УДК 004.42:004.891

ВЕРИФИКАЦИЯ МЕТОДИКИ КЛАССИФИКАЦИИ БИБЛИОГРАФИЧЕСКИХ ДАННЫХ НА ОСНОВЕ УСЛОВНО-СЛУЧАЙНЫХ ПОЛЕЙ

[1,3]Черников Б.В., [2]Петров Е.Н., [3]Борисова Е.А.

[1]ООО «Газпром ВНИИГАЗ», Москва, e-mail: bor-cher@yandex.ru;
[2]Национальный исследовательский университет «МИЭТ», Зеленоград;
[3]Российский экономический университет им. Г.В. Плеханова, Москва

По оценкам специалистов, от 70 до 90% всей потенциально полезной деловой информации представлено в неструктурированной форме, что, учитывая взрывной рост объемов информации за последние годы, связанный с развитием интернета, снижает возможности ее эффективной обработки и анализа. В частности, существуют проблемы обработки библиографических данных, используемые в библиотеках, вузах, в научно-исследовательских институтах и в других организациях, связанных с издательской или научной деятельностью. При ручной обработке данных результат работы подвержен влиянию человеческого фактора: неточности при перепечатке, пропуск опциональных и обязательных полей библиографического описания, нарушение порядка следования, недостаток или избыток пунктуационных символов. Указанные отступления от стандартов усложняют поиск и определение использованных источников, затрудняя научную и отчетную работу. С другой стороны, выполнение данной работы средствами вычислительной техники гарантирует отсутствие ошибок по невнимательности, универсальность обработки различных форматов данных, структуризацию и быстроту обработки и, как следствие, своевременную отчетность. В статье представлена характеристика программной реализации и верификации методики классификации библиографических данных на основе условно-случайных полей, которая была разработана для решения этой задачи. В результате вычислительного эксперимента, проведенного с помощью программного комплекса классификации библиографических данных, получено увеличение качества классификации модели, обученной с помощью разработанной методики, на 9–10%.

Ключевые слова: библиографические данные, программное обеспечение, классификация, точность, полнота

VERIFICATION OF CRF-BASED BIBLIOGRAPHIC DATA CLASSIFICATION METHOD

[1,3]Chernikov B.V., [2]Petrov E.N., [3]Borisova E.A.

[1]LLC «Gazprom VNIIGAZ», Moscow, e-mail: bor-cher@yandex.ru;
[2]Moscow Institute of Electronic Technology, Zelenograd;
[3]Plekhanov Russian University of Economics, Moscow

According to recent surveys, 70 to 90% of business data is stored unstructured, and while the volume of such data has increased, the processing ability has not. This fact affects the efficiency of the data processing and analysis. In particular, bibliographic data that is used in libraries, universities and other organizations is still manually processed. Thus human factor has a dramatic impact on processing results. Typos, missing bibliographical elements, wrong sequence order and punctuation errors complicate proper citation search and obstruct scientific work and report composition. Meanwhile computer processing guarantees faster and error free data processing. The article presents software implementation and verification of CRF-based bibliographic data classification method that was developed to solve the described problem in earlier studies of the authors. The experiment carried out with the bibliographic data classification software showed classification quality improvement by 9-10% due to the applied method. In the current state of development, operator involvement is still required. This proves the need for further research to improve the methodology.

Keywords: bibliographic data, software, classification, precision, recall

В настоящее время обработка библиографических данных с целью учета научных трудов в высших учебных заведениях и других организациях производится вручную, что существенно сказывается на временных и трудовых затратах сотрудников. Как и во всякой рутинной работе, выполняемой вручную, качество выполняемой работы страдает из-за человеческого фактора – невнимательности, ошибок и опечаток. В рамках предыдущих исследований для решения исследуемой задачи была разработана методика классификации библиографических данных с помощью условно-случайных полей [1]. В настоящей статье рассматривается программная реализация методики в виде программного комплекса и верификация качества классификации с помощью метрик точности, полноты и F-меры [2].

Компьютерная обработка библиографических данных с минимальным участием человека лишена описанных недостатков и может выполняться в кратчайшие сроки. Выполнение данной работы компьютером гарантирует отсутствие ошибок, вызванных

невнимательностью, обеспечивает структуризацию и высокую скорость обработки и, как следствие, своевременное формирование отчетности.

Таким образом, задача разработки программного комплекса классификации библиографических данных, способного выполнять такую обработку, является актуальной.

Цель исследования заключается в разработке программного комплекса для верификации того, что разработанная методика [1] позволяет выполнять компьютерную обработку библиографических данных с минимальным участием оператора, обеспечивая прирост качества классификации по сравнению с моделью, обученной без использования этой методики.

Материалы и методы исследования

Для программной реализации алгоритма и методики классификации библиографических данных с помощью условно-случайных полей разработан программный комплекс, архитектура которого представлена на рис. 1.

Разработанный программный комплекс включает в себя следующие модули:

– модуль предобработки и вычисления функций признаков;

– модуль обучения, оценки и визуализации;

– модуль классификации и преобразования.

Модуль предобработки данных и вычисления функций признаков выполняет разбиение блока библиографических данных на отдельные фрагменты библиографической записи и их дальнейшую токенизацию (разбиение на слова). Каждое слово в библиографической записи сохраняется в двух представлениях: со знаками препинания и без них. Такая обработка необходима для корректной работы модуля вычисления функций признаков, так как среди них присутствуют признаки, основанные на наличии в слове определенных символов. Кроме того, этот модуль выполняет преобразование библиографической записи в набор векторов, являющихся результатами вычисления функций признаков для каждого слова обрабатываемой записи.

Модуль обучения, оценки и визуализации используется только в случае, если классификационная модель, представляющая собой бинарный файл с результатами работы модуля, еще не создана или нуждается в переобучении. В противном случае работа программного комплекса может производиться на основе модели, полученной при предыдущих запусках. Данный модуль получает на вход набор размеченных библиографических записей и их векторизованное представление, полученное с помощью модуля предобработки данных и вычисления функций признаков. Полученные библиографические записи случайным образом перемешиваются и разбиваются на записи для обучения и тестирования в соотношении 80 к 20 [3]. После этого в соответствии с методикой [1] производится обучение классификационной модели, которое заключается в вычислении оптимальных весов функций признаков и вероятностей переходов между отдельными классами данных, выбранных для обучения. Вычисления проводятся на наборе обучающих данных с помощью алгоритма оптимизации L-BFGS (модификация алгоритма Бройдена – Флетчера – Гольдфарба-Шанно с ограниченным использованием памяти, представленная в статье [4]), выбор которого обусловлен результатами исследований последних лет [5].

Для проверки адекватности обучения модуль выполняет вычисление метрик классификатора на наборе тестовых данных и представляет результаты в графическом виде. Классификатор оценивается по следующим параметрам: точность, полнота и F-мера [2] для каждого класса и в среднем. В графическом виде выводится матрица ошибок каждого класса, дающая наглядное представление о том, в каких классах ошибается обученная модель и с какими классами чаще всего происходит путаница.

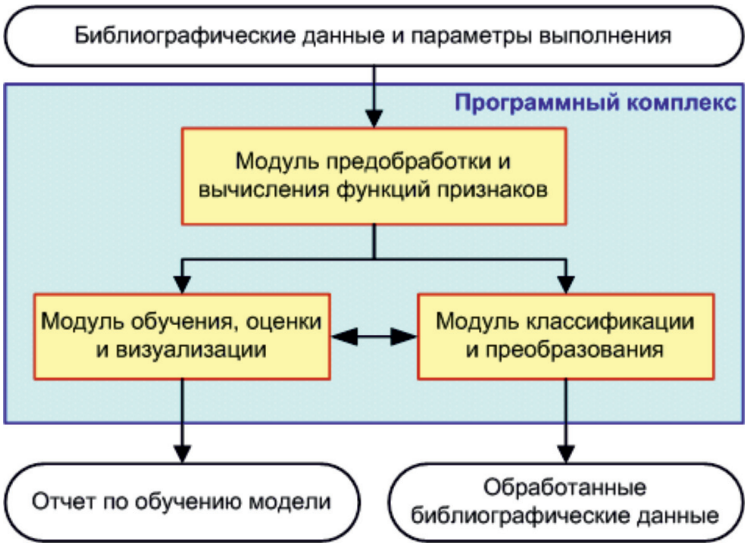

Рис. 1. Структура программного комплекса

Информация, которая может быть получена с помощью этого модуля, позволяет оценить эффективность работы классификатора и внести необходимые правки в набор функций признаков и правила предобработки. Полученная модель вместе с данными для тестирования передается в модуль классификации для подтверждения адекватности обучения модели и вывода отчета по обучению.

Модуль классификации и преобразования используется в случае, когда классификационная модель уже обучена. Данный модуль выполняет извлечение весов функций признаков и вероятностей межклассовых переходов из обученной модели, после чего вычисляет взвешенные значения функций признаков, полученных из векторизованного представления. Вычисленные значения признаков используются для определения последовательности классов с помощью алгоритма Витерби, впервые сформулированного в статье [6]. Кроме того, модуль выполняет постобработку классифицированных библиографических записей, зависящую от параметров программного комплекса, указанных пользователем, и рассматриваемых ниже.

Разработанный программный комплекс классификации библиографических данных на основе условно-случайных полей осуществляет работу в трех основных режимах:

– вывод справочной информации;
– обучение модели;
– собственно классификация.

Вывод справочной информации выполняется в случае вызова программы с некорректными параметрами (например, не указан формат записи в выходной файл). Выбор режима работы и входные параметры передаются аргументами командной строки и включают в себя следующие флаги и настройки:

– вывод справочной информации;
– тип входных данных;
– тип выходных данных;
– логин и пароль от базы данных (если БД – выходной источник);
– путь к файлу с входными данными;
– путь к файлу с выходными данными;
– режим работы (обучение или классификация).

Режим вывода справочной информации игнорирует остальные параметры и выводит на экран строку, представленную ниже и сообщающую о правилах вызова программы и возможных наборах параметров:

usage: python3 pybico_crf.py [−h] [−l <input_mode> −i<input_file>][−s <output_mode>[−o <output_file>] | [−u <user_name> −p <user_password>]].

Под типом входных данных понимается тип файла, содержащего библиографические данные для обучения или классификации. В качестве типа входного источника могут выступать текстовый файл (*.txt) или текстовый документ (*.doc, *.docx), при этом в текстовом документе информация может быть представлена в виде таблиц.

Под типом выходных данных понимается тип файла, содержащего классифицированные библиографические данные, или база данных, в которую эти данные следует сохранить. Выходными могут быть текстовый файл (*.txt), табличный документ (*.xls, *xlsx), текстовый документ (*.doc, *.docx) или база данных. Параметр режима работы определяет, какая задача решается программным комплексом: обучение или классификация. Формат входных данных зависит от выбранного режима работы. В случае если осуществляется обучение модели, входные данные представляют собой файл с размеченными библиографическими данными, где после каждого блока однотипных элементов указывается метка класса, к которому этот блок принадлежит, а каждая библиографическая запись начинается с новой строки. Пример размеченной библиографической записи выглядит следующим образом:

Петров Е.Н., Чумаченко П.Ю. T_AUTHOR Программное средство конфигурационных файлов сетевого оборудования. T_TITLE Молодой учёный. T_JOURNAL – T_DELIMETER 2016. T_YEAR –T_DELIMETER №10.T_VOLUME – T_DELIMETERC.110–115.T_PAGES

В том случае, если осуществляется непосредственно классификация, входные данные представляют собой файл с библиографическими записями. Каждая запись в таком файле начинается с новой строки. В зависимости от типа входного источника файл может являться текстовым файлом или текстовым документом, что облегчает работу пользователя с программным комплексом. Библиографические записи могут не иметь четкого формата и содержать ошибки оформления, связанные с перестановками блоков однотипных элементов и пунктуацией. Пример такой неразмеченной библиографической записи неопределенного формата может быть представлен таким образом:

Программное средство верификации конфигурационных файлов сетевого оборудования.// Молодой учёный. – С. 110–115. 2016. – Петров Е. Н., Чумаченко П. Ю. № 10.

Несмотря на несоответствие приведенного примера какому-либо стандарту составления библиографических записей, разработанный программный комплекс в состоянии провести автоматическую классификацию с помощью разработанной методики на основе условно-случайных полей [1]. Пример работы программного комплекса в режиме классификации представлен на рис. 2.

В зависимости от режима работы и выходного источника результат работы программного комплекса может преобразовываться в различные форматы выходных данных.

В случае если выбранным режимом работы является обучение, выходные данные представляют собой обученную модель и результаты оценки этой модели в соответствии с выбранными метриками. Если выбранным режимом работы является непосредственно классификация, формат выходных данных зависит от параметров программы. Когда в качестве выходного файла указан текстовый документ или текстовый файл, выходные данные представляют собой набор классифицированных и преобразованных к стандарту ГОСТ 7.0.100-2018 [7] библиографических записей.

Результат классификации может быть сохранен в базе данных, тогда выходные данные представляют собой набор классифицированных и сгруппированных элементов для каждой библиографической записи.

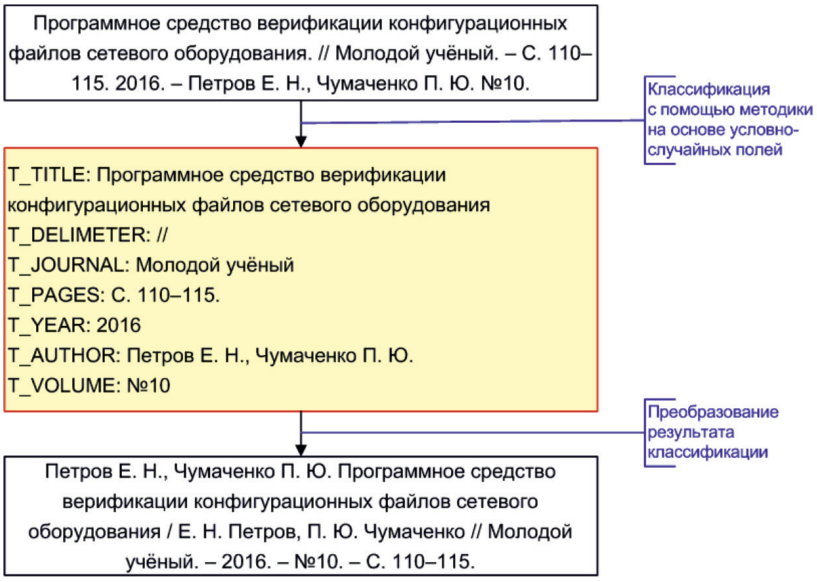

Рис. 2. Результат работы программного комплекса в режиме классификации

После этого группы заносятся в базу данных с сохранением связей и принадлежности к исходной библиографической записи. В том случае если в качестве выходного объекта выбран табличный документ, выходные данные представляют собой набор классифицированных и сгруппированных элементов, сохраненных в виде таблицы, где каждая строка соответствует отдельной библиографической записи, а столбцы содержат сгруппированные элементы.

С помощью режима обучения программного комплекса была проведена оценка эффективности разработанной методики на основе сравнения качества классификации, определяемого метриками точности, полноты и F-меры моделей, обученных с учетом разработанной методики и на стандартном наборе признаков для работы с текстом на естественном языке. Для верификации полученных результатов проведен вычислительный эксперимент.

В ходе вычислительного эксперимента в качестве входных данных использовалась тестовая выборка размеченных библиографических записей, составленных на основе научных трудов сотрудников института СПИНТех за последние пять лет. Входные данные делились на две части в соотношении 80 к 20, для обучения и проверки соответственно. Верификация исследований производилась на основе сравнения метрик точности, полноты и F-меры модели с использованием разработанного признакового представления и стандартной модели для обработки текста.

Результаты исследования и их обсуждение

В соответствии с разработанной методикой была проведена поклассовая оценка качества классификации с помощью вычисления метрик точности, полноты и F-меры:

$$\text{Точность} = \frac{TP}{TP + TN};$$

$$\text{Полнота} = \frac{TP}{TP + FN};$$

$$F = 2 * \frac{\text{Точность} \times \text{Полнота}}{\text{Точноть} + \text{Полнота}},$$

где TP – число истинно положительных результатов классификации (таких результатов, в которых объект рассматриваемого класса соотнесен с этим классом), TN – истинно отрицательных (таких результатов, в которых объект другого класса не соотнесен с рассматриваемым классом), FN – ложно отрицательных (таких результатов, в которых объект рассматриваемого класса не соотнесен с этим классом).

В ходе вычислительного эксперимента, включающего в себя оценку результатов классификации по формальным критериям, отобранным в рамках разработанной методики, получены значения точности, полноты и F-меры для каждого из рассматриваемых классов, а также проведено их сравнение при обучении с использованием разработанной методики и без нее (таблица). Визуализация сравнения F-мер представлена на рис. 3.

Исходя из данных таблицы и представленной диаграммы, можно сделать вывод об уменьшении качества классификации года, города и издательства. Понижение качества классификации года издания свидетельствует о необходимости добавления характерных функций признаков для этого класса в разработанную методику. Уменьшение качества классификации города издания может быть компенсировано таблицей рас-

пространенных значений, определение которых является темой для отдельного исследования. Судя по значениям и соотношению точности и полноты классификации названия издательства, снижение в этом классе связано с ошибочным включением города издательства в название. Таким образом, улучшение качества классификации города положительно скажется на полноте классификации названия издания.

Несмотря на снижения качества классификации в перечисленных случаях, улучшения в классификации авторов, названий статей и названий журналов, а также средний прирост качества классификации в 9–10%, позволяют сделать вывод о положительном эффекте от применения разработанной методики.

Заключение

1. На основе методики классификации библиографических данных с помощью условно-случайных полей, полученной в рамках предыдущих исследований, проведена разработка и программная реализация программного комплекса классификации библиографических данных и проведена оценка качества классификации разработанной методики с помощью метрик точности, полноты и F-меры. Средний положительный прирост при использовании методики составил 9–10% по сравнению с моделью, обученной на той же тестовой выборке без использования этой методики. Причины снижения качества классификации по отдельным классам обозначены и будут учтены в дальнейших исследованиях.

Сравнение моделей по метрикам качества классификации

Класс	Модель, обученная без использования разработанной методики			Модель, обученная с помощью разработанной методики		
	Точность	Полнота	F-мера	Точность	Полнота	F-мера
T_AUTHOR (автор)	1,000	0,750	0,857	1,000	1,000	1,000
T_DELIMITER (разделитель)	1,000	1,000	1,000	0,938	1,000	0,968
T_JOURNAL (журнал)	0,794	0,862	0,827	0,940	0,959	0,949
T_LOCATION (город издания)	1,000	0,500	0,667	0,800	0,800	0,800
T_PAGES (страницы)	0,935	1,000	0,967	1,000	1,000	1,000
T_PUBLISHER (издательство)	1,000	1,000	1,000	1,000	0,500	0,667
T_TITLE (название)	0,827	0,893	0,859	1,000	1,000	1,000
T_TJSEP (разделитель названия статьи и названия журнала)	1,000	1,000	1,000	1,000	1,000	1,000
T_VOLUME (том)	0,938	0,833	0,882	1,000	0,895	0,944
T_YEAR (год)	1,000	0,870	0,930	0,800	0,842	0,821
Среднее	0,881	0,872	0,871	0,969	0,968	0,967

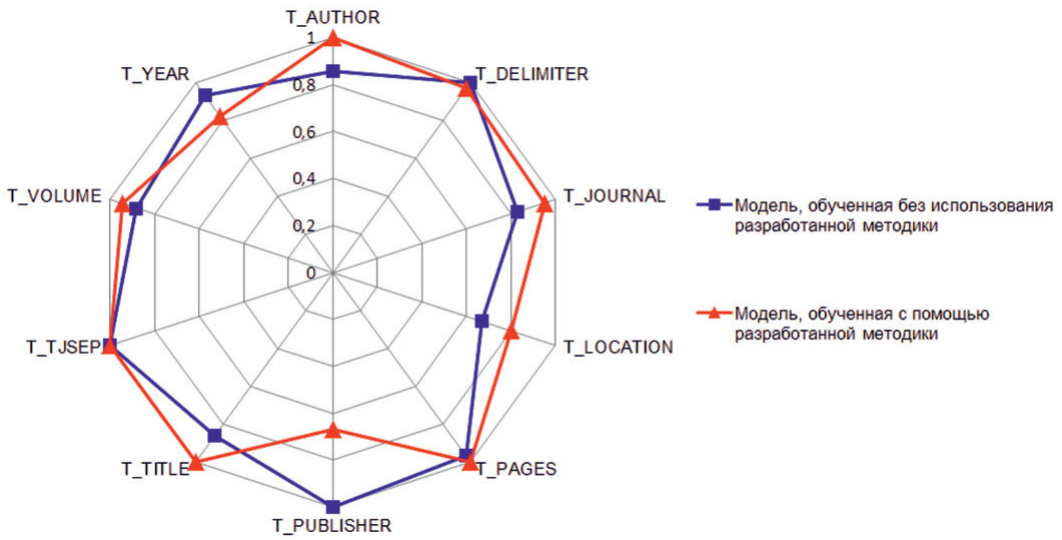

Рис. 3. Визуализация F-меры обученных моделей

2. Средний результат качества классификации, составивший 96 %, позволяет сделать вывод о практической применимости разработанного программного комплекса при обработке библиографических данных в автоматизированном, но не автоматическом режиме. Для устранения возможных ошибок обработки все еще требуется участие оператора, что оставляет пространство для улучшений системы классификации библиографических данных.

Список литературы

1. Петров Е.Н. Исследование и разработка методики и алгоритма классификации библиографических данных с помощью условно-случайных полей // Итоги диссертационных исследований: материалы X Всероссийского конкурса молодых ученых. Т. 2. М.: РАН, 2018. С. 91–98.

2. Powers D.M. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies. 2011. Vol. 2. Issue 1. P. 37–63.

3. Sutton C., McCallum A. An Introduction to Conditional Random Fields. Foundations and Trends in Machine Learning. 2012. Vol. 4. no. 4. P. 267–373.

4. Liu D.C., Nocedal J. On the limited memory BFGS method for large scale optimization. Mathematical programming. 1989. Vol. 45. no. 1–3. P. 503–528.

5. Quoc V. Le, Jiquan Ngiam, Adam Coates, Abhik Lahiri, Bobby Prochnow, Andrew Y.Ng. On optimization methods for deep learning. In Proceedings of the 28th International Conference on International Conference on Machine Learning (ICML'11). Omnipress, USA, 2011. P. 265–272.

6. Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE transactions on Information Theory. 1967. Vol. 13. no. 2. P. 260–269.

7. ГОСТ 7.0.100-2018. Библиографическая запись. Библиографическое описание. Общие требования и правила составления. М.: Изд-во стандартов, 2018. 128 с.