УДК 004.931

АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ СЛАБОСТРУКТУРИРОВАННЫХ ДОКУМЕНТОВ, УЧАСТВУЮЩИХ В НАУЧНО-ОБРАЗОВАТЕЛЬНОМ ПРОЦЕССЕ

А. М. Гудов, С. Ю. Завозкин, В. А. Шевнин

AUTOMATIC CLASSIFICATION OF SEMISTRUCTURED DOCUMENTS IN SCIENTIFIC AND EDUCATIONAL PROCESS

A. M. Gudov, S. Yu. Zavozkin, V. A. Shevnin

Работа выполнена в рамках задания № 2014/64 на выполнение государственной работы «Организация проведения научных исследований».

Ежедневно в научно-образовательном процессе любого учебного учреждения используется множество слабоструктурированных документов. Одним из подходов, позволяющих единообразно обрабатывать такие документы, является работа не с самими документами, а с их метаданными. Однако эффективность такого подхода в случае большого числа слабоструктурированных документов может быть достигнута лишь при наличии эффективного, с точки зрения использования вычислительных ресурсов, механизма автоматического извлечения метаданных из содержимого документов, который можно разбить на три этапа: определение класса документа; кластеризация документов, класс которых не удалось определить; извлечение метаданных из документа уже известного класса. Данная работа посвящена поиску возможных решений на первом этапе – автоматической классификации слабоструктурированных документов. В работе введено понятие слабоструктурированного документа, представлены критерии эффективности методов классификации, проведен сравнительный анализ методов в соответствии с первыми пятью критериями. Для оценки по дополнительно разработанным двум критериям были реализованы методы: многослойные нейронные сети, Роккио, k-ближайших соседей. Результаты проведенного анализа показали, что наибольшую эффективность при решении данной задачи с точки зрения соотношения точность/скорость показывают нейронные сети, но точность классификации на слабоструктурированных документах не является достаточной. Выдвинута гипотеза, что точность методов можно повысить, используя при классификации не только ключевые слова, но и известную структуру документа.

Numerous semi-structured documents are used daily in education and research activities at universities. Dealing with metadata rather than documents themselves is one of the ways of processing documents uniformly. However, as far as many semi-structured documents are concerned, this method is considered to be efficient only in case of the existing procedure of automatic extraction of documents content metadata. The procedure includes 3 stages: document class identification, clusterization of the documents whose classes could not be identified, extraction of metadata from the documents of identified classes. The paper is dedicated to possible solutions for the first stage, i.e. automatic classification of semi-structured documents. The paper includes the definition of a semi-structured document, criteria of methods efficiency classification, comparative analysis of different methods regarding 5 top criteria. To estimate 2 additionally developped criteria the following methods are used: multilayer neural networks, Rocchio algorithm, k-nearest neighbor method. Based on the analysis results, the neural networks method appears to be the most efficient in the context of accuracy and speed correlation. However, classification accuracy is not enough when dealing with semi-structured documents. The authors suppose the accuracy of the methods can be improved by using not only key words but also determined document structure during classification process.

Ключевые слова: слабоструктурированные документы, метаданные, автоматическое извлечение, автоматическая.

Keywords*:* semi-structured documents, metadata, automatic extraction, automatic.

На современном этапе развития мирового сообщества информация является стратегическим ресурсом, таким же, как традиционные материальные или энергетические ресурсы. При этом становление информационного общества уже немыслимо без использования информационных ресурсов (ИР) в электронном виде. Наибольший экономический и социальный успех сегодня сопутствует странам, активно развивающим и использующим современные информационные технологии и системы управления ИР. Это направление является приоритетным для многих развитых стран, в том числе и России, где оно отнесено к критическому направлению развития общества.

В настоящее время используется множество систем, обеспечивающих информационную поддержку различных областей деятельности, которые, как правило, работают со строго структурированными данными. Однако существенная часть документов в электронном документообороте многих организаций являются слабоструктурированными. К примеру, ежедневно в научно-образовательном процессе учебного учреждения используется множество слабоструктурированных документов – решения, рефераты, отчеты, статьи, тезисы, курсовые и дипломные работы, электронные письма, рабочие и учебные планы и т. д. Работа с такими документами вызывает целый ряд проблем, связанных со сложностью и трудозатратностью поиска, классификации и обработки слабоструктурированных данных.

Одним из решений описанной проблемы является обработка метаданных документа известной структуры

и, основываясь на ее результатах, осуществление основных действий с данными. Однако эффективность такого подхода в случае большого числа слабоструктурированных документов может быть достигнута лишь при наличии механизма автоматического извлечения метаданных из содержимого документа, который позволяет снизить затраты на обработку новых поступающих в систему документов. В частности, для учреждения высшего образования механизм автоматического извлечения метаданных позволит ускорить процессы регистрации и обработки документов: в системе, обеспечивающей информационную поддержку проведения приемной компании (паспортные данные, заявление и т. д.); в системе электронного документооборота (письмо, приказ и т. д.); в системе информационной поддержки научной деятельности (публикации, грамота за победу в конференции и т. д.); в системе информационной поддержки образовательной деятельности (лабораторная работа, сочинение и т. д.); в системе информационной поддержки библиотеки (новые литературные поступления и т. д.).

Еще одно преимущество применения механизма автоматического извлечения метаданных – возможность компьютеризированного анализа текста, в том числе и интеллектуального. Рассмотрим типовую ситуацию, когда на электронный почтовый ящик кафедры отправлено письмо, содержащее документы с результатами выполнения самостоятельной работы студента очной или очно-заочной формы обучения. После автоматического извлечения метаданных становится возможной регистрация и обработка каждого документа в соответствующей информационной системе с заданными для данного класса документов регламентом.

Метода для универсального решения задачи извлечения метаданных на сегодняшний день не существует [2]. Поэтому чрезвычайно актуальным является разработка общих научно-методологических подходов к решению указанной проблемы.

Основной проблемой для автоматических методов извлечения метаданных является достижение приемлемой точности определения как самих метаданных, так и их значений. Определение путей повышения скорости работы механизма извлечения метаданных при сохранении указанной точности получаемых значений является основной целью предлагаемой работы. Один из способов достижения поставленной цели – первоначальная классификация документа, а затем извлечение значений метаданных из документа определенного класса.

Таким образом, процесс автоматического извлечения метаданных из документов можно разбить на следующие этапы.

1. Автоматическая классификация документа. Данный шаг обоснован тем, что разные типы документов содержат различные наборы и определения метаданных.

2. Кластеризация документов, класс которых не удалось определить на предыдущем шаге. Данный этап требуется для того, чтобы в процессе классификации выделять документы тех классов, которые не участвовали в первоначальном обучении алгоритмов классификации.

3. Непосредственное извлечение метаданных из документа определенного класса.

Был проведен сравнительный анализ большинства известных методов классификации документов:

а) опорные вектора (О);
б) «наивный» алгоритм Байеса (Б);
в) метод k-ближайших соседей (БС).
г) многослойные нейронные сети (НС);
д) метод Роккио (Р);
е) модифицированный метод k-ближайших соседей (МБС).

Описание методов (а – г) и (д – е) можно найти соответственно в работах [6; 7].

Для измерения эффективности методов при решении задачи извлечения метаданных использовались следующие характеристики, разработанные на основе работы [6]: 1) возможность модификации метода для учета структуры документа; 2) алгоритмическая сложность; 3) независимость скорости классификации от числа классифицированных документов; 4) возможность оценки степени близости к классу; 5) возможность классификации на уровне словосочетаний; 6) точность классификации; 7) скорость классификации.

Для каждой из характеристик определен способ измерения и шкала принимаемых ею значений. Характеристики (1 – 5) были измерены на основе анализа научной литературы. Соответствующие оценки представлены в таблице 1.

Полученные результаты позволили сделать вывод о том, что методы (г – е) наиболее эффективны при решении рассматриваемой задачи [3]. Для измерения критериев (6 – 7) анализа литературы оказалось недостаточно и методы (г – е) были реализованы в исполняемом коде.

Все рассматриваемые методы относятся к числу обучаемых – для них необходимо задать исходный набор заранее классифицированных документов, который изначально делится на две части: обучающее и тестовое множества. Эти множества были сформированы из 300 структурированных и слабоструктурированных документов.

Под структурированными в работе понимаются документы, структура которых позволяет однозначно определить необходимый набор метаданных. В качестве примера таких документов можно привести: распоряжение, приказ, служебная записка, заявление, объяснительная, отзыв на выпускную работу, рабочая программа.

Под слабоструктурированными в работе понимаются документы, структура которых не позволяет однозначно определить необходимый набор метаданных. Использовались следующие типы слабоструктурированных документов: реферат, отчет по лабораторной работе, дипломная работа, курсовая работа, отзыв на автореферат.

Таблица 1

Сравнительный анализ методов классификации документов

Характеристики эффективности	*Шкала, балл*	*Методы классификации*					
		БС	**О**	**Б**	**НС**	**Р**	**МБС**
Возможность модификации метода для учета структуры документа	1 – да; 0 – нет.	1	1	0	1	1	1
Алгоритмическая сложность	1 – алгоритмическая сложность O(n); 0 – алгоритмическая сложность O(k*n).	0	1	1	1	1	1
Независимость скорости классификации от числа классифицированных документов	1 – скорость классификации не зависит от числа уже классифицируемых документов; 0 – скорость классификации падает при увеличении количества классифицируемых документов.	0	1	1	1	1	1
Возможность оценки степени близости к классу	1 – метод позволяет оценить вероятность принадлежности к классу; 0 – метод не позволяет оценить вероятность принадлежности к классу.	1	0	1	1	1	1
Возможность классификации на уровне словосочетаний	1 – метод позволяет учитывать словосочетания; 0 – метод не позволяет учитывать словосочетания.	1	1	0	1	1	1
Итоговая оценка		**3**	**4**	**3**	**5**	**5**	**5**

Все исследуемые методы работают с векторами из пространства признаков R^n, где n – количество признаков, которыми характеризуется каждый документ. Это приводит к необходимости преобразования документов к векторному виду, для чего каждому слову документа ставится в соответствие координата из пространства признаков (вес слова в контексте документа), значение которой находилось с помощью частотной характеристики TF-IDF [4]:

$$w_i = \frac{n_i}{N}\left(1 - \frac{l_i}{L}\right), \qquad (1)$$

где w_i – вес i-того слова в контексте документа; n_i – частота встречаемости слова в документе; N – общее количество слов в документе; l_i – количество документов из обучающей выборки, в которых встречается i-слово; L – количество документов в обучающей выборке.

Учет всех слов, встречаемых в документах из обучающей выборки, приводит к высокой размерности пространства признаков, что может быть причиной низкой скорости работы алгоритмов. С целью сокращения размерности пространства признаков для классификации документов использовались лишь такие ключевые слова, которые встречаются не менее чем в 75 % документов одного класса из обучающей выборки. Кроме того, все рассматриваемые слова приводились к нормальным формам (именительный падеж, единственное число) с помощью словарей «iSpell» [5], что позволило объединить одинаковые слова, имеющие различные морфологические формы, в одну координату пространства признаков.

Таким образом, алгоритм приведения документов к векторному виду состоит из следующих шагов.

1. Из каждого класса документов, входящих в обучающую выборку, извлекаются и приводятся к нормальным формам ключевые слова.
2. Из множества полученных ключевых слов удаляются те, которые встречаются менее чем в 75 % документов рассматриваемого класса, что позволяет избавиться от слов, не характерных для документов класса.
3. Множества ключевых слов, которые были извлечены из каждого класса документов обучающей выборки, объединяются в одно общее множество.
4. Из полученного множества удаляются все повторяющиеся ключевые слова.
5. Векторное представление документов (входящих в обучающую выборку или поступающих на вход алгоритма для классификации), получается путем нахождения веса каждого слова из полученного множества в контексте рассматриваемого документа при помощи меры (1).

Представленный алгоритм каждому рассматриваемому документу ставит в соответствие некоторый вектор из векторного пространства размерности n, что позволяет сравнивать между собой различные документы.

При реализации метода, основанного на нейронных сетях, использовалась трехслойная нейронная сеть типа «Персептрон». Первый (входной) слой нейронной сети принимал координаты векторов документов из обучающей выборки, поэтому количество нейронов на данном слое совпадало с количеством координат. Оптимальное число нейронов на втором (скрытом) слое подбиралось в процессе тестирования алгоритма. Каждый нейрон на третьем (выходном) слое отвечал за определенный класс документов, поэтому количество нейронов на данном слое равнялось количеству имеющихся классов. В качестве функции

активации использовалась сигмоидальная функция следующего вида: $out = \frac{1}{1+e^{-sum}}$,

где *out* – выходной сигнал нейрона; *sum* – сумма произведений выходных сигналов нейронов предыдущего слоя на соответствующие весовые коэффициенты данного нейрона.

Для обучения нейронной сети использовался алгоритм обратного распространения ошибки [1]. Обучение прекращалось, если для каждого документа из обучающей выборки выходной сигнал нейрона, соответствующего классу данного документа, принимал значение больше 0.999, в то время как остальные нейроны возвращали сигналы меньше 0.001.

При реализации метода Роккио для каждого класса документов из обучающей выборки находился центроид (типичный представитель класса) по формуле:

$$x_i = \frac{1}{n}\sum_{k=1}^{n} x_i^k ,$$

где x_i – *i*-ая координата центроида; *n* – количество документов в классе; x_i^k – значение *i*-й координаты для *k*-го документа класса. Классифицируемый документ относился к тому классу, к которому принадлежал ближайший к нему центроид.

При реализации классического метода ближайших соседей определялось расстояние от классифицируемого документа до каждого документа из обучающей выборки с целью определения *k* наиболее близких к нему документов. Новый документ относят к самому многочисленному из классов, на которые разбиваются данные *k* документов. В модифицированном методе ближайших соседей хранятся расстояния между всеми документами из обучающей выборки, что позволяет каждый новый документ сравнивать лишь с теми документами из обучающей выборки, которые расположены заведомо близко от него.

В процессе классификации методами (д-е) необходимо находить расстояние между документами, для чего использовались следующие метрики: евклидово расстояние (Е)

$$P_e(x,y) = \sum_{i=1}^{n} x_i y_i ,$$

косинусная мера сходства (К)

$$\cos(a) = \frac{(x,y)}{|x||y|},$$

хеммингово расстояние (Х)

$$P_h(x,y) = \sum_{i=1}^{n} |x_i - y_i| .$$

Выбор для тестирования нескольких метрик был обусловлен предположением, что различные способы нахождения расстояния могут давать разные точности классификации.

Тестирование реализованных методов производилось на компьютере со следующей аппаратной конфигурацией: Intel Core i5 Haswell 3.4 GHz, 128 GB SSD, 4 GB ОЗУ. Результаты анализа представлены в таблице 2.

Таблица 2

Точность классификации рассматриваемых методов

Метод	***Метрика***	***Кол-во нейронов***	***Кол-во соседей***	***Точность классификации на структурированных документах, %***	***Точность классификации на слабоструктурированных документах, %***	***Общая точность, %***	***Время, мс***
НС		24		97,14	84	91,67	8
НС		25		97,14	80	90	11
МБС	Е		5	94,29	84	90	444
МБС	Х		5	94,29	84	90	269
НС		26		100	72	88,33	7
Р	Е			94,29	80	88,33	26
Р	К			91,43	84	88,33	25
МБС	К		5	85,71	76	81,67	389
Р	Х			57,14	64	60	24

Представленные результаты показывают, что наибольшую эффективность при решении данной задачи с точки зрения соотношения точность/скорость показывают нейронные сети, но точность классификации на слабоструктурированных документах не является достаточной. Возможно, точность методов можно повысить, если при классификации использовать не только ключевые слова, но и структуру документа. Для проверки этой гипотезы разрабатываются соответствующие модификации рассмотренных методов с целью сравнения точности и скорости классификации слабоструктурированных документов.

Литература

1. Галушкин А. И. Синтез многослойных систем распознавания образов. М.: Энергия. 1974.

2. Гудов А. М., Завозкин С. Ю., Меньшиков А. С. Модуль автоматического определения метаданных документа в системе электронного документооборота вуза // Вестник КемГУ. 2006. № 1(25). С. 31 – 36.

3. Гудов А. М., Завозкин С. Ю., Шевнин В. А. Автоматическое извлечение метаданных из слабоструктурированных документов, участвующих в научно-образовательном процессе // Информационные технологии и

математическое моделирование (ИТММ-2013): материалы XII Всероссийской научно-практической конференции с международным участием (им. А. Ф. Терпугова), 29 – 30 ноября 2013 г. Ч. I.

4. Кристофер Д. Маннинг, Правхакар Рагхаван, Хайнрих Шютце. Введение в информационный поиск [пер. с англ.] М.: И. Д. Вильямс, 2011.

5. Лебедев А. Словарь русского языка ispell // Кафедра физики полупроводников. 2014. Режим доступа: scon155.phys.msu.su/~swan/orthography.html (дата обращения: 30.01.2014).

6. Пескова О. В. Методы автоматической классификации текстовых электронных документов // НТИ. (Серия 2: Информационные процессы и системы). 2006. № 3. С. 13 – 20.

7. Толчеев В. О. Модифицированный и обобщенный метод ближайшего соседа для классификации библиографически текстовых документов // Заводская лаборатория, диагностика материалов. 2009. Т. 75. № 7. С. 63 – 70.

Информация об авторах:

Гудов Александр Михайлович – доктор технических наук, доцент, заведующий кафедрой ЮНЕСКО по новым информационным технологиям КемГУ, good@kemsu.ru.

Alexander M. Gudov – Doctor of Technical Sciences, Associate Professor, Head of the UNESCO Department for New Information Technologies, Kemerovo State University.

Завозкин Сергей Юрьевич – кандидат технических наук, доцент кафедры ЮНЕСКО по новым информационным технологиям КемГУ, shade@kemsu.ru.

Sergey Yu. Zavozkin – Candidate of Technical Sciences, Assistant Professor at the UNESCO Department for New Information Technologies, Kemerovo State University.

Шевнин Василий Алексеевич – аспирант кафедры ЮНЕСКО по новым информационным технологиям КемГУ, vaaaasyaaaa@gmail.com.

Vasily A. Shevnin – post-graduate student at the UNESCO Department for New Information Technologies, Kemerovo State University.

(**Научный руководитель – *С. Ю. Завозкин***).

Статья поступила в редколлегию 17.10.2014 г.