nomial algorithms for scheduling in GRID-systems], *Informatsionnye tekhnologii* (Information Technologies], 2012, no. 9, pp. 28—32.

21. **Saak A. Eh.** Urovnevye algoritmy dispetcherizatsii massivamy zayavok krygovogo tipa v Grid-sistemakh [Level algorithms of scheduling by circle type task sets in grid systems], *Izvestiya YuFU. Tekhnichcskie nauki* [Izvestiya SFedU. Engineering Sciences], 2015, no. 6 (167), pp. 223—231.

22. **Saak A. Eh.** Kolcivie algoritmy dispetcherizatsii massivamy zayavok v Grid-sistemakh [Ring algorithms for scheduling in grid systems by sets of tasks], *Informatsionnye tekhnologii* [Information Technologies], 2016, no. 3, pp. 163—169.

23. **Saak A. Eh.** Dispetcherizatsiya zayavok krygovogo tipa v Grid-sistemakh [Circular-typed multiprocessor tasks scheduling in grid systems], *Informatsionnye tekhnologii* (Information Technologies], 2016, no. 1, pp. 37—41.

24. **Jepsen C.** Dissections into 1:2 rectangles, *Discrete Mathematics*. 1996, vol. 148, pp. 107—117.

25. **Friedman E.** *Dominoes in squares*. 2014. URL: http://www2.stetson.edu/~efriedma/domino/

26. **Korf R., Moffitt M., Pollack M.** Optimal rectangle packing, *Annals of Operations Research*, 2010, vol. 179, Iss. 1, pp. 261—295.

27. **Huang E., Korf R.** Optimal rectangle packing: an absolute placement approach, *Journal of Artificial Intelligence Research*, 2013, vol. 46, pp. 47—87.

28. **Friedman E.** *Tiling Almost Squares*. 2013. URL: http://www2.stetson.edu/~efriedma/almost/

УДК 004.912; 519.254

А. С. Мохов[1], ассистент, e-mail: asmokhov@mail.ru,
В. О. Толчеев[2], д-р техн. наук, проф., e-mail: tolcheevvo@mail.ru
НИУ "МЭИ"

Способы учета структуры научных документов в задачах обработки и анализа текстовой информации

Рассмотрены процедуры обработки и анализа текстовой информации на основе учета структуры документа. Приведены основные модели представления текстов в задачах машинного обучения. Показана эффективность использования частично структурированных моделей для информационного поиска, автоматического аннотирования, выявления нечетких дубликатов и классификации. Наряду с известными подходами в работе излагаются предложенные авторами процедуры, учитывающие особенности двуязычных библиографических документов и позволяющие проводить высокоточную классификацию.

Ключевые слова: *интеллектуальный анализ текстовых данных, модель текстового документа, информационный поиск, автоматическое аннотирование, выявление нечетких дубликатов, классификация двуязычных библиографических документов*

Введение

В рамках работ по интеллектуальному анализу текстовых данных (Text Mining) решаются задачи поиска информации, кластеризации, фильтрации и классификации документов, автоматического аннотирования и извлечения ключевых понятий из текстов, выявления плагиата и нечетких дубликатов в документальных массивах, создания онтологий и построения моделей предметных областей. Общей особенностью вышеуказанных задач является их слабая формализованность, сложность математического описания текстов.

Под структурой документа далее понимается его логическое построение в виде взаимосвязанных фрагментов. Структура текстов во многом обусловливается функциональным стилем, которому соответствует публикация. Различают публицистический, официально-деловой, научно-технический, разговорный, художественный стили. В данной работе рассматривается научно-технический стиль, который используется в монографиях, статьях, учебниках, диссертациях, отчетах по НИР и т. д. Для научных статей в качестве фрагментов целесообразно рассматривать название, аннотацию, ключевые слова, введение, основную часть, заключение. Для учебного пособия или монографии это могут быть главы, подглавы, параграфы и т. п.

Характерной особенностью научно-технических документов является наличие наряду с полным текстом короткого библиографического описания (БО), включающего имена и фамилии авторов, название, аннотацию, ключевые слова, место издания и другую вспомогательную информацию. Чаще всего БО представлены на двух языках: русском и английском. В библиографических описаниях, как и в других коротких текстах (сообщения электронной почты, новостная информация, описания товаров, инструкции по применению, анкеты, справочники), наиболее четко прослеживается общая структура. В простейшем случае библиографический научный документ может быть описан кортежем $<T, A, K>$, в который включены название (title) — T, аннотация (abstract) — A, ключевые слова (key

words) — *K*. Именно научные публикации, заданные своими двуязычными (русско-английскими) БО, рассматриваются в данной статье, и применительно к ним показана целесообразность использования заранее известной структуры для более эффективного решения прикладных задач Text Mining.

Основные модели представления текстовых документов

Результаты поиска, обработки и анализа текстовой информации существенным образом зависят от выбранного способа представления и математического описания документа. Чаще всего используется широко известная в литературе модель "мешок слов" ("bag of words"), которая представляет публикацию в виде набора не связанных между собой слов [1]:

$$X_j = [x_j^{(1)}, ..., x_j^{(i)}, ..., x_j^{(M)}]^{\mathrm{T}}, \qquad (1)$$

где $x_j^{(i)}$ — вес термина i в документе j ($j = 1, ..., N$; N — число документов в выборке; $i = 1, ..., M$; M — число информативных терминов после удаления служебных слов). Значение $x_j^{(i)}$ будет существенно зависеть от способа взвешивания терминов (в задачах Text Mining обычно используются *tf*-, *tf-idf*-, *tfc*-, *ltc*-взвешивания) [2, 3]. К информативным словам относятся:

- общенаучные (высокочастотные) термины, единые для различных предметных областей;
- междисциплинарные и узкоспециализированные понятия (среднечастотные слова), характеризующие тематику статьи;
- редкие (низкочастотные) термины, соответствующие конкретному научному направлению или отражающие авторский стиль.

Выборка текстовых документов может быть представлена в виде матрицы "документ—термин":

$$X = \begin{bmatrix} x_1^{(1)} & \cdots & x_1^{(i)} & \cdots & x_1^{(M)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_j^{(1)} & \cdots & x_j^{(i)} & \cdots & x_j^{(M)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_N^{(1)} & \cdots & x_N^{(i)} & \cdots & x_N^{(M)} \end{bmatrix}. \qquad (2)$$

Более полное описание документов благодаря использованию всей доступной информации достигается в *частично структурированных* и *полностью структурированных* моделях. В частично структурированных моделях при определении весов терминов учитываются дополнительные характеристики: порядок следования слов (символов), их расположение в документе (заголовок, аннотация, ключевые слова, первый абзац и т. д.), наличие специального оформления термина стилистически или с помощью шрифта. Кроме того, в ряде случаев проводится выделение гиперонимов — слов с более широким значением (нож — (это) оружие, бульдог — собака — животное) и словосочетаний — устойчивых групп терминов, которые образуют общие понятия для данной предметной области [4—6]. Частично структурированные модели весьма эффективны при обработке и анализе библиографических документов, новостной и справочной информации, анкет, т. е. коротких текстов, которые обладают заранее известной структурой.

В полностью структурированных моделях применяются заранее сформированные базы знаний, содержащие ключевые слова, их сочетания, а также иерархические связи, свойственные предметной области. Традиционно для этого разрабатываются онтологии и тезаурусы, на верхних уровнях которых находятся общие термины, уточняемые на более низких уровнях [7].

В специализированной литературе активно обсуждается вопрос, насколько высокая сложность и существенные трудозатраты при разработке частично структурированных и полностью структурированных моделей гарантируют значимое улучшение целевых показателей (например, точности-быстродействия классификации документов или полноты-точности поиска информации). Необходимо признать, что в профессиональном сообществе пока не выработано единой позиции по этой проблеме [4, 7, 8]. Справедливо отмечается, что выбор модели, прежде всего, зависит от области применения (сложности решаемой задачи Text Mining), требований целевого критерия, имеющейся априорной информации [2].

Решение проблемы обработки и анализа больших массивов информации (так называемой проблемы больших данных — Big Data), требующее распределения колоссального числа документов по многочисленным классам, целесообразно организовывать на базе полностью структурированных моделей — онтологий. В настоящее время для описания в онтологиях семантики лексических единиц (слов, словосочетаний, предложений и дискурсов) разработаны эффективные подходы, в частности, теория K-представлений (концептуальных представлений) [18—20].

Для широкого круга задач классификации, предусматривающих разнесение небольших выборок по сравнительно малому числу классов, достаточно использовать модель "мешок слов". При этом более высокая точность может быть достигнута путем проведения комплексного обучения решающего правила на большом числе выборок, тщательного отбора информативных терминов с помощью специализированных процедур, формирования коллективов решающих правил, использования дополнительной априорной информации (например, структуры БО) [2, 9].

Частично стуктурированные модели в задачах поиска, обработки и анализа информации

В центре внимания данной статьи находятся такие задачи Text Mining, для решения которых, на наш взгляд, целесообразно применять частично структурированные модели. Рассмотрим эти задачи более подробно.

Информационный поиск. На раннем этапе разработка частично структурированных моделей проводилась, прежде всего, в рамках работ по теории информационного поиска в целях повышения релевантности выдаваемых документов. В этих исследованиях проверялась эффективность применения цитирования (перекрестных ссылок), учета местоположения и частоты появления терминов, анализа стиля и шрифта написания слов для улучшения показателя полнота-точность. В результате были разработаны достаточно универсальные подходы, в частности PageRank, широко известная процедура Google, предназначенная для оценки качества Web-страниц на основании информации о количестве имеющихся на нее ссылок [5]. Более комплексный подход, применимый как к Web-документам, так и журнальным публикациям, предложили В. Фресно и А. Рибейро (V. Fresno, A. Ribero) [10].

В данной модели рассчитываются следующие характеристики текстов:

- частота i-го термина в документе: $f_f(i) = n_f(i)/N_{tot}$;
- частота i-го термина в заголовках документа: $f_t(i) = n_t(i)/N_{tit}$;
- частота i-го термина среди выделенных слов (жирное начертание, курсив): $f_e(i) = n_e(i)/N_{emph}$;
- функция положения i-го термина в документе:

$$f_p(i) = \frac{\frac{3}{4}n_{1,4}(i) + \frac{1}{4}n_{2,3}(i)}{\sum_{l=1}^{k}\left(\frac{3}{4}n_{1,4}(l) + \frac{1}{4}n_{2,3}(l)\right)}.$$

Дадим расшифровку обозначений: $n_{1,4}(i)$ — число вхождений слова в первую и четвертую четверти документа; $n_{2,3}(i)$ — число вхождений слова во вторую и третью четверти документа; k — число различных терминов в соответствующих частях документа; $n_f(i)$, $n_t(i)$, $n_e(i)$ — число раз, которые i-й термин встречался в документе, заголовке, среди выделенных слов; $N_{tot}(i)$, $N_{tit}(i)$, $N_{emph}(i)$ — число терминов в документе, заголовке, общее число выделенных слов.

Итоговый вес вычисляется как линейная комбинация вышеуказанных показателей:

$$r_i = C_f f_f(i) + C_t f_t(i) + C_e f_e(i) + C_p f_p(i). \qquad (3)$$

Здесь C_f, C_t, C_e, C_p — настраиваемые коэффициенты, определяющие степень влияния различных характеристик термина на значение результирующего веса.

Эмпирически авторы получили следующие значения коэффициентов:

$$r_i = 0{,}3f_f(i) + 0{,}15f_t(i) + 0{,}25f_e(i) + 0{,}3f_p(i). \qquad (4)$$

В модели В. Фресно и А. Рибейро настройка параметров проводится на основе экспериментальных исследований. Поэтому в новых условиях (другая предметная область, измененный размер и структура Web-страниц) придется заново настраивать весовые коэффициенты и искать их наилучшее сочетание. Возможно, этим объясняется весьма ограниченное практическое применение формулы (4) в задачах Text Mining.

Автоматическое аннотирование. Автоматическое аннотирование (реферирование) заключается в выявлении наиболее информативных фрагментов полнотекстового документа и их объединении в короткую аннотацию-реферат (отметим, что такие рефераты во многом являются аналогом библиографического описания) [5]. Кроме анализа частоты появления термина, его стиля и шрифта, контекста (окружение термина, очередность следования слов вокруг него), при автоматическом аннотировании анализируется местоположение термина или фрагмента в документе. Это позволяет корректно оценить их значимость и полезность для включения в короткое описание. В литературе предложены различные способы определения ценности фрагмента, например можно использовать формулу [5]

$$Weight(r) = Location(r) + KeyPhrase(r) + StatTerm(r) + AddTerm(r), \qquad (5)$$

где r — фрагмент исходного документа ($r = 1, ..., R$, R — общее число фрагментов в тексте); *Location* — весовой коэффициент, значение которого зависит от места появления фрагмента (начало, середина или конец параграфа, введение, заключение); *KeyPhrase* — весовой коэффициент, увеличивающий значимость фрагмента, который входит в ключевые конструкции-клише (например, для научных публикаций: статья посвящена...., рассматривается актуальная задача...., результаты эксперимента... и т. п.); *StatTerm* — весовой коэффициент, учитывающий частоту встречаемости фрагмента в тексте; *AddTerm* — весовой коэффициент, повышающий вес фрагмента в случае наличия в нем важных терминов (определенных пользователем или экспертом).

После расчета весов отбираются фрагменты, обладающие наибольшими значениями *Weight*(r). Слабое место данного подхода заключается в неформализованности расчета весовых коэффициентов, которые в ряде случаев могут назначаться самими разработчиками исходя лишь из собственного опыта и предпочтений.

Выявление нечетких дубликатов. К *дубликатам (неуникальным публикациям)* принято относить документы с идентичным содержанием. *Нечеткими (неполными) дубликатами* или *почти дубликатами*

считаются документы, в содержательную часть которых внесены незначительные изменения. Проблема выявления нечетких дубликатов возникает при поиске информации (устранение одинаковых документов, выдаваемых информационно-поисковой системой), формировании цифровых библиотек и банков данных (исключение повторяющихся сведений, полученных из разнородных источников), проверке на оригинальность представленных для издания научных статей и докладов (обнаружение работ, которые были уже ранее опубликованы авторами в других журналах) [11, 12].

В литературе по Text Mining предложены и исследованы различные способы выявления нечетких дубликатов (специализированные расстояния, меры близости и коэффициенты ассоциативности, метод шинглов и его модификации, процедуры на основе расчета весов терминов и т. п.) [11]. В известных подходах практически не используется априорная информация о строении документа. Вместе с тем идентификация дубликатов часто проводится именно среди коротких документов, хорошо поддающихся структурированию (библиографические описания научных статей, новостные сообщения, анкеты и т. п.).

Рассматриваемый в данной работе метод — *Обобщенный Коэффициент Ассоциативности* (*ОКА*) — предназначен для обнаружения нечетких дубликатов по библиографическим описаниям научных статей. Его специфика заключается в выявлении не только совпадающих терминов двух сравниваемых текстов, но и в учете того, в каком месте БО (название или аннотация) это совпадение произошло. Расчет ОКА основан на вычислении различных коэффициентов ассоциативности, определяемых отдельно для названий и аннотаций [13].

Введем следующие обозначения, используемые при анализе таблиц сопряженности размера 2 × 2: $A_{\text{назван}}$ и $A_{\text{аннот}}$ — число совпавших терминов соответственно в названиях и аннотациях двух документов X_j и X_l; $B_{\text{назван}}$ и $B_{\text{аннот}}$ — число терминов (соответственно из названий и аннотаций), имеющихся в X_j и отсутствующих в X_l; $C_{\text{назван}}$ и $C_{\text{аннот}}$ — число терминов (соответственно из названий и аннотаций), имеющихся в X_l и отсутствующих в X_j.

ОКА определяется по формуле [13] :

$$OKA = \frac{1}{2}(K_1 + K_2) =$$

$$= \frac{1}{2}\left\{\frac{A_{\text{назван}}}{\max(A_{\text{назван}}, B_{\text{назван}}, C_{\text{назван}})} + \right.$$

$$\left. + \min\left(\frac{A_{\text{аннот}}}{A_{\text{аннот}} + B_{\text{аннот}}}; \frac{A_{\text{аннот}}}{A_{\text{аннот}} + C_{\text{аннот}}}\right)\right\}. \qquad (6)$$

Здесь K_1 — коэффициент, который вычисляется только по терминам, встречающимся в названиях статей ($K_1 \in [0; 1]$). Такой расчет K_1 позволяет более качественно выявлять статьи с терминологически близкими названиями и публикации с "непохожими" заголовками; K_2 — коэффициент, который рассчитывается только по терминам из аннотаций ($K_2 \in [0; 1]$), равен наименьшему из выражений, стоящих в скобках. Это позволяет учесть наличие в двух аннотациях несовпадающих терминов. Уменьшение значений K_1 и (или) K_2 снижает ОКА, возрастание значений K_1 и (или) K_2 увеличивает ОКА. Усреднение суммы коэффициентов K_1 и K_2 приводит значения ОКА в диапазон [0; 1] (равенство ОКА единице означает обнаружение полного дубликата).

Результаты исследования ОКА (расчет показателя полнота-точность на тестовых массивах) и сопоставление ОКА с известными процедурами (метод шинглов, коэффициент Жаккара, расстояния Джаро и Джаро-Винклера) приведены в работе [13]. Сравнительный анализ показал, что учет структуры документа в ОКА позволяет увеличить точность распознавания нечетких дубликатов без уменьшения показателя полноты.

Классификация библиографических текстовых документов. Несмотря на широкое использование модели "мешок слов" в задачах классификации, усиливается интерес к применению частично структурированных моделей для снижения ошибки группировки документов [2, 6, 14]. В данной работе рассматриваются подходы, которые позволяют повысить точность за счет анализа информации о структуре текстового документа, в частности, сведений о местоположении терминов. Так, в работе [9] предлагается следующая формула линейного взвешивания:

$$x_j^{(1)} = \alpha t_j^{(i)} + \beta a_j^{(i)} + \gamma k_j^{(i)}, \qquad (7)$$

где $x_j^{(i)}$ — результирующий вес термина i в документе j; $t_j^{(i)}$ — вес термина i в названии документа j; $a_j^{(i)}$ — вес термина i в аннотации документа j; $k_j^{(i)}$ — вес термина i в ключевых словах документа j; α, β, γ — настраиваемые весовые коэффициенты.

Как и в других рассмотренных выше подходах, наиболее неформализованной проблемой при использовании формулы (7) является расчет весовых коэффициентов. В работе [9] для определения величин α, β, γ были использованы две стратегии, основанные на экспериментальных исследованиях. Согласно первой стратегии коэффициенты рассчитывались как отношение ошибок, получаемых при классификации текстовых документов при раздельном использовании терминов из названий, аннотаций и ключевых слов (для классификации использовались метод k-ближайших соседей и косинусоидальная мера близости). При использовании второй стратегии коэффициенты определялись как отношение числа терминов, встречавшихся в аннотациях, названиях и ключевых словах

обучающей выборки. Настройка коэффициентов α, β, γ по второй стратегии позволила увеличить точность классификации на 5 % [9].

В отличие от задач информационного поиска и автоматического аннотирования, для которых определение местоположения термина является стандартным приемом, при классификации документов этот подход не стал общепринятым, возможно, из-за большей трудоемкости и негарантированного улучшения точности. В данной работе исследуется целесообразность учета структуры текстов при классификации двуязычных (русско-английских) библиографических документов, к которым относятся, например, научные публикации, размещаемые на сайтах журналов и в цифровых библиотеках.

Классификация двуязычных библиографических текстовых документов

Дадим краткое описание используемых методов классификации на основе вычисления профилей [15]. Профиль класса состоит из информативных (классоразделяющих) терминов, способных характеризовать (представлять при классификации) все остальные элементы класса. Одним из наиболее известных в литературе профилей является центроид [1].

Для выявления информативных терминов и вычисления их весов на этапе обучения применяются специальные статистические, теоретико-информационные и эвристические критерии, рассчитываемые, как и в случае метода ОКА, на основе таблиц сопряженности размера 2 × 2. Введем следующие обозначения: A — число раз, когда термин $x^{(i)}$ и класс Q_g встречаются вместе; B — число раз, когда $x^{(i)}$ встречается без Q_g; C — число раз, когда Q_g встречается без $x^{(i)}$; D — число раз, когда ни Q_g, ни $x^{(i)}$ не встречаются; N — общее число наблюдений в выборке, χ^2 — величина Хи-квадрат критерия [6].

Приведем формулы для вычисления профилей разными методами с учетом введенных обозначений [15]:

РО-профиль (ρ):

$$\rho = \sqrt{\frac{\chi^2}{N}} = \frac{(AD - CB)}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}. \quad (8)$$

Нормированный МИ-профиль (НМИ):

$$NMI = \frac{A\log\frac{AN}{(A+B)(A+C)}}{(A+B)\log\frac{N}{A+B}}. \quad (9)$$

Профиль Жаккара (J-профиль):

$$J = \frac{A}{A+B+C}. \quad (10)$$

В работе [15] также исследуются специально разработанные комбинированные профили UNI3 и UNI6, созданные на основе РО-, НМИ- и J-профилей. В профиль UNI3 прежде всего отбираются наиболее информативные русские термины, общие для РО- и J-профилей. Затем профиль дополняется самыми важными общими английскими терминами. При этом вес выбирается наибольший из РО- и J-профилей [16].

UN13-профиль:

$$\omega_{UNI3} = \max(\omega_{PO}, \omega_J). \quad (11)$$

Для составления UNI6 используется аналогичная процедура, но на основе трех профилей (РО-, НМИ- и J-) веса исходных профилей суммируются:

UNI6-профиль:

$$\omega_{UNI6} = \omega_{PO} + \omega_{\text{НМИ}} + \omega_J. \quad (12)$$

В формулах (11) и (12) использованы следующие обозначения: ω_{UNI3}, ω_{UNI6} — веса слов в профилях UNI3 и UNI6, ω_{PO}, $\omega_{\text{НМИ}}$, ω_J — веса слов, рассчитанных по формулам (8)—(10).

На этапе классификации вычисляются значения весов классов для каждого нового документа (документа, который не использовался для построения профиля на стадии обучения):

$$W_g = \sum_{i=1}^{Mg} tf_i \mathrm{Pr}of(x^{(i)}, Q_g). \quad (13)$$

Здесь tf_i — частота встречаемости i-го признака в новом документе; M_g — число информативных терминов, включенных в профиль g-го класса на этапе обучения (в наших исследованиях все классы имели профиль одинакового размера $L = M_g$); $\mathrm{Pr}of(x^{(i)}, Q_g)$ — вес i-го термина в профиле, вычисленном по одной из формул (8)—(12).

Согласно решающему правилу (см. соотношение (13)) новый документ относится к тому классу, которому соответствует наибольший вес: $W_g = \max(W_g)$, $g = 1, \ldots, G$.

Рассмотрим возможность использования априорной информации о структуре двуязычных библиографических описаний для проведения классификации. Логично предположить, что термины из различных фрагментов (название, аннотация, ключевые слова) имеют неодинаковую ценность и учет этого факта в формуле взвешивания улучшает результирующую точность. Далее проверяется предположение, что значимость (ценность, информативность) термина зависит не только от величин, вычисляемых по формулам (8)—(12), но и от весовых коэффициентов, характеризующих место появления слова. Предлагается общий вес i-го термина определять следующим образом:

$$\omega_{part}^{(i)} = k_{part}\mathrm{Pr}of_{part}(x^{(i)}, Q_g), \quad (14)$$

где "part" обозначает раздел БО: T — название, A — аннотация, K — ключевые слова и слова, выделенные стилем или шрифтом в тексте БО; k_{part} — весовой коэффициент, повышающий или уменьшающий значимость соответствующих разделов БО; $\mathrm{Pr}of_{part}(x^{(i)}, Q_g)$ — профиль класса, который используется для данного раздела БО.

Решающее правило (13) модифицируется следующим образом:

$$W_g = \sum_{i=1}^{Mg} tf_T^{(i)}\omega_T^{(i)} + tf_A^{(i)}\omega_A^{(i)} + tf_K^{(i)}\omega_K^{(i)} =$$

$$= \sum_{i=1}^{Mg} \alpha tf_T^{(i)}\mathrm{Pr}of_T(x^{(i)}, Q_g) + \beta tf_A^{(i)}\mathrm{Pr}of_A(x^{(i)}, Q_g) +$$

$$+ \gamma tf_K^{(i)}\mathrm{Pr}of_K(x^{(i)}, Q_g), \qquad (15)$$

где α, β, γ — коэффициенты, повышающие или уменьшающие значимость термина в зависимости от его появления в одном из фрагментов документа; $\omega_T^{(i)}$, $\omega_A^{(i)}$, $\omega_K^{(i)}$, $tf_T^{(i)}$, $tf_A^{(i)}$, $tf_K^{(i)}$ — соответственно вес и частота i-го термина в названии, аннотации, ключевых словах классифицируемого документа;

$$\omega_T^{(i)} = \mathrm{Pr}of_T(x^{(i)}, Q_g),$$

$$\omega_A^{(i)} = \mathrm{Pr}of_A(x^{(i)}, Q_g),\ \omega_K^{(i)} = \mathrm{Pr}of_K(x^{(i)}, Q_g).$$

Отметим также, что нами рассматриваются двуязычные БО, поэтому в профили, описывающие разные фрагменты, входят как русские, так и английские информативные слова. Далее проводятся экспериментальные исследования профильных методов; рассчитанных по формулам (8)—(12) и (14), на специально сформированных выборках.

Экспериментальное исследование разработанных методов

Для проведения исследований было составлено по 20 обучающих и экзаменационных выборок, состоящих из двуязычных БО научных статей. Выборки были сформированы из электронных библиотек eLibrary, Киберленинка, а также электронных журналов по 36 инженерно-техническим, естественно-научным и гуманитарным тематикам. Каждая выборка состояла из семи классов, содержащих по 65 и 15 документов для обучения и экзамена. Ошибка рассчитывалась как число неправильных решений метода на экзаменационной выборке.

В ходе экспериментальных исследований ставилась задача подтвердить (или опровергнуть) предположения:

а) в формуле (15) целесообразно использовать различные профили для взвешивания терминов из разных фрагментов БО;

б) введение (и настройка) дополнительных коэффициентов α, β, γ для расчета общего веса термина в формуле (15) увеличивают точность классификации.

В табл. 1 приведены ошибки классификации, полученные с помощью профильных (РО, НМИ, J) и комбинированных (UNI3, UNI6) методов. Для них использовалось решающее правило, рассчитываемое по формуле (13). Кроме того, в табл. 1 содержатся ошибки составных профилей, в которых применяются различные профильные методы для взвешивания терминов из названий, аннотаций и ключевых слов. Например, UNI3-J-UNI6 означает, что для расчета веса слова в названии использовался метод UNI3, в аннотации — J, ключевых слов — UNI6. При проведении исследований значимость трех фрагментов при определении класса документа считалась одинаковой $\alpha = \beta = \gamma = 1$ и использовалось решающее правило, заданное формулой (15).

Как следует из табл. 1, использование различных способов взвешивания фрагментов (с одинаковыми коэффициентами $\alpha = \beta = \gamma = 1$) не дает значимого прироста в точности по сравнению с профилями, рассчитанными по формулам (8)—(12). В связи с этим главный акцент в экспериментальных исследованиях был сделан на снижении ошибки классификации за счет настройки весовых коэффициентов α, β, γ в соотношении (15). В качестве начальных приближений были взяты следующие значения коэффициентов по методу Фишберна <α, β, γ>; <1/6, 2/6, 3/6>; <2/6, 1/6, 3/6>; <1/6, 3/6, 2/6>; <2/6, 3/6, 1/6>; <3/6, 2/6, 1/6>; <3/6, 1/6, 2/6>, при этом ($\alpha + \beta + \gamma = 1$) [17].

Наилучшие результаты для всех методов показал вариант, при котором названиям статей присваивался коэффициент 3/6, аннотациям 1/6 и ключевым словам 2/6, т. е. $\alpha = 3/6$; $\beta = 1/6$; $\gamma = 2/6$. Результаты классификации для данных коэффициентов и выбранных методов приведены в табл. 2.

Целесообразность применения НМИ-профиля, обладающего достаточно большой ошибкой (см. табл. 1), для ключевых слов объясняется тем, что этот профиль устанавливает высокие веса для редких,

Таблица 1

Минимальные, максимальные и средние ошибки классификации

Метод	Ошибки классификации, %		
	Минимальная	Максимальная	Средняя
РО	5,71	20,95	12,04
НМИ	8,57	18,09	12,75
J	7,61	16,19	11,52
UNI3	8,57	16,19	12,13
UNI6	3,81	18,09	10,52
UNI3-J-UNI6	5,71	16,19	10,52
UNI3-J-НМИ	5,71	14,28	10,23
J-UNI6-НМИ	5,71	16,19	10,37

Таблица 2

Минимальные, максимальные и средние ошибки при классификации с использованием коэффициентов Фишберна

Метод	Ошибки классификации, %		
	Минимальная	Максимальная	Средняя
UNI3-J-UNI6-Fishburn	5,71	15,23	10,33
UNI3-J-НМИ-Fishburn	6,67	14,28	9,71
J-UNI6-НМИ-Fishburn	3,81	15,23	9,56

специализированных терминов (и это представляется важным в контексте проводимого исследования). Остальные профили (прежде всего J и UNI6) были отобраны для исследования, так как они относятся к наиболее точным профильным методам. Кроме того, UNI3 и UNI6 являются комбинированными процедурами, учитывающими преимущества статистического, теоретико-информационного и эвристического взвешивания (см. формулы (11)—(12)).

Наши экспериментальные исследования показали, что способ расчета профилей для различных фрагментов БО практически не влияет на результирующую точность. Вместе с тем, введение дополнительных коэффициентов, присваивающих различный вес терминам из названий, аннотаций, ключевых и выделенных слов, позволяет улучшить точность на 0,96 % по сравнению с наиболее точным профильным методом UNI6 (на 1,96, 2,48 и 3,19 % соответственно для "классических" J-, PO- и НМИ-профилей).

Заключение

На наш взгляд, широко распространенная практика использования априорных знаний о структуре документов при решении задач информационного поиска и автоматического аннотирования может быть успешно применена для выявления нечетких дубликатов и увеличения точности классификации текстовых данных. Резюмируя основные результаты, приведенные в статье, отметим следующие:

- исследованы профильные методы, на базе которых предложены новые модификации;
- сформированы обучающие и экзаменационные выборки, состоящие из двуязычных (русско-английских) библиографических описаний;
- экспериментально установлено снижение ошибки классификации для комплексного решающего правила (см. формулу (15)), в котором учитывается структура текста (местоположение термина в БО).

При проведении дальнейших исследований планируется особое внимание уделить уточнению значений коэффициентов α, β, γ в процессе машинного обучения (Machine Learning) на сформированных выборках и проверке полученных результатов на новых коллекциях двуязычных библиографических документов.

Список литературы

1. **Сэлтон Г.** Автоматическая обработка, хранение и поиск информации. М.: Советское радио, 1973. 560 с.

2. **Маннинг К., Рагхаван П., Шютце Х.** Введение в информационный поиск. М.: ВИЛЬЯМС, 2014. 528 с.

3. **Aas K., Eikvil L.** Text Categorization: A Survey. Technical Report Raport NR 941. Norwegian Computing Center. Oslo. 1999. P. 1—37.

4. **Scott S., Matwin S.** Text Classification using WordNet hypernyms // Proceedings of the COLING/ACL WorkShopon Usage of WordNet in Natural Language Processing Systems. Montreal. 1998. P. 45—51.

5. **Барсегян А. А., Куприянов М. С., Холод И. И., Тесс М. Д., Елизаров С. И.** Анализ данных и процессов. СПб: БХВ-Петербург, 2009. 512 с.

6. **Толчеев В. О.** Модели и методы классификации текстовой информации // Информационные технологии. 2004. № 5. С. 6—14.

7. **Iwazume M., Takeda H., Nishida T.** Ontology-based Information Gathering and Text Categorization from the Internet // Proceedings of the 9th ACM International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems. 1996. P. 305—314.

8. **Peng X., Choi B.** Document classifications based on word semantic hierarchies. IASTED International Conference on Artificial Intelligence and Applications 2005. February 14—16. Innsbruck, Austria. 2005. P. 362—367.

9. **Некрасов И. В., Толчеев В. О.** Построение модели представления библиографического документа // Информационные технологии. 2005. № 11. С. 57—63.

10. **Fresno V., Ribero A.** An Analytical Approaches to Concept Extraction in HTML Environments // Journal of Intelligent Information Systems. 2004. N. 22. P. 213—236.

11. **Зеленков Ю. Г., Сегалович И. В.** Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Труды 9-й Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции", Переславль-Залесский: Изд. ИПС РАН. 2007. С. 166—174.

12. **Толчеев В. О.** Анализ проблемы и разработка процедуры выявления нечетких дубликатов научных статей по библиографическим описаниям // Информационные технологии. 2011. № 2. С. 17—21.

13. **Дербенев Н. В., Толчеев В. О.** Разработка метода выявления нечетких дубликатов по библиографическим описаниям // Тр. Междунар. конф. "Интеллектуализация обработки информации", Черногория, Будва: Изд-во Торус. 2012. С. 613—616.

14. **Губин М. В.** Модели и методы представления текстового документа в системах информационного поиска. Автореферат дисс. канд. физ.-мат. наук. СПб., 2005. 15 с.

15. **Мохов А. С., Толчеев В. О.** Разработка методов высокоточной классификации двуязычных текстовых библиографических документов // Информационные технологии. 2014. № 5. С. 8—13.

16. **Мохов А. С., Толчеев В. О., Юров Р. С.** Разработка процедуры взвешивания терминов в зависимости от структуры двуязычного библиографического документа // Тр. XX Байкальской Всероссийской конф. "Информационные и математические технологии в науке и управлении". Ч. 3. Иркутск: ИСЭМ СО РАН. 2015. С. 43—50.

17. **Фишберн П.** Теория полезности для принятия решений: Пер. с англ. Под ред. Н. Н. Воробьева. М.: Наука, 1978. 352 с.

18. **Фомичев В. А.** Математические основы представления смысла текстов для разработки лингвистических информационных технологий. Часть I // Информационные технологии. 2002. № 10. С. 16—25.

19. **Фомичев В. А.** Математические основы представления смысла текстов для разработки лингвистических информационных технологий. Часть II // Информационные технологии. 2002. № 11. С. 34—45.

20. **Fomichov V. A.** Semantics-Oriented Natural Language Processing: Mathematical Models and Algorithms. New York, Dordrecht, Heidelberg, London: Springer, 2010. 354 p.

A. S. Mokhov, Assistant, e-mail: asmokhov@mail.ru,
V. O. Tolcheev, Professor, e-mail: tolcheevvo@mail.ru,
NRU "MPEI"

Approaches to Considering Patterns of Scientific Documents in Processing and Analysis of Text Information

The procedures of processing and analysis of text information, based on document structure are considered in the article. The main models of text document representation are reviewed. The effectiveness of partly structured models is shown for such tasks as information retrieval, automatic annotation and fuzzy duplicates identification. Along with the known approaches the new authors' methods are presented. These methods successfully consider distinctive features of bilingual bibliographic documents and decrease classification error.

Keywords: *Text mining, text document representation, informational retrieval, automatic annotation, fuzzy duplicates identification, classification of bilingual bibliographic documents*

References

1. **Salton G.** *Avtomaticheskaya obrabotka, hraneniye i poisk informatcii*. Moscow: Sovetskoe radio. 1973. 560 p.
2. **Manning K., Raghavan P., Shutce H.** *Vvedenie v informatcionniy poisk*. Moscow: Wiliams, 2014. 528 p.
3. Aas K., Eikvil L. *Text Categorization: A Survey*. Norwegian Computing Center. Oslo. 1999, pp. 1—37.
4. **Scott S., Matwin S.** Text Classification using WordNet hypernyms, *Proceedings of the COLING/ACLWorkShop on Usage of WordNet in Natural Language Processing Systems*. Montreal, 1998, pp. 45—51.
5. **Barsegjan A. A., Kupriyanov M. S., Holod I. I., Tess M. D., Elizarov S. I.** *Analiz dannih i processov*. SPb: BHV-Peterburg, 2009. 512 s.
6. **Tolcheev V. O.** Modeli i metodi classifikacii tekstovoi informatcii. *Informacionnie tekhnologii*, 2004, no. 5, pp. 6—14.
7. **Iwazume M., Takeda H., Nishida T.** Ontology-based Information Gathering and Text Categorization from the Internet. *Proceedings of the 9th ACM International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, 1996, pp. 305—314.
8. **Peng X., Choi B.** *Document classifications based on word semantic hierarchies IASTED International Conference on Artificial Intelligence and Applications*, 2005, pp. 362—367.
9. **Nekrasov I. V., Tolcneev V. O.** Postroenie modeli predstavleniya bibliograficheskogo dokumenta, *Informatcionnie tekhnologii*, 2005, no. 11, pp. 57—63.
10. **Fresno V., Ribero A.** An Analytical Approaches to Concept Extraction in HTML Environments. *Journal of Intelligent Information Systems*, 2004, no. 22, pp. 213—236.
11. **Zelenkov Y. G., Segalovich I. V.** Sravnitelniy analiz metodov opredeleniya nechetkih dublikatov dlya Web-documentov. Trud*y 9-y Vserossiyskoy nauchnoy konferencii "Electronnie biblioteki: perspektivniye metody tekhnologii, electronnie kollektcii"*, Pereslavl-Zalesskiy, Izd-vo IPS RAN, 2007, pp. 166—174.
12. **Tolcheev V. O.** Analiz problemi i razrabotka procedury viyavleniya nechetkih dublikatov po bibliograficheskim opisaniyam, *Informatcionnie tekhnologii*, 2011, no. 2, pp. 17—21.
13. **Derbenev N. V., Tolcheev V. O.** Razrabotka metoda viyavleniya nechetkih dublikatov po bibliograficheskim opisaniyam. *Trudy mezhdunarodnoy konferentcii "Intellektualizatciya obrabotki informatcii",* Chernogoriya, Budva: Izd-vo Tous. 2012, pp. 613—616.
14. **Gubin M. V.** *Modeli i metody predstavleniya tekstvogo dokumenta v sistemah infoimatcionnogo poiska*. Avtoreferat diss. kand. fiz-mat nauk. SPb., 2005 g. 15 p.
15. **Molhov A. S., Tolcheev V. O.** Razrabotka metodov visokotochnoy classifikacii dvuyazichnikh tekstovikh bibliographicheskikh dokumentov, *Informatcionnie tekhnologii*. 2014, no. 5, pp. 8—13.
16. **Mokhov A. S., Tolcheev V. O., Yurov R.S.** Razrabotka procedury vzveshivaniya terminov v zavisimosti ot struktury dvuyazychnogo bibliographicheskogo dokumenta, *Trudy XX Baykalstoy vserossiyskoy konferentcii "Informatcionnie i matematicheskie technologii v nauke I upravlenii"*. Ch. 3. Irkutsk: ISEM SO RAN, 2015, pp. 43—50.
17. **Fishburn P.** *Teoriyapoleznosti dlya prinyatiya resheniy*. Per. s angl. pod red. N. N. Vorobieva. Moscow: Nauka, 1978. 352 p.
18. **Fomichev V. A.** Matematicheskie osnovi predstavleniya smysla tekstov dlya razrabotki lingvisticheskikh informatcionnyh tehnologiy. Chast' I, *Informatsionnie tekhnologii*. 2002, no. 10, p. 16—25.
19. **Fomichev V. A.** Matematicheskie osnovi predstavleniya smysla tekstov dlya razrabotki lingvisticheskikh informatcionnyh tehnologiy. Chast' II, *Informatsionnie tekhnologii*. 2002, no. 11, p. 34—45.
20. **Fomichov V. A.** *Semantics-Oriented Natural Language Processing: Mathematical Models and Algorithms*. New York. Dordrecht. Heidelberg. London: Springer, 2010, 354 p.