различной мощности:

$$V_{i,j} = \frac{\sum S(T_j)}{K(T_i)},$$

где V_{ij} – результат поиска j-го термина i-го процесса, $S(T_j)$ – количество вхождений термина T_j в вектор классифицируемого документа, $K(T_i)$ – количество терминов в модели документов i-го процесса;

- итерация повторяется для каждого термина каждой модели.

В результате получится вектор, компонентами которого являются веса дуг графа, характеризующие степень принадлежности классифицируемого документа к каждому процессу:

$V = \{V_i\}$.

На третьем шаге производится выбор тех процессов, которые соответствуют наибольшим значениям вектора V:

$$\exists i, j, V_i, V_j : \frac{V_i}{V_j} >> 1, i \neq j \text{ и } V_i > V_{min},$$

где V_{min} – минимальное значение результата поиска, устанавливаемое экспертом, при котором V_i считается решением.

В связи с тем, что документ может принадлежать нескольким процессам и, соответственно, несколько V_i будут иметь приблизительно равные значения, необходимо определить интервал, в котором значение V_i будет считаться решением

$$\exists i, j, V_i, V_j : V_i >> V_j + const, i \neq j.$$

В случае, если ни один вектор V_i не удовлетворяет приведенным выше условиям, необходимо произвести классификацию данного документа вручную с последующим дообучением системы, т.е. указанием экспертом тех терминов, которые позволили отнести данный документ именно к соответствующему технологическому процессу.

Осуществлена практическая реализация прототипа системы, показавшая возможность произведения автоматизированной классификации. В результате обработки контрольной выборки документов произведена классификация 2574 документов, из них 2316 – правильно, что соответствует 90% и является хорошим показателем как в относительном, так и в абсолютном выражении. Для 203 документов потребовалось вмешательство эксперта для определения терминов, которые затем были добавлены в модели соответствующих процессов. В основном документы, потребовавшие вмешательства, были неинформативными или содержали противоречивые термины.

Полученные при проведении тестовой классификации результаты позволяют сделать вывод о возможности перехода к автоматической классификации.

Решена задача автоматической классификации разнородных объектов, в данном случае – документов по технологическим процессам. Разработана модель документа технологического процесса в виде кортежа из матрицы общих терминов документов процесса и вероятности их появления в документах процессов. Разработан алгоритм автоматической классификации документов с использованием предложенной модели документов процесса. На примере прототипа программного обеспечения показана эффективность предлагаемых подходов к автоматизации классификации документов по технологическим процессам.

СПИСОК ЛИТЕРАТУРЫ

1. **Тимофеев, К. А.** Технология Workflow в управлении ТУ Банка России / Г. Г. Куликов, К. А. Тимофеев, А. А. Хуснутдинов // Повышение функциональной роли банковской системы через улучшение качества ее деятельности. Управление бизнес-процессами в Банке России и кредитных организациях: сб. науч. тр. М. : Наука, 2006. С. 172–175.

2. **Тимофеев, К. А.** Интеграция системы поддержки процессного управления с существующими документообразующими системами // Автоматизированные системы обработки информации и управления : сб. тр. шк. аспирантов. Уфа : УГАТУ, 2006. Т. 1. С. 27–33.

3. **Афанасьев, С. И.** О создании стандартов и протоколов взаимодействия систем автоматизации документационного обеспечения управления [Электронный ресурс] / С. И. Афанасьев // ВКСС Connect. 2004. № 6. (http:// eos.ru/ eos/ 130987).