Об одной реализации системы распознавания факсов

О. А. Славин

В работе рассмотрена система распознавание факсов, базирующаяся на аппарате теории мультимножеств. Рассмотрена версия системы, развернутая в ООО «Городисский и партнеры». Предложена концепция развития системы, существенно использующая многопоточность и многопроцессность, обеспечивающая повышение быстродействия обработки факсов и дополнительные возможности обработки критических ситуаций.

Введение

Распознавание факсов является достаточно популярной задачей [1, 2].

В настоящей работе мы будем рассматривать факсы, содержащие переписку клиентов с организацией, оказывающей услуги в патентовании интеллектуальной собственности в РФ. Особенности автоматизированной системы управления документооборота (АСУД) рассматриваются на примере АСУД ООО «Городисский и партнеры», описанной в [3].

Потоки входящих факсов обладают следующими особенностями:

- переменный состав факса, выражающийся в том, что одноименные свойства располагаются в документе в произвольных местах, а также могут отсутствовать;
- в факсе может присутствовать несколько одноименных реквизитов;
- метаданные, являющиеся характеристиками всего факса в целом, также могут содержать несколько одноименных реквизитов, например, один и тот же документ может быть отправлен по нескольким маршрутам.

То есть факсы, обрабатываемые в АСУД ООО «Городисский и партнеры» являются структурированными документами с переменным составом, определенные в [4]. В [4] было показано, что обработка структурированных документов с переменным составом (поля которых вообще нельзя описать соотношениями с другими реквизитами, а реквизиты не обязаны быть уникальными) может производиться с помощью аппарата теории мультимножеств.

Работа является развитием [3] и [4] и состоит из двух частей. В первой из них приводится теория распознавания документов с переменным составом, применимая к многостраничным факсам. Вторая часть посвящена реализации системы распознавания факсов.

1. Методология распознавания факсов документов с переменным составом

Далее *документом* мы называем как документ с переменным составом, так и факс, являющийся образом документа с переменным составом. В изложении используются определения и понятия теории мультимножеств, взятые из книги [5].

Пусть $F = \{f\}$ — множество *документов*, каждый из которых может существовать в виде множества представлений (графического образа, текста). Представления позволяют находить состав документа, то есть реквизиты, представляющие интерес для последующей обработки документов. В множестве документов F существует «пустой» документ Θ_F, в состав которого реквизиты не входят. Пусть $T = \{t\}$ — множество *типов документов* и задано соответствие *типизации документов* $\tau \subseteq F \times T$.

Каждый документ составляется из реквизитов, принадлежащих множеству $R_0 = \{r_1, \ldots, r_M\}$. Назовем *составом документа* мультимножество $R = \{N_1 \cdot r_1, \ldots, N_M \cdot r_M\}$, носителем которого является множество реквизитов $Supp(R) = R_0$, а кратности $N_1, \ldots, N_M$ соответствуют встречаемости реквизитов в документе. Состав пустого документа Θ_F без реквизитов выглядит следующим образом: $\{0 \cdot r_1, \ldots, 0 \cdot r_M\}$. Для *структурированных документов с переменным составом* не существует ограничений на кратности реквизитов, а также не рассматриваются предположения об отношениях между геометрическими характеристиками реквизитов в документе.

Пусть f_1 и f_2 — два документа, содержащих реквизиты $\mathrm{Re}(f_1) \subseteq R$ и $\mathrm{Re}(f_2) \subseteq R$. *Суммой документов* f_1 и f_2 назовем документ, обладающий мультимножеством реквизитов, равным $\mathrm{Re}(f_1) + \mathrm{Re}(f_2)$. На определении суммы документов основываются алгоритмы поиска реквизитов в многостраничных документах. Многостраничные факсы рассматриваются как суммы документов, являющихся отдельными страницами. Найденные в различных страницах мультимножества реквизитов R_i, суммировались в одно общее мультимножество в соответствии с определением суммы документов:

$$R = \sum R_i.$$

Рассмотрим множество *типов реквизитов* Ω и множество *имен реквизитов* N_R, с помощью которых определяются соответствие *типи-*

зации реквизитов $\omega \subseteq R_0 \times \Omega$, соответствие *идентификации реквизитов* $\eta \subseteq R_0 \times N_R$ и соответствие *разыменования реквизитов* $\theta \subseteq N_R \times R_0$. В множестве типов реквизитов Ω содержится тип Θ_Ω, олицетворяющий случай, когда подобрать тип невозможно.

Каждый реквизит обладает значением и метаданными, то есть дополнительной информацией, необходимой в процессе обработки программными приложениями. Совокупность значения и метаданных реквизита мы будем называть *характеристиками реквизита*. Метаданные документа, то есть информация, которая не может быть отнесена к реквизитам документа, мы будем называть *характеристиками документа*. Характеристики реквизитов, будучи производным понятием, также образуют мультимножество X, характеристики документа — мультимножество Y. Рассмотрим соответствие *характеризации* $\chi \subseteq R_0 \times X$.

Над множеством однотипных документов $O(t) = \{O \mid \tau(o) = t\}$ строится набор *ограничений* над характеристиками отдельных реквизитов и образуемых реквизитами групп. Множество однотипных документов $O(t)$, реквизиты которых идентифицированы соответствием η, и набор ограничений C определяют модель $\langle O(t), \eta, C\rangle$ документов с типом t.

Распознавание структурированного документа определяется как процесс нахождения характеристик (значений и метаданных) реквизитов в документах, а также характеристик документа, в ограничениях фиксированной модели или группы моделей. Иными словами, модель распознавания задается соответствием $F \times (X + Y)$.

В отличие от реквизитов, определяющихся в процессе распознавания, *свойствами* мы будем называть элементы документа и его составных частей, необходимые в системе хранения документов. Например, свойствами является группа полей базы данных, в которой сохраняются результаты распознавания. Типы и значения свойств задаются и контролируются системой хранения документов. Например, для реляционных систем управления базами данных (СУБД) имеют смысл значения типа целых и действительных чисел, символьных строк, дат и времени. В одном документе возможно существование нескольких одноименных свойств, то есть в общем случае свойства образуют мультимножество $V = \{n_1 \cdot v_1, \ldots, n_K \cdot v_K\}$, аналогичное мультимножеству реквизитов R. Свойства обладают характеристиками, аналогичными характеристикам реквизитов, то есть каждому из свойств соответствуют значения и метаданные.

Таким образом, отображение найденных реквизитов на свойства является интерпретацией, правила которой не отражаются в моделях документов $O(t)$. Процесс определения свойств в множестве реквизитов назовем *конвертацией*. *Модель конвертации* задается соответствием $(X + Y) \times V$.

Выделим мультимножество $V' \subseteq V$, в котором часть свойств имеет нулевые кратности, и зафиксируем среди свойств V' несколько *ключевых* свойств. Переупорядочим элементы мультимножества V' таким образом, чтобы первыми стояли N ключевых свойств. Рассмотрим отношение φ между ключевыми и всеми остальными свойствами

$$\varphi = \langle V'_K, V'_{K'} \rangle,$$

$$V'_K = \{n_1 \cdot v_1, \ldots, n_N \cdot v_N\}, \quad V'_{K'} = \{n_{N+1} \cdot v_{N+1}, \ldots, n_{K'} \cdot v_{K'}\}.$$

Наконец, рассмотрим сужения $\varphi_V = \langle V, V'_{K'} \rangle$ отношения φ на подмультимножествах V'_K, в частности, подмультимножества следующего вида $W(i, \rho) = \{V'_K \mid r_i = \text{const}\}$. Отношения φ_V, описывающие зависимости неключевых свойств от ключевых, назовем *справочниками*.

Определение справочников, то есть фиксация мультимножеств V', выбор ключевых свойств V'_K и конструирование отношений φ_V, осуществляется в рамках решения содержательных задач. Целями создания справочного отношения являются:

- описание взаимозависимостей между свойствами, составляющими отношение;
- хранение значений элементов мультимножества в виде пар ⟨код элемента, значение элемента⟩.

Конвертация некоторых результатов носит не формальный характер, а определяется правилами документооборота, входным контуром которого служат средства распознавания документов. В частности, если найдено несколько экземпляров реквизита r (то есть $k_R(r) > 1$, где R — мультимножество реквизитов), то могут быть приняты различные решения, например:

- создание некоторого свойства с кратностью, отличной от 1;
- сохранение $k_R(r)$ различных свойств;
- отказ Θ_Ω от типизации свойства.

Ведение в хранилищах справочников φ_V открывает возможности извлечения с помощью найденных ключевых свойств всех остальных свойств отношения. Это позволяет определять значения ненайденных свойств и повышать надежность определения у найденных. Отметим, что для баз данных реляционного типа понятие отношения является базовым и полностью соответствует отношениям свойств документа.

Алгоритмы преобразования найденных реквизитов в свойства являются финальным этапом сохранения представления документа (графического или текстового) в хранилище. После этого жизненный цикл документа продолжается в виде стадий документооборота или архивного хранения.

2. Реализация распознавания факсов, содержащих документы с переменным составом

На рис. 1 приведена схема обработки факсов, применяемая в ООО «Городисский и партнеры». Образы факсов распознаются, после поиска в результатах распознавания реквизитов и маркировки образа факса производится публикация в БД архива и распечатка. Прикладное ПО, установленное на сервере системы обработки факсов, состоит из нескольких компонент:

— сервис, управляющий запуском и остановкой диспетчера;
— диспетчер, контролирующий состояние RecoUnit-а и PrintUnit-а;
— RecoUnit, который распознает поток образов принятых факсов. В задачи RecoUnit-а входят: распознавания всех страниц образа факса, маркировка образа факса штрих-кодом, поиск реквизитов в результатах распознавания, публикация реквизитов и маркированного образ в архиве, печать с помощью принтюнита;
— PrintUnit, осуществляющий вывод образа факса в очередь принтера.

RecoUnit осуществляет распознавание и поиск реквизитов в каждой из страниц документа в соответствии с алгоритмами, приведенными в [4]. А именно, в образе каждой из страниц производится поиск мультимножества реквизитов, мультимножество реквизитов многостраничного факса является суммой мультимножества реквизитов отдельных страниц. Библиотека доступа к архиву GiPSQL, используемая в RecoUnit-е, реализует как операции извлечения данных из архива, так и отображения мультимножества реквизитов на мультимножество свойств. Маркировка факса, то есть внедрение штрих-кода с уникальным номером, обеспечивается библиотекой, описанной в [6].

Печать образов факсов производится компонентой, использующей библиотеку обработки изображений, например AccuSoft, обмен информацией с другим приложением осуществляется с помощью обмена сообщениями типа WM_COPYDATA.

Наличие нескольких устройств и особенности ПО могут привести к возникновению нескольких критических ситуаций:

— потеря доступа к файловым серверам и серверам БД из-за их собственных неустранимых сбоев и проблем в локальной сети;
— остановка принтера из-за сбоев и отсутствия бумаги;
— теоретически возможные критические ошибки, зависания, удаление из памяти программных компонент;
— превышения тайм-аутов обработки (например, обработки одного факса);
— перезагрузка серверов из-за сбоев электропитания.

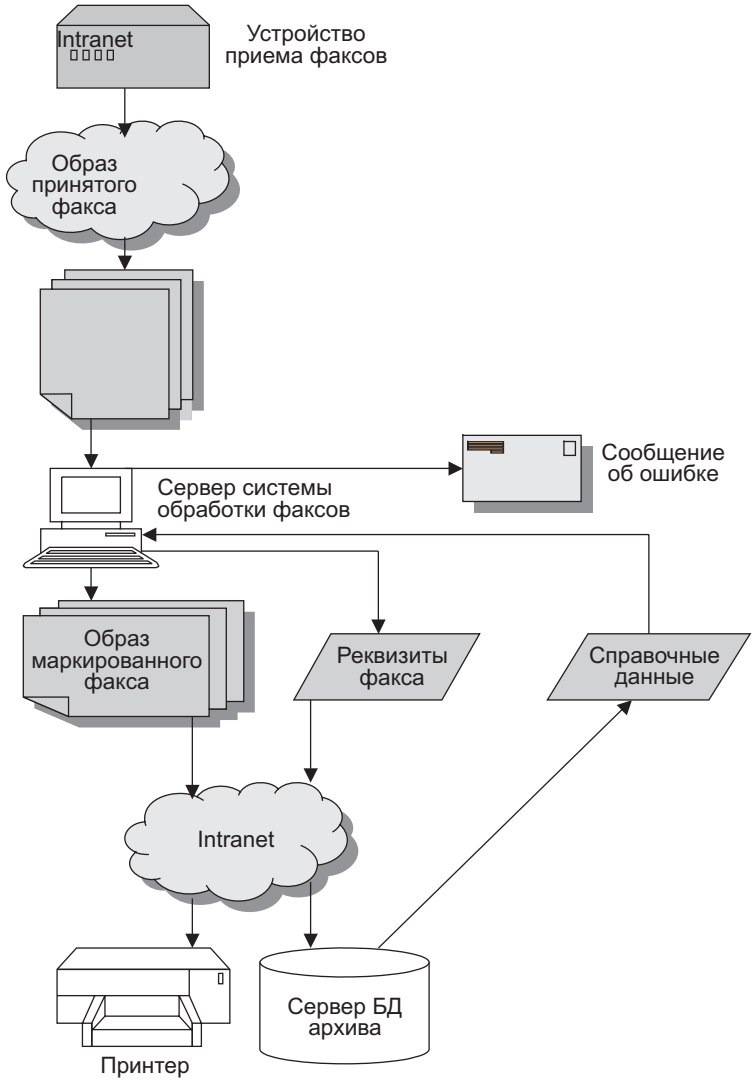

Рис. 1. Обработка факса в АСУД ООО «Городисский и партнеры»

Диспетчер контролирует состояние RecoUnit-а и PrintUnit-а на предмет возникновения приведенных критических ситуаций с помощью обмена сообщениями типа WM_COPYDATA и функции теста на зависание IsHungAppWindow (из библиотеки User32). При перезагрузке компью-

тера, на котором производится обработка факсов, запускается сервис, который инициирует работу диспетчера с восстановленными из контрольной точки параметрами.

Критериями качества работы системы распознавания факсов являются:

— функционирование с максимальной точностью;
— быстродействие;
— восстановление работы при возникновении критических ситуаций.

Недостатки описанной версии системы распознавания факсов состоят в неиспользованных возможностях современных процессоров, позволяющих производить параллельное распознавание на нескольких процессорах или ядрах процессоров. Также неоптимальным является совмещением в RecoUnit-е функций распознавания с операциями доступа к БД архива.

Указанные недостатки могут быть преодолены с помощью новой архитектуры системы распознавания факсов CTFR_GiP (Cognitive Technologies Fax Reader for Gorodissky & Partners), описание которой приводится ниже.

В прикладном ПО CTFR_GiP присутствуют следующие компоненты:

— сервис, функционирование которого по отношению к предыдущей версии не изменилось;
— диспетчер;
— RecoUnit;
— PrintUnit;
— DBUnit, осуществляющий взаимодействие с БД архива.

Рассмотрим работу RecoUnit-а. RecoUnit выполняет операции, связанные с распознаванием, такие как загрузка изображения, установка опций, сегментация страницы, распознавание страницы или блока, передача результатов распознавания в различных представлениях. Взаимодействие с БД архива отсутствует, необходимые подмножества справочников должны передаваться извне вместе с другими опциями. Запуск операции производится в синхронном режиме с помощью функции SendMessageTimeout. Обмен данными ограниченного объема осуществляется посредством статической общей области (#pragma data_seg), создаваемой в одной из библиотек. Обмен данными произвольного объема производится с помощью сообщений типа WM_COPYDATA. Для управления RecoUnit-ом создается отдельная библиотека CTFR_GiP_RecoUnit.dll, обеспечивающая создание, использование и удаление одного RecoUnit-а с помощью «канала». Количество каналов ограничено объемом свободной оперативной памяти, каждый из каналов реализуется в виде отдельного потока. Средства канала контролируют состояние (в том числе

отсутствие) RecoUnit-а, в свою очередь, RecoUnit контролирует отсутствие создавшего его приложения. Характеристики состояния RecoUnit-а хранятся в статической общей области. Каждый из RecoUnit-ов использует следующие библиотеки:

— CT_Puma (распознавание образов);

— CTFR_GiP_DataChange (общая область и обмен данными).

Таким образом, предлагаемая реализация RecoUnit-ов обеспечивает многопоточный режим распознавания с точки зрения приложения, создающего каналы и многопроцессный режим функционирования нескольких RecoUnit-ов. Контроль критических ситуаций упрощается из-за адекватных системных средств и объектного характера реализации каналов.

PrintUnit также реализуется в виде отдельного приложения. Для управления PrintUnit-ом создается отдельная библиотека CTFR_GiP_PrintUnit.dll, обеспечивающая создание, использование и удаление одного PrintUnit-а с помощью канала. Обмен данными осуществляется посредством статической общей области и с помощью сообщений типа WM_COPYDATA. Каждый из PrintUnit-ов использует следующие библиотеки:

— CTFR_Imp (обработка изображений);

— CTFR_GiP_DataChange (общая область и обмен данными).

Использование нескольких PrintUnit-ов целесообразно для печати на нескольких принтерах, переключении на дополнительный принтер при отказе основного и т. п.

В предыдущей версии CTFR_GiP для доступа к БД использовалась библиотека GiPSQL, которую в предлагаемой реализации целесообразно организовать в виде отдельного процесса DBUnit, в состав которого будет входить библиотека CTFR_GiP_DataChange. Это позволит более оперативно отслеживать критические ситуации, а также с помощью библиотеки CTFR_GiP_DBUnit.dll организовывать доступ к БД с различными account-ами.

Описание диспетчера начнем с его состава, в который входят библиотеки

CTFR_GiP_RecoUnit.dll, CTFR_GiP_PrintUnit.dll, CTFR_GiP_DBUnit.dll

для организации каналов различного типа, а также библиотеки CT_GLB (маркировка факса штрих-кодом) и CT_TAN (поиск мультимножества реквизитов в результатах распознавания). Количество каналов типа PrintUnit и DBUnit определяется конфигурацией технического обеспечения (количество принтеров, дополнительных серверов с репликациями БД). Количество каналов типа RecoUnit определяется числом ядер процессора. Стратегия работы диспетчера состоит в ускорении распознавания страниц факсов за счет использования нескольких каналов, а также

взаимодействия с дополнительными ресурсами при отказах основных технических средств.

Таким образом предлагаемая разбивка системы распознавания факсов на модули и организация межмодульного взаимодействия обеспечивает гибкость взаимодействия с техническими средствами, повышение быстродействия и упрощение в обработке критических ситуаций.

Выводы

Предыдущая версия системы распознавания факсов CTFR успешно развернутая в ООО «Городисский и партнеры», содержит следующие программные компоненты: сервис, диспетчер, RecoUnit и PrintUnit.

Предлагаемая реализация CTFR сохранит достоинства предыдущей версии и добавит ряд новых свойств, обеспечивающих повышение быстродействия обработки факсов и дополнительные возможности обработки критических ситуаций.

Литература

1. *Цопкало Н. Н.* Разработка и исследование методов и средств распознавания текста факсимильных сообщений // Автореферат на соискание ученой степени к. т. н., Таганрог, 2002.
2. *Дудушкин С. В.* Документооборот в юридических кругах (интервью) // Intelligent Enterprise (Корпоративные системы). 2004. № 2(91).
3. *Постникова М. В.*, *Славин О. А.* Развитие концепции электронного документооборота на примере системы документооборота патентного ведомства // Сборник трудов ИСА РАН «Организационное управление и искусственный интеллект». М.: УРСС, 2003. С. 30–51.
4. *Славин О. А.* Алгоритмы распознавания структурированных документов с переменным составом // Программирование. 2005. № 4. С. 1–18
5. *Петровский А. Б.* Основные понятия теории мультимножеств. М.: УРСС, 2002.
6. *Славин О. А.*, *Федоров Г. О.* Об использовании штрих-кодирования и специализированных устройств в корпоративном электронном документообороте // Сборник трудов ИСА РАН «Организационное управление и искусственный интеллект». М.: УРСС, 2003. С. 185–197.