УДК 002.66

АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ ФОРМАЛИЗОВАННЫХ ДОКУМЕНТОВ В СИСТЕМЕ ЭЛЕКТРОННОГО ДОКУМЕНТООБОРОТА

Носенко Сергей Владимирович

Королев Игорь Дмитриевич
д.т.н., проф.
Филиал Военной академии связи (г. Краснодар), Краснодар, Россия

В данной статье представлен способ автоматической классификации электронных документов, основанный на анализе метаданных документа при помощи алгебры конечных предикатов

Ключевые слова: ЭЛЕКТРОННЫЙ ДОКУМЕНТООБОРОТ, ЭЛЕКТРОННЫЙ ДОКУМЕНТ, ФОРМАЛИЗОВАННЫЙ ДОКУМЕНТ, ТЕОРИЯ КОНЕЧНЫХ ПРЕДИКАТОВ, СЛОВОФОРМА

UDC 002.66

AUTOMATIC CLASSIFICATION OF FORMAL DOCUMENTS IN ELECTRONIC DOCUMENT MANAGEMENT SYSTEM

Nosenko Sergey Vladimirovich

Korolyov Igor Dmitrievich
Dr.Sci.Tech., professor
Krasnodar branch of Military Academy of network, Krasnodar, Russia

This article presents a method for automatic classification of electronic documents, based on the analysis of document metadata with using the final predicates algebra

Keywords: ELECTRONIC DOCUMENT MANAGEMENT SYSTEM, ELECTRONIC DOCUMENT, FORMAL DOCUMENT, THEORY OF FINAL PREDICATES, WORD-FORM

Автоматическая классификация документов используется в автоматизированных системах (далее – АС) электронного документооборота, базах данных, электронных хранилищах (электронных архивах) в случаях, когда существует необходимость классификации формализованных документов, поступающих из внешних АС, по тематическим признакам, формам (структурам), значениям реквизитов документов.

Описываемый в статье способ заключается в описании документов с помощью математического аппарата теории конечных предикатов. Способ предназначен для:

реализации возможности классификации формализованных документов по формам;

повышения оперативности извлечения заданных метаданных;

повышения оперативности извлечения метаданных документа, позволяющих классифицировать документ по информационным областям

за счет проведения анализа не всего содержимого документа, а только его информативной части;

определения относимости документа к информационной области с обеспечением возможности априорного задания информационных областей, к которым относится электронный документ, в том числе с учетом всевозможных взаимосвязей таких информационных областей.

Результат классификации формализованных документов по формам(структурам), а также повышение оперативности выделения метаданных (в том числе информативной части документа) получаем за счет того, что осуществляем выделение характеристик одинаковых участков текста $Z=\{z_1, z_2,\ldots, z_n\}$ (реквизитов) формализованного документа. Реквизиты выражаем конечным предикатом, где T –множество характеристик текста t, $L=\{l_1, l_2,\ldots, l_q\}$ – множество конечных предикатов узнавания ключевых слов реквизита l, q – количество всех используемых ключевых слов[1].По количеству используемых реквизитов документов вАС создаем систему конечных предикатов для идентификации всех реквизитов.

Правило построения предиката узнавания реквизита формализованного документа, выразится следующей формулой [1]:

где t_h^a – предикат узнавания значения ah-той переменной текста; m – количество переменных текста, n – величина алфавита h-той переменной текста; l_i^b – предикат узнавания значения b ключевого слова соответствующего i-той зоне.

ГОСТ Р 6.30-2003 подразумевает перечень 30 реквизитов документов. Вместе с этим, некоторые реквизиты не определяют индивидуальность формы документа, например те, которые свойственны всем формам документов (например – текст) или вообще не свойственны в данных условиях применения (например – Государственный герб

Российской Федерации в частной организации), что приводит к еще большему сокращению размеров системы предикатов узнавания реквизитов. Количество различных используемых реквизитов определяет размерность системы предикатов узнавания реквизитов.

Форму(структуру) документа выражаем конечным предикатом $P_{\text{ф}}(V, Z, L)$, где $V=\{v_1, v_2,\ldots, v_m\}$ – множество форм документа, m – количество всех используемых форм документов, $Z=\{z_1, z_2,\ldots, z_n\}$ – множество конечных предикатов узнавания реквизитов документа, n – количество реквизитов документов, $L=\{l_1, l_2,\ldots, l_q\}$ – множество ключевых слов, q – количество всех используемых ключевых слов. По количеству используемых форм документов вАС создаем систему конечных предикатов для идентификации всех форм.

Правило построения предиката узнавания формы документа выразится следующей формулой [1]:

где – i-тый предикат узнавания реквизита документа системы предикатов узнавания реквизитов; $l_j^{\text{с}}$ - предикат узнавания уникального значения ключевого слова, соответствующего j-той форме документа.

Созданная система предикатов с использованием (2) применяется в АС для классификации формализованных документов по формам (структурам).

Форма документа однозначно задает места расположения реквизитов документа и область поиска значений конкретного реквизита ограничивается областью имеющую положительное значение соответствующего предиката из системы предикатов, построенной по (1) такая система предикатов используются вАС для повышения

оперативности выделения метаданных (в том числе информативной части документа).

Вышеизложенное позволяет:

реализовать возможность классификации формализованных документов по формам за счет однозначного определения формы документа;

повысить оперативность извлечения заданных метаданных, за счет определения области поиска значений реквизитов, например, из списка возможных значений при малом словаре значений реквизитов или по маске реквизита;

выделять информативную часть документа (например – текстна естественном языке) для последующего анализа не всего содержимого документа, а только информативной части с целью повышения оперативности отнесения документа к той или иной информационной области.

С целью реализации возможности классификации по информационным областям формализованных документов слова текста на естественном языке информативной части документа преобразуем в базовые словоформы, отбросим незначимые слова, осуществим подсчет весов слов в тексте в соответствии с частотами их появления и тем самым сформируем предикаты узнавания информационной области.

Правило построения системы предикатов $P(U,W)$ узнавания информационной области $u_j \in U = \{u_1, u_2, \dots, u_s\}$; s – количество информационных областей АС, выразится следующей формулой:

$$P(U,W) = \bigwedge_{\forall w_i \sim u_j} \bigvee_{\forall f \sim u_j} w_{ig}^{f}, \qquad (3)$$

где w_{ig}^{f} – предикат узнавания значения веса f значимого слова $w_i \in W = \{w_1, w_2, \dots, w_p\}$ – множество значимых слов текстов, в тексте

документа d^{u_f}-той информационной области по g-тому значению веса значимого слова; p – количество значимых слов текстов.

На этапе обучения системы по предъявленному набору классифицированных вручную текстов сформируем систему предикатов идентификации признаков текста, где количество предикатов в системе предикатов определяется количеством информационных областей, на которые необходимо классифицировать документы (количество исполнителей или пользователей в автоматизированной системе).

На этапе работы системы, при классификации текста на естественном языке, преобразуем слова текста в базовые словоформы, отбросим незначимые слова, осуществим подсчет весов слов в тексте, получившиеся значения подставим в систему предикатов, построенных по (3) на этапе обучения. По предикатам в системе предикатов принявшим значение истинности «1» определим принадлежность к соответствующей информационной области или областям. При этом, в случае необходимости использования априорной информации о зависимостях информационных областей друг от друга, например для: задания дерева информационных областей; создания составной области знаний; исключения части области знаний, нет необходимости проводить этап обучения вновь, а используя алгебру конечных предикатов [2], проводим полный спектр операций над логическими выражениями, а соответственно и над информационными областями, описанными конечными предикатами (добавление, исключение, сложение информационных областей и т.д.). Данный способ классификации позволяет с учетом этого по входному документу определить, каким информационным областям он принадлежит, а каким нет.

Вес f^{w_i} словоформы в тексте документа d_j, рассчитаем по формуле:

$$f_{w_i d_j} = \frac{c_{w_i d_j}}{N_{d_j}}, \qquad (4)$$

Здесь $c_{w_i d_j}$ - количество раз, которое w_i-я словоформа встречается в d_j-м тексте документа, N_{d_j} - общее количество словоформ в i-м тексте документа.

Информативная часть документа для классификации должна быть представлена в виде, допускающем выделение из нее текстового содержания. Каждый документ (либо обучающий, либо подвергающийся классификации) предварительно проходит стадию первичной обработки, на которой производится определение формата документа и установление того, возможно ли извлечение текста из документа данного формата. После разбиения текста на слова определяем для каждого слова его базовую словоформу по одному из способов [3-6]. Наиболее часто для решения подобных задач используется алгоритм Портера [6], заключающийся в использовании специальных правил отсечения и замены окончаний слов.

Согласно предлагаемому способу каждый документ d_i представляем декартовым произведением переменных из множеств $T \times L \times W$, где для инициализации классификатора и построения классификационных признаков служит этап обучения классификатора. При этом должно быть задано множество обучающих документов, заранее классифицированных вручную. После извлечения из них текстового содержания производим построение словаря значимых слов. Словарь содержит базовые словоформы всех слов, встречающихся в обучающих документах.

При классификации документа в расчет берутся не все словоформы из словаря документов, а лишь те из них, которые входят в рабочий словарь классификатора данной информационной области (данного исполнителя), что и использует (3). В рабочий словарь классификатора включаются наиболее информативные словоформы с точки зрения определения принадлежности документа данной категории, не попавшие в

стоп-словарь. Информативность словоформы w_i для классификатора по информационной области u_j определяется по известной формуле [7]:

$$H(w_i, u_j) = \sum_{w \in \{w_i, \bar{w}_i\}} \sum_{u \in \{u_j, \bar{u}_j\}} P(uw) \log \frac{P(uw)}{P(u)P(w)}, \quad (5)$$

При этом устанавливается порог информативности ε; в рабочий словарь классификатора включаются все словоформы, не попавшие в стоп-словарь, информативность которых превышает этот порог. Стоп-словарь состоит из словоформ, частоты встречаемости которых во множестве обучающих документов превышают заранее установленный порог δ. При этом отсекаются слова, не несущие смысловой нагрузки, такие как предлоги, союзы, вводные и общие слова и т.д. Значения коэффициента δ, согласно данному способу, устанавливаются в пределах от 0.05 до 0.7 в зависимости от специфики использования способа. Значения порога информативности δ могут быть различны в различных условиях использования способа.

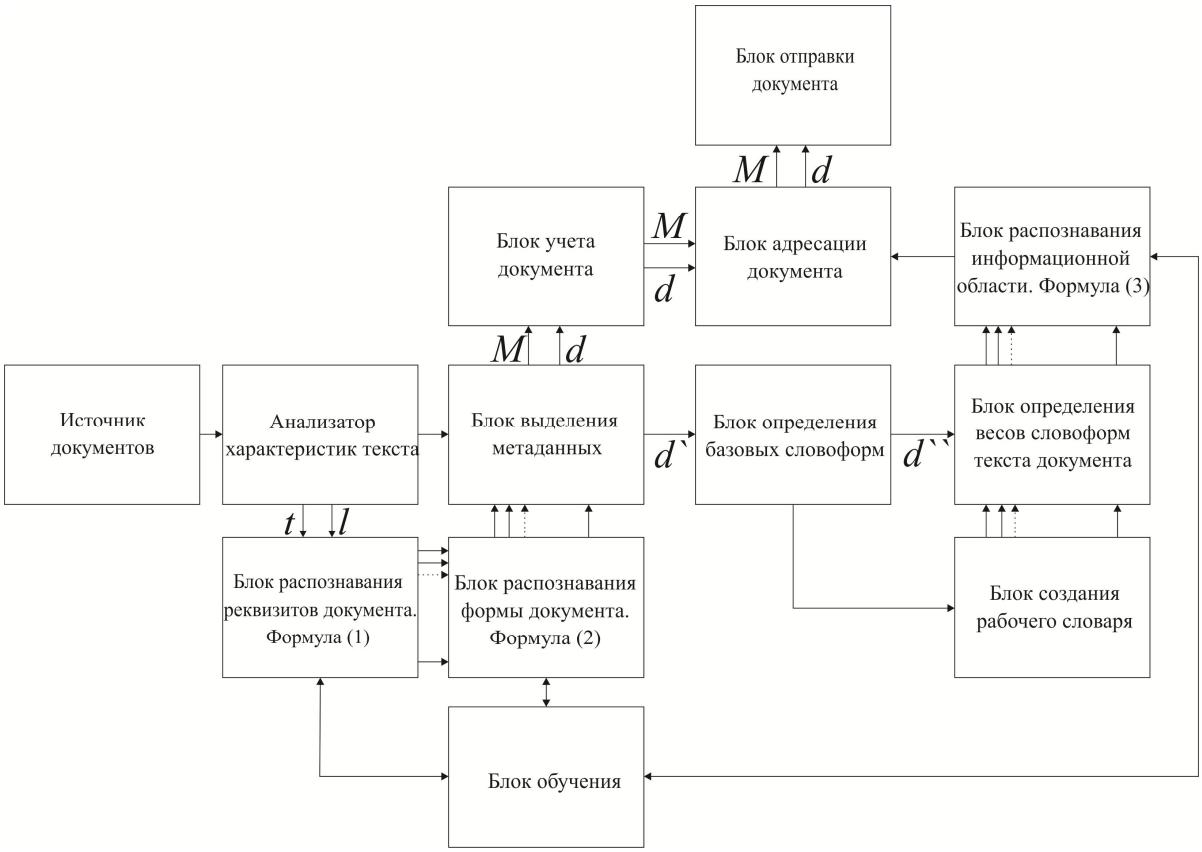

Рис. - Блок-схема для реализации способа автоматической классификации формализованных документов в системе электронного документооборота.

Классификация текстов (информативных частей) документов производится путем вычисления значений системы предикатов, описывающей информационные области. Система предикатов строится по (3).

На рис. представлена блок-схема для реализации способа.

Описание работы:

В режиме классификации.

При появлении в источнике документов нового документа он поступает в блок анализа характеристик текста, который выявляет значения переменных t участков документа и ключевых слов l в них. Значения t и l участков документа поступают в блок распознавания

реквизитов документа, где с помощью системы предикатов, построенных по (1) распознаются реквизиты документа. Информация о распознанных реквизитах документа поступает в блок распознавания формы документа, где система предикатов, построенная по (2) осуществляет распознавание.

В блоке выделения метаданных из поступившего документа от анализатора текста, используя сведения об определенной форме документа из блока распознавания формы документа, которая однозначно задает места расположения значений реквизитов документа, выделяются требуемые значения реквизитов, которые являются метаданными документа. Документ и соответствующие ему метаданные поступают в блок учета документа, и организуется хранение его эталонной копии. Также однозначно определенная в блоке выделения метаданных информативная часть документа поступает в блок определения базовых словоформ. Полученные словоформы поступают в блок создания рабочего словаря из значимых слов по (5).

Полученные словоформы документа, попавшие в рабочий словарь, поступают в блок определения весов слов документа (4), где производится расчет весов f словоформ информативной части документа. Далее значения весов полученных словоформ поступают в блок распознавания информационной области u_i путем вычисления значений предикатов системы предикатов, построенной по (3).

Поступившему документу и метаданным из блока учета документов в блок адресации документов, с использованием полученных значений из блока распознавания информационной области присваиваются соответствующие адреса (классификация в соответствии с информационной областью).

В режиме обучения.

Режим обучения системой используется в трех случаях:

в случае невозможности распознавания системой предикатов реквизитов документа в блоке распознавания реквизитов документа по значениям переменных документа *t* и *l*. В этом случае оператором системы через блок обучения вносятся изменения в систему предикатов блока распознавания документов или определяется реквизит документа «вручную»;

в случае невозможности распознавания системой предикатов формы документа по значениям системы предикатов блока распознавания реквизитов документа. В этом случае оператором системы через блок обучения вносятся изменения в систему предикатов блока распознавания формы документа или определяется форма документа «вручную»;

в случае невозможности распознавания системой предикатов информационной области по значениям весов значимых слов из рабочего словаря, извлеченных из информативной части документа. В этом случае оператором системы через блок обучения вносятся изменения в систему предикатов блока распознавания информационной области или определяется информационная область документа «вручную».

Таким образом, способ позволяет классифицировать документы с учетом любых значений реквизитов, анализировать только информативную часть документа с учетом априорных зависимостей между информационными областями, что достигается результатом классификации по информационным областям и использованием алгебры конечных предикатов.

Список литературы:

1. Подходы к оперативной идентификации формализованных электронных документов в автоматизированных делопроизводствах / И.Д. Королев, С.В. Носенко // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета (Научный журнал КубГАУ) [Электронный ресурс]. – Краснодар: КубГАУ, 2013. – №08(092). – IDA [article ID]: 0921308074. – Режим доступа: http://ej.kubagro.ru/2013/08/pdf/74.pdf, 0,875 у.п.л.

2. М.Ф. Бондаренко, Ю.П. Шабанов-Кушнаренко. Об алгебре конечных предикатов. [Текст]// Научно-технический журнал «Бионика интеллекта». ХНУРЭ, г. Харьков, Украина – 2011 № 3(77).

3. Porter M.F. "An algorithm for suffix stripping", Program, Vol.14, No.3, 1980, pp.130-137.

4. Пат. 2096825 Российская Федерация, МПК G 06 F 17/00, G 06 F 17/30. Устройство обработки информации для информационного поиска [Текст] / Ковалев М.В., Виргунов И.В., Наймушин И.А., Четверов В.В.; заявитель и патентообладатель Общество с ограниченной ответственностью "Информбюро". - № 96119820/09; заявл. 14.10.96; опубл. 20.11.97, Бюл. № 14.

5. Пат. 6308149 Соединенные Штаты Америки, МПК G 06 F 17/27. Grouping words with equivalent substrings by automatic clustering based on suffix relationships [Текст] / Gaussier E., Grefenstette G., Chanod J.-P.; заявительипатентообладательXerox Corporation. - № 09/213309; заявл.16.12.98; опубл. 23.10.01.

6. Пат. 6430557 Соединенные Штаты Америки, МПК G 06 F 017/30; G 06 F 017/27; G 06 F 017/21. Identifying a group of words using modified query words obtained from successive suffix relationships [Текст] / Gaussier E., Grefenstette G., Chanod J.-P.; заявительипатентообладательXerox Corporation. - № 09/212662; заявл.16.12.98; опубл.06.08.02.

7. Craven M., DiPasquo D., Freitag D. et al. "Learning to construct knowledge bases from the World Wide Web", Artificial Intelligence, Vol.118(1-2), 2000, pp.69-113.

References:

1. Podhody k operativnoj identifikacii formalizovannyh jelektronnyh dokumentov v avtomatizirovannyh deloproizvodstvah / I.D. Korolev, S.V. Nosenko // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta (Nauchnyj zhurnal KubGAU) [Jelektronnyj resurs]. – Krasnodar: KubGAU, 2013. – №08(092). – IDA [article ID]: 0921308074. – Rezhim dostupa: http://ej.kubagro.ru/2013/08/pdf/74.pdf, 0,875 u.p.l.

2. M.F. Bondarenko, Ju.P. Shabanov-Kushnarenko. Ob algebre konechnyh predikatov. [Tekst]// Nauchno-tehnicheskij zhurnal «Bionika intellekta». HNURJe, g. Har'kov, Ukraina – 2011 № 3(77).

3. Porter M.F. "An algorithm for suffix stripping", Program, Vol.14, No.3, 1980, pp.130-137.

4. Pat. 2096825 Rossijskaja Federacija, MPK G 06 F 17/00, G 06 F 17/30. Ustrojstvo obrabotki informacii dlja informacionnogo poiska [Tekst] / Kovalev M.V., Virgunov I.V., Najmushin I.A., Chetverov V.V.; zajavitel' i patentoobladatel' Obshhestvo s ogranichennoj otvetstvennost'ju "Informbjuro". - № 96119820/09; zajavl. 14.10.96; opubl. 20.11.97, Bjul. № 14.

5. Pat. 6308149 Soedinennye Shtaty Ameriki, MPK G 06 F 17/27. Grouping words with equivalent substrings by automatic clustering based on suffix relationships [Tekst] / Gaussier E., Grefenstette G., Chanod J.-P.; zajavitel'ipatentoobladatel'Xerox Corporation. - № 09/213309; zajavl.16.12.98; opubl. 23.10.01.

6. Pat. 6430557 Soedinennye Shtaty Ameriki, MPK G 06 F 017/30; G 06 F 017/27; G 06 F 017/21. Identifying a group of words using modified query words obtained from successive suffix relationships [Tekst] / Gaussier E., Grefenstette G., Chanod J.-P.; zajavitel'ipatentoobladatel'Xerox Corporation. - № 09/212662; zajavl.16.12.98; opubl.06.08.02.

7. Craven M., DiPasquo D., Freitag D. et al. "Learning to sonstruct knowledge bases from the World Wide Web", Artificial Intelligence, Vol.118(1-2), 2000, pp.69-113.