УДК 519.248.6: 519.766.4

И. В. Поляков, **Т. В. Соколова**, **А. А. Чеповский**, **А. М. Чеповский**

Национальный исследовательский университет
Высшая школа экономики
ул. Мясницкая, 20, Москва, 101000, Россия

achepovskiy@hse.ru

ПРОБЛЕМА КЛАССИФИКАЦИИ ТЕКСТОВ И ДИФФЕРЕНЦИРУЮЩИЕ ПРИЗНАКИ

Описан метод классификации текстов на естественных языках, основанный на методе взаимной информации. Показано, что псевдоосновы, выделенные аналитическим алгоритмом морфологического анализа, являются универсальными дифференцирующими признаками при классификации текстовых сообщений.

Ключевые слова: классификация текстов, метод взаимной информации.

Введение

В информационных системах различного типа, предназначенных для обработки в автоматическом режиме больших объемов текстов на естественных языках, актуальны различные задачи распознавания текстовой информации [1; 2]. Задача классификации текстовых сообщений по тематике – это задача распознавания образов, которая может решаться с единых позиций теории информации и алгебраической теории [3–5].

Требование автоматизации процессов обработки текстовой информации придают особую важность проблемам классификации текстов на естественном языке по тематике, авторству, стилю и жанру письма.

Актуальной проблемой является классификация по тематической, психолингвистической направленности коротких текстов [6; 7], особенно состоящих из одного-двух предложений или даже нескольких слов. Такие тексты встречаются в комментариях к интернет-блогам, на форумах.

Принципы построения систем классификации больших объемов текстовой информации довольно универсальны [1; 5; 6]. Принадлежность к тому или иному классу определяется выделенными наборами признаков. Поэтому интерес представляют как алгоритмы решения данной задачи [6; 8–10], так и выбор тех дифференцирующих признаков, которые определяют отнесение текстов к заданным рубрикам [6; 8; 10]. Выбор дифференцирующих признаков является ключевым для создания методик классификации текстов на естественных языках, если алгоритм не опирается на сложную словарную систему, как в [9]. Отметим, что задача классификации и рубрикации текстов чаще всего привязана к конкретному естественному языку [6; 8; 10].

В случае текстов в качестве признаков обычно рассматриваются слова и взаимосвязанные наборы слов, содержащиеся в текстах. Отметим, что для решения задачи классификации текстов по авторству предлагалось использовать частоты проявления символов и сочетаний символов языка, применение буквосочетаний символов [8]. Исследования различных грам-

Поляков И. В., *Соколова Т. В.*, *Чеповский А. А.*, *Чеповский А. М.* Проблема классификации текстов и дифференцирующие признаки // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2015. Т. 13, вып. 2. С. 55–63.

матических форм в качестве дифференцирующих признаков для тематической классификации текстов на русском языке проводились в [6; 10; 11]. Представляет интерес определение таких универсальных методов и дифференцирующих признаков, которые могут рассматриваться с позиций единых методик для различных естественных языков.

В [10; 12] задача идентификации языка и классификации текстов решалась на базе вероятностной модели строки текста и Байесовского классификатора.

В данной работе мы рассматриваем методику классификации текстовых сообщений на различных естественных языках на базе метода взаимной информации с целью выявить нарушения законодательства в сети Интернет и анализируем выбор наиболее эффективных и универсальных дифференцирующих признаков.

Сформулируем основные положения метода взаимной информации, основанного на алгебраической теории информации [5] и алгебраической теории распознавания образов [3; 4].

Информационная матрица

Введем обозначения для количественных характеристик документов, которые характеризуются L признаками X_j (j = 1, ..., L) и могут принадлежать K классам Y_i (i = 1, ..., K):

R^X_{ij} – количество документов, принадлежащих классу Y_i, содержащих признак X_j;

B^X_{ij} – количество документов, не релевантных классу Y_i, содержащих признак X_j;

R^e_{ij} – количество документов, принадлежащих классу Y_i, не содержащих признак X_j;

B^e_{ij} – количество документов, не релевантных классу Y_i и не содержащих признак X_j.

Общее количество текстов, принадлежащих к классу Y_i:

$$R_i = R^X_{ij} + R^e_{ij}. \tag{1}$$

Общее количество текстов, не принадлежащих к классу Y_i:

$$B_i = B^X_{ij} + B^e_{ij}. \tag{2}$$

Условная вероятность того, что текст, содержащий признак X_j, принадлежит классу Y_i:

$$P\left(Y_i \middle| X_j\right) = \frac{R^X_{ij}}{R^X_{ij} + B^X_{ij}}. \tag{3}$$

Условная вероятность того, что текст, содержащий признак X_j, не принадлежит классу Y_i:

$$\overline{P}\left(Y_i \middle| X_j\right) = 1 - P\left(Y_i \middle| X_j\right) = \frac{B^X_{ij}}{R^X_{ij} + B^X_{ij}}. \tag{4}$$

Условная вероятность того, что текст, не содержащий признак X_j, принадлежит классу Y_i:

$$Q\left(Y_i \middle| X_j\right) = \frac{R^e_{ij}}{R^e_{ij} + B^e_{ij}}. \tag{5}$$

Условная вероятность того, что текст, не содержащий признак X_j, не принадлежит классу Y_i:

$$\overline{Q}\left(Y_i \middle| X_j\right) = 1 - Q\left(Y_i \middle| X_j\right) = \frac{B^e_{ij}}{R^e_{ij} + B^e_{ij}}. \tag{6}$$

Для оценки соответствия документа заданному классу вводится критерий релевантности, который показывает соответствие получаемого результата желаемому. Предполагаем, что признак X_j тем более релевантен классу Y_i, чем выше вероятность непринадлежности классу при отсутствии признака и принадлежности классу при наличии признака. С другой стороны, признак X_j тем более релевантен классу Y_i, чем ниже вероятность принадлежности классу при отсутствии признака и непринадлежности классу при наличии признака.

Тогда коэффициент релевантности определится как отношение произведений вероятностей:

$$\rho_{ij} = \frac{P\left(Y_i \middle| X_j\right)\overline{Q}\left(Y_i \middle| X_j\right)}{\overline{P}\left(Y_i \middle| X_j\right)Q\left(Y_i \middle| X_j\right)} = \frac{R^X_{ij} \,/\, R^e_{ij}}{B^X_{ij} \,/\, B^e_{ij}} = \frac{R^X_{ij} B^e_{ij}}{B^X_{ij} R^e_{ij}}. \tag{7}$$

Для оценки взаимосвязи между признаками используется коэффициент корреляции, который определим как фита-коэффициент (phi-coefficient) Пирсона [13]:

$$\varepsilon_{ij} = \frac{\left(R_{ij}^X B_{ij}^e - R_{ij}^e B_{ij}^X\right)}{\sqrt{\left(R_{ij}^X + R_{ij}^e\right)\left(B_{ij}^X + B_{ij}^e\right)\left(R_{ij}^X + B_{ij}^X\right)\left(R_{ij}^e + B_{ij}^e\right)}}. \quad (8)$$

Значения данного коэффициента изменяются в интервале $[-1,1]$. Значения, близкие к 1, показывают сильную взаимосвязь признаков класса.

Для обоих введенных коэффициентов (релевантности и корреляции) справедливо утверждение, что большие их значения соответствуют признакам, наиболее точно характеризующим данный класс.

Элементы информационной матрицы I_{ij} определяются как пара – коэффициент релевантности и коэффициент корреляции:

$$I_{ij} = \left\{\rho_{ij}, \varepsilon_{ij}\right\}. \quad (9)$$

Принадлежность документа к данному классу будет определяться наличием в нем признаков, релевантных данному классу и коррелирующих с признаками рассматриваемого класса, а также отсутствием нерелевантных признаков и признаков, не коррелирующих с признаками данного класса.

Алгоритмы обучения и классификации

На основании (1)–(9) можно сформулировать алгоритм обучения системы классификации.

Алгоритм 1. Алгоритм обучения системы классификации.

Для каждой пары Y_i, X_j (класс, признак) по формулам изложенным выше вычислим коэффициенты релевантности (7) и корреляции (8):

$$\rho_{ij} = \frac{R_{ij}^X B_{ij}^e}{B_{ij}^X R_{ij}^e};$$

$$\varepsilon_{ij} = \frac{\left(R_{ij}^X B_{ij}^e - R_{ij}^e B_{ij}^X\right)}{\sqrt{\left(R_{ij}^X + R_{ij}^e\right)\left(B_{ij}^X + B_{ij}^e\right)\left(R_{ij}^X + B_{ij}^X\right)\left(R_{ij}^e + B_{ij}^e\right)}}.$$

Сформируем информационную матрицу, состоящую из найденных коэффициентов:

$$I_{ij} = \left\{\rho_{ij}, \varepsilon_{ij}\right\}.$$

Процедура обучения завершается. Результатом процедуры обучения является данная матрица.

Определим показатели соответствия классам для анализируемого n-го текста, имеющего m_{nj} – количество вхождений j-го признака в анализируемый n-й текст.

Коэффициент релевантности текста относительно класса Y_i запишется как скалярное произведение вектора признаков n-го текста $\{m_{nj}\}$ и строки коэффициентов релевантности, соответствующей классу Y_i в информационной матрице:

$$\sigma_{in} = \sum_{j=1}^{L}\left(\rho_{ij} \cdot m_{nj}\right).$$

Считается, что n-й текст релевантен классу Y_i, если

$$\sigma_{in} \in D_i = \left[d_i^{\min}, \infty\right),$$

где D_i – область значений коэффициента релевантности, характеризующих принадлежность к i-му классу.

Порог релевантности:

$$\delta_i = d_i^{\min}.$$

Коэффициент корреляции текста относительно класса Y_i запишется как скалярное произведение вектора признаков n-го текста $\{m_{nj}\}$ и строки коэффициентов корреляции, соответствующей классу Y_i в информационной матрице:

$$\theta_{in} = \sum_{j=1}^{L}\left(\varepsilon_{ij} \cdot m_{nj}\right).$$

Считается, что *n*-й текст коррелирует с классом Y_i, если

$$\theta_{in} \in G_i = \left[g_i^{\min}, \infty \right),$$

где G_i – область значений коэффициента корреляции, характеризующих принадлежность к *i*-му классу.

Порог корреляции:

$$\gamma_i = g_i^{\min}.$$

Пороги релевантности и корреляции служат параметрами, управляющими точностью обучения и классификации.

В процессе классификации каждого текста вычисляются его коэффициенты релевантности и корреляции относительно каждого из классов. Документ относится к тем классам, для которых произошло превышение пороговых значений как по коэффициенту корреляции, так и по коэффициенту релевантности.

Алгоритм 2. Алгоритм классификации *n*-го документа:

для каждого признака X_j найдем вектор $\{m_{nj}\}$,

для каждого класса Y_i вычислим векторы

$\sigma_{in} = \sum_{j=1}^{L} \left(\rho_{ij} \cdot m_{nj} \right)$;

$\theta_{in} = \sum_{j=1}^{L} \left(\varepsilon_{ij} \cdot m_{nj} \right)$;

if $\sigma_{in} \in D_i$ and $\theta_{in} \in G_i$:

считаем, что *n*-й текст релевантен *i*-му классу.

Суть алгоритма состоит в нахождении признаков, имеющих вхождения в данный текст. После этого коэффициенты релевантности и корреляции текста с данным классом вычисляются как суммы соответствующих коэффициентов для данного класса по всем вхождениям признаков. Если по обеим характеристикам произошло превышение пороговых значений, то текст относится к данному классу.

Пороговые значения для каждого класса могут быть заданы пользователем или же рассчитаны автоматически по обучающей выборке (таким образом, чтобы высокий процент документов, принадлежащих классу Y_i, при последующей классификации были отнесены к этому классу).

Показатели качества классификации

Оценку качества распознавания кодировки естественного языка и тематики текстов будем производить по аналогии с оценками качества документальных информационно-поисковых систем, предлагаемых в [1; 2].

В результате процедуры распознавания каждый из анализируемых файлов относится к конкретному *i*-тому классу Y_i (набору файлов). В каждом *i*-м наборе N_i файлов содержится R_i файлов, которые соответствуют данному *i*-му набору файлов (с текстами соответствующей тематики).

Для тестирования выбирается некоторая исходная коллекция. Исходная тестовая коллекция содержит вполне определенные наборы объемом N_i^{IS} каждого *i*-го типа файлов, принадлежащих классу Y_i. Для каждого используемого тестового корпуса можно подсчитать и число N_i^B файлов, не соответствующих *i*-му типу файлов (не принадлежащих *i*-му классу Y_i). Данные величины определяются только для размеченной тестовой коллекции.

Для оценки качества распознавания используются следующие показатели: коэффициент релевантности, коэффициент полноты, усредненная точность.

Точность классификации для данного класса (коэффициент релевантности) при определении заданного *i*-го типа файлов измеряет в результирующем *i*-м наборе объемом N_i файлов долю файлов, которые действительно являются файлами данного типа и измеряются количеством R_i файлов с текстами, соответствующих *i*-му классу:

$$A_i = \frac{R_i}{N_i}. \tag{10}$$

В задачах информационного поиска аналог введенного коэффициента релевантности (10) иногда называют точностью информационного поиска, а в задачах классификации аналогичный коэффициент носит название точность классификации для данного класса.

Коэффициент полноты при создании *i*-го набора файлов измеряет, какую долю количество R_i файлов данного типа из результирующего *i*-го набора составляет в исходном тестовом наборе файлов *i*-го типа объемом N_i^{IS}:

$$C_i = \frac{R_i}{N_i^{IS}}.$$

Усредненная точность определяется как взвешенная гармоническая средняя коэффициента релевантности и коэффициента полноты. Будем использовать для оценок сбалансированную точность, которая предполагает равный вес, как для коэффициента релевантности, так и для коэффициента полноты:

$$F_i = \frac{2A_iC_i}{A_i + C_i}. \tag{11}$$

Все приведенные коэффициенты – коэффициенты релевантности, полноты, усредненная точность – можно рассматривать как вероятностные оценки качества работы процедур, программного обеспечения распознавания.

Рассмотрим способ оценки взаимного качества классификации различными алгоритмами с различными дифференцирующими признаками. Будем называть конкретным «классификатором» один из методов классификации (байесовский классификатор или классификатор на основе взаимной информации) с конкретным набором признаков.

Для сравнения работы различных классификаторов на реальном потоке данных введем коэффициент сравнения результатов работы двух классификаторов, обозначив их символами *a* и *b*:

$$P_i\left(b\middle|a\right) = \frac{\left|A_i \cap B_i\right|}{\left|A_i\right|} = \frac{R_i\left[ab\right]}{R_i\left[a\right]}, \tag{12}$$

где

A_i – множество текстов, принадлежащих к классу *i* согласно классификатору *a*;

B_i – множество текстов, принадлежащих к классу *i* согласно классификатору *b*;

$R_i\left[a\right]$ – количество текстов, отнесенных классификатором *a* к классу *i*;

$R_i\left[ab\right]$ – количество текстов, отнесенных классификаторами *a* и *b* к классу *i*.

$P_i\left(b\middle|a\right)$ показывает отношение количества текстов, отнесенных к классу *i* классификаторами и *a* и *b*, к количеству всех текстов, отнесенных к *i* классификатором *a*.

Результаты экспериментальных исследований

Рассмотрим следующие варианты дифференцирующих признаков для русского языка с введенными обозначениями: *N* – существительные; *NA* – существительные и прилагательные; *NAV* – существительные, прилагательные и глаголы; *NNP* – существительные и именные группы; *VVP* – глаголы и глагольные группы; *Stem* – псевдоосновы словоупотреблений текста, полученные алгоритмами аналитического морфологического анализа [14; 15].

Для перечисленных выше дифференцирующих признаков рассмотрим два алгоритма классификации: байесовский метод, реализованной на основе предложенной в [10; 12] модели строки текста; метод взаимной информации, описанный в данной статье. Первый метод обозначим префиксом *B*, а второй – префиксом *I*. Конкретным классификатором будем называть совокупность метода классификации (префиксы *B* и *I*) и набор дифференцирующих признаков (идентификаторы *N*, *NA*, *NAV*, *NNP*, *VVP*, *Stem*).

Обучение классификаторов проводилось на созданных экспертами выборках текстов объемами: для русского языка – 57,3 Мб; башкирского – 1,87 Мб; татарского – 2,68 Мб. Обуче-

ние осуществлялось для следующих классов: наркотики; насилие, жестокость; национализм, социальная рознь; отрицание традиционных ценностей; порнография; терроризм; фашизм; экстремизм. Классы определялись необходимостью выявления тематик, которые интересны для задачи определения нарушений законодательства в текстах в мировой сети электронной коммуникации и носят несколько условные названия.

Результаты классификации текстов методом на основе взаимной информации по рубрикам нарушения законодательства в сети Интернет для перечисленных выше дифференцирующих признаков приведены в табл. 1. Эксперименты проводились на подготовленном экспертами тестовом наборе данных, составляющем в сумме 300 файлов текстовой информации. В табл. 1 приведены значения усредненной точности (11) для полноты и точности, подсчитанных для случаев отнесения текста хотя бы к одному из рассматриваемых классов.

Очевидно, что учет различных морфологических признаков оказывает различное влияние на показатели классификации в зависимости от класса. Для некоторых тематик могут оказывать положительное влияние на показатели классификации существительные и именные группы (например рубрика «фашизм»), а на определение некоторых тематик (например рубрики «наркотики», «фашизм») отрицательное влияние оказывает учет глагольных групп.

В табл. 2 приведены значения коэффициента сравнения (12) для набора классификаторов, которые демонстрируют, насколько совпадают результаты классификации для разных методов с различными наборами дифференцирующих признаков. Результаты получены для конкретной рубрики «терроризм» на реальном наборе текстов с сайтов Интернета общим объемом текстов около 100 000 суммарным объемом около 2 Гб. При обработке такого объема текстов отнесенными к каждой рубрике получалось от 200 текстов байесовским классификатором на основе псевдооснов до почти 5 000 текстов при учете глаголов и глагольных групп.

Таблица 1

Значение усредненной точности для классификации текстов
методом на основе взаимной информации для различных дифференцирующих признаков

Рубрика	N	NA	NAV	NNP	VVP	Stem
Наркотики	0,666	0,666	0,654	0,719	0,617	0,765
Насилие, жестокость	0,901	0,893	0,863	0,896	0,823	0,913
Национализм, социальная рознь	0,796	0,827	0,841	0,755	0,808	0,804
Отрицание традиционных ценностей	0,806	0,763	0,780	0,778	0,865	0,862
Порнография	0,832	0,876	0,902	0,829	0,847	0,754
Терроризм	0,938	0,929	0,923	0,888	0,828	0,905
Фашизм	0,974	0,903	0,968	0,977	0,681	0,909
Экстремизм	0,638	0,660	0,707	0,724	0,750	0,740

Таблица 2

Сравнения классификаторов для текстов на русском языке,
отнесенных к классу «терроризм»

P	*B_NA*	*B_NNP*	*B_VVP*	*I_NA*	*I_NNP*	*I_VVP*	*B_Stem*	*I_Stem*
B_NA	1	0,875	0,609	0,336	0,447	0,358	0,87	0,32
B_NNP	0,875	1	0,638	0,340	0,456	0,366	0,852	0,303
B_VVP	0,750	0,786	1	0,364	0,447	0,398	0,852	0,345
I_NA	0,839	0,857	0,739	1	0,942	0,829	0,833	0,908
I_NNP	0,821	0,839	0,667	0,681	1	0,650	0,815	0,648
I_VVP	0,786	0,804	0,710	0,729	0,777	1	0,833	0,761
B_Stem	0,839	0,821	0,667	0,321	0,427	0,336	1	0,317
I_Stem	0,821	0,804	0,710	0.921	0,893	0,878	0,833	1

Таблица 3

Сравнения классификаторов для текстов на башкирском языке, отнесенных к классам «терроризм» и «экстремизм»

P	*B_Stem*	*I_Stem*	*Ngram*	*B_Stem*	*I_Stem*	*Ngram*
	терроризм			экстремизм		
B_Stem	1,000	0,500	0,625	1,000	0,800	0,769
I_Stem	0,571	1,000	0,500	0,667	1,000	0,675
Ngram	0,714	0,500	1,000	0,833	0,800	1,000

В табл. 3 приведены значения коэффициента сравнения (12) результатов распределения текстов на башкирском языке байесовским классификатором и методом взаимной информации с использованием в качестве дифференцирующих признаков псевдооснов, выделенных в автоматическом режиме алгоритмом аналитического морфологического анализа. Для сопоставления классификаторов в табл. 3 помещены результаты сравнения с классификатором на основе учета буквосочетаний символов (обозначено *Ngram*) по методике работ [10; 12].

Рассматривая табл. 2, можно сделать вывод о том, что большинство классификаторов дают близкое отнесение файлов к одному классу для различных дифференцирующих признаков. Явно завышается по сравнению с другими классификаторами количество отнесенных к рубрике файлов при учете глагольных групп. Очевидно, что использование псевдооснов в качестве дифференцирующих признаков дает результаты, хорошо согласующиеся с классификаторами на основе других дифференцирующих признаков, что подтверждается приведенными в табл. 2 и 3 значениями. Поэтому можно делать вывод о возможности использования псевдооснов словоупотреблений в качестве дифференцирующих признаков при тематической классификации текстов.

Заключение

Тестирование на реальных данных проводилось для текстов на русском, английском, татарском и башкирском языках по тематикам наркотики, насилие, национализм, отрицание ценностей, порнография, терроризм, фашизм, экстремизм.

Тестирование на реальном потоке русскоязычных текстов показывает достаточно хорошее совпадение результатов одного и того же классификатора с различными наборами дифференцирующих признаков, включающих такие признаки, как существительные, именные группы, прилагательные. Явно ухудшают результаты классификации по некоторым тематикам учет глагольных групп. Сравнительно высокие результаты получаются при использовании псевдооснов в качестве дифференцирующего признака.

По результатам исследований утверждается, что псевдоосновы, выделенные аналитическим алгоритмом морфологического анализа, являются универсальными дифференцирующими признаками при классификации текстовых сообщений на различных естественных языках.

Список литературы

1. *Корнеев В. В.*, *Гареев А. Ф.*, *Васютин С. В.*, *Райх В. В.* Базы данных. Интеллектуальная обработка информации. М.: Нолидж, 2001. 496 с.

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

3. *Журавлев Ю. И.* Избранные научные труды. М.: Магистр, 1998. 420 с.

4. *Журавлев Ю. И.* Об алгебраическом подходе к решению задач распознавания и классификации // Проблемы кибернетики. 1978. Вып. 33. С. 5–68.

5. *Гоппа В. Д.* Введение в алгебраическую теорию информации. М.: Наука, 1995. 112 с.

6. *Бабенко М.*, *Куршев Е.*, *Одинцов О.*, *Сулейманова Е.*, *Чеповский А.* Система классификации текстов информационных сообщений на русском языке «АКТИС» // Тр. Междунар. конф. «Программные системы: теория и приложения». М.: Физматлит, 2004. Т. 2. С. 7–20.

7. *Мбайкоджи Э.*, *Драль А. А.*, *Соченков И. В.* Метод автоматической классификации коротких текстовых сообщений // Информационные технологии и вычислительные системы 2012. № 3. С. 93–102.

8. *Батура Т. В.* Формальные методы определения авторства текста // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2012. Т. 10, вып. 4. С. 81–94.

9. *Боярский К. К.*, *Каневский Е. А.*, *Саганенко Г. И.* К вопросу автоматической классификации текстов // Экономико-математические исследования: математические модели и информационные технологии. VII. СПб.: Нестор-История, 2009. С. 252–273.

10. *Гусев С. В.*, *Поляков И. В.*, *Чеповский А. М.* Применение статистической модели текста в информационных системах // Ершовская конференция по информатике 2011. Рабочий семинар «Наукоемкое программное обеспечение». Новосибирск, 2011. С. 69–72.

11. *Андреев А. М.*, *Березкин Д. В.*, *Сюзев В. В.*, *Шабанов В. И.* Модели и методы автоматической классификации текстовых документов // Вестн. МГТУ им. Н. Э. Баумана. Сер. «Приборостроение». 2003. № 4. С. 64–94.

12. *Гусев С. В.*, *Чеповский А. М.* Модель для идентификации естественного языка текста // Бизнес-информатика. 2011. № 3 (17). С. 31–35.

13. *Chen P. Y.*, *Popovich P. M.* Correlation. Parametric and Nonparametric Measures. Sage university papers series // Quantitative applications in the social sciences, 07-139. Thousand Oaks, CA: Sage, 2002. 95 p.

14. *Болховитянов А. В.*, *Чеповский А. М.* Методы автоматического анализа словоформ // Информационные технологии. 2011. № 4 (176). С. 24–29.

15. *Болховитянов А. В.*, *Чеповский А. М.* Алгоритмы морфологического анализа компьютерной лингвистики: Учеб. пособие. М., 2013. 198 с.

Материал поступил в редколлегию 22.04.2015

I. V. Polyakov, **T. V. Sokolova**, **A. A. Chepovskiy**, **A. M. Chepovskiy**

National Research University Higher School of Economics
20 Myasnitskaya Str., Moscow, 101000, Russian Federation

achepovskiy@hse.ru

TEXT CLASSIFICATION PROBLEM AND FEATURES SET

This paper presents a text classification method based on mutual information method. It was shown that word stems are universal features for text classification problem.

Keywords: Text classification, Mutual information method.

References

1. Korneev V. V., Gareev A. F., Vasutin S. V., Rihe V. V. Bazi dannikh. Intellektual'naya obrabotka informacii. Moscow, Izdatel'stvo Nolidg, 2001, 496 p. (In Russ.)

2. Manning K., Ragkhavan P., Shutze K. Vvedenie v informacionniy poisk. Moscow, ID Vilyams, 2014, 528 p. (In Russ.)

3. Guravlev Yu. I. Izbrannie nauchnie trudi. Moscow, Izdatel'stvo Magistr, 1998, 420 p. (In Russ.)

4. Guravlev Yu. I. Ob algebraicheskom podkhode k resheniu zadach raspoznovaniya i klassifikacii. *Problemi kibernetiki*, 1978, no. 33, p. 5 68. (In Russ.)

5. Goppa V. D. Vvedenie v algebraicheskuyu teoriu informacii. Moscow, Nauka, 1995, 112 p. (In Russ.)

6. Babenko M., Kurshev E., Odincov O., Syleimanova E., Chepovskiy A. Systemi klassifikacii tekstov informacionnikh soobsheniy na russkom yazike «AKTIS». *Trudi megdunarodnoi*

konferencii «Programmnie systemi: teoriya i prilogeniya», Moscow, Fizmatlit, 2004, vol. 2, p. 7–20. (In Russ.)

7. Mbaikodgy A., Dral A. A., Sochenkov I. V. Metod avtomaticheskoi klassifikacii korotkikh tekstovikh soobsheniy. *Informacionnie tekhnologii i vichislitel'nie systemi*, 2012, no. 3, p. 93–102. (In Russ.)

8. Batura T. V. Formal'nie metodi opredeleniya avtorstva teksta. *Vestnik of Novosibirsk State University. Series: Information Technology*, 2012, vol. 10, no. 4, p. 81–94. (In Russ.)

9. Boyarskiy K. K., Kanevskiy E. A., Sagaenko G. I. K voprosu avtomaticheskoi klassifikacii. *Akonomiko-matematicheskie issledovaniya: matematicheskie modeli i informacionnie tekhnologii*. VII. St.-Petersburg, Nestor-Istoriya, 2009, p. 252–273. (In Russ.)

10. Gusev S. V., Polyakov I. V., Chepovskiy A. Primenenie statisticheskoi modeli teksta v informacionnikh systemakh. *Ershovskaya konferenciya po informatike 2011. Rabochiy seminar «Naukoemkoe programmnoe obespechenie»*. Novosibirsk, 2011, p. 69–72. (In Russ.)

11. Andreev A. M., Berezkin D. V., Suzev V. V., Shabanov V. I. Modeli i metodi avtomaticheskoi klasssifikacii tekstovikh dokumentov. *Vestnik Bauman MGTU. Ser. «Priborostroenie»*, 2003, no. 4, p. 64–94. (In Russ.)

12. Gusev S. V., Chepovskiy A. M. Modeli dlya identifikacii estestvennogo yazika teksta. *Biznes-informatika*, 2011, no. 3 (17), p. 31–35. (In Russ.)

13. Chen P. Y., Popovich P. M. Correlation. Parametric and Nonparametric Measures. Sage university papers series. *Quantitative applications in the social sciences*, 07-139. Thousand Oaks, CA, Sage, 2002, 95 p.

14. Bolkhovityanov A. V., Chepovskiy A. M. Metodi avtomaticheskogo analiza slovoform. *Informacionnie tehnologii*, 2011, no. 4 (176), p. 24–29. (In Russ.)

15. Bolkhovityanov A. V., Chepovskiy A. M. Algoritmi morfologicheskogo analiza komp'uternoi lingvistiki. Moscow, 2013, 198 p. (In Russ.)