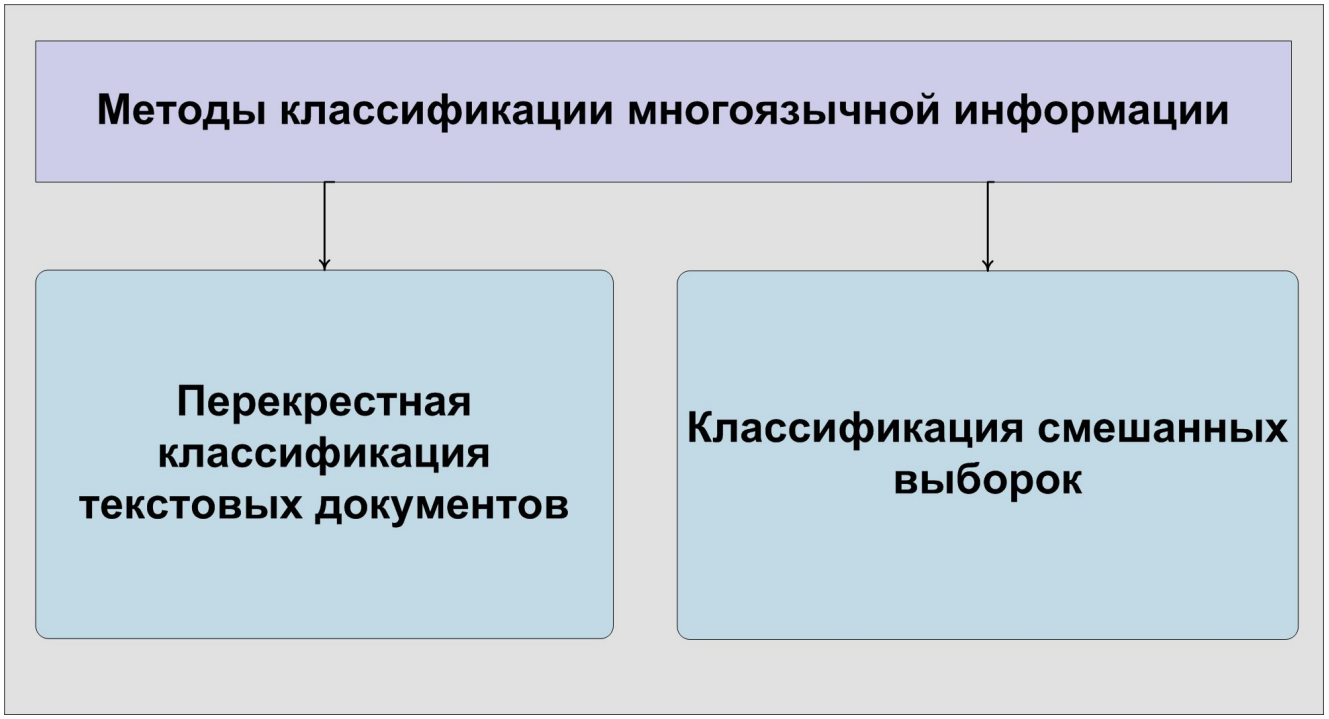
Мохов А.С. Толчеев В.О., НИУ МЭИ

РАЗРАБОТКА МЕТОДОВ ВЫСОКОТОЧНОЙ КЛАССИФИКАЦИИ ДВУЯЗЫЧНЫХ ТЕКСТОВЫХ БИБЛИОГРАФИЧЕСКИХ ДОКУМЕНТОВ

ЦЕЛЬ РАБОТЫ:

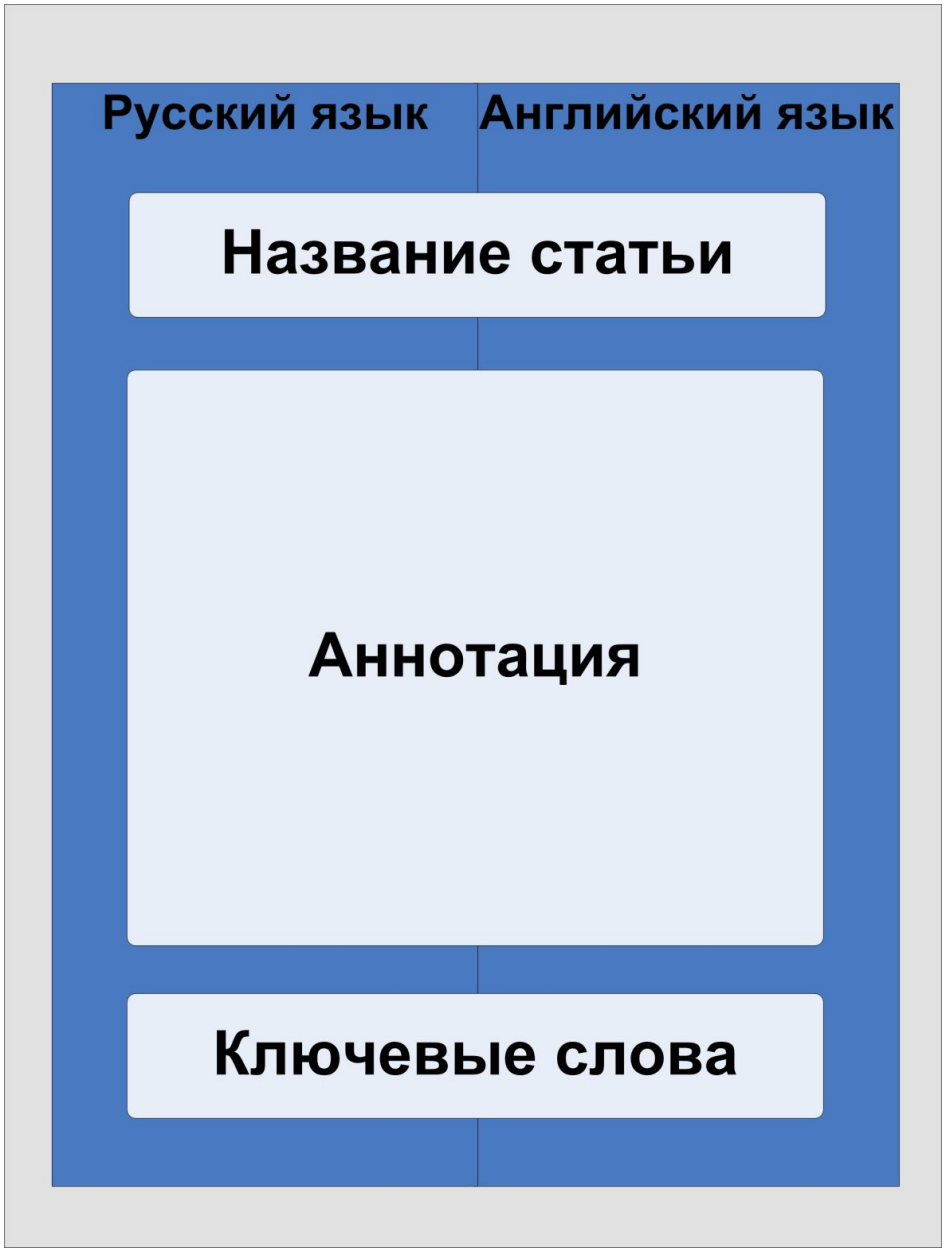
ИССЛЕДОВАНИЕ И РАЗРАБОТКА ПОДХОДОВ К ПОВЫШЕНИЮ ТОЧНОСТИ КЛАССИФИКАЦИИ ДВУЯЗЫЧНЫХ НАУЧНЫХ ТЕКСТОВ.

Основные направления работ



1) **Обучающая выборка – на исходном языке (английский, французский), экзаменационная – на целевом (вьетнамский, венгерский);**
2) **Смешанная обучающая выборка - документы представлены одновременно на двух языках.**

Структура библиографического описания



Библиографические описания научных статей – обычно состоят из названия статьи, аннотации и ключевых слов, приведенные одновременно на русском и английском языках.

Описания могут быть неполными – название может быть приведено только на русском, а ключевые слова отсутствовать полностью.

Состав выборок

Название класса	Размер класса			Состав выборки						
	Русский	Английский	Смешанный	1	2	3	4	5	6	7
Text mining	1048	846	1866	x		x		x		x
Soft Computing	840	820	1651	x		x		x		
Наукометрия	1243	791	1990	x		x				x
Экспертные оценки	802	670	1463	x	x	x				
Стат. анализ	974	757	1714	x	x					x
Контроллеры	849	783	1612	x	x				x	
Многоагентные системы	886	766	1639	x	x					x
Роботы	1043	962	1983		x	x			x	
Базы данных	925	727	1619		x	x	x			x
Автоматическое управление	1027	822	1883		x	x	x			
САПР	946	840	1733				x	x	x	x
Информационная безопасность	922	766	1656				x	x	x	
Программирование	949	831	1720				x	x	x	x
Стандартизация и метрология	1176	1027	2174				x	x	x	
Операционные системы	1112	953	1976				x	x	x	

Объем обучающих выборок: 383 документов, экзаменационных: 84 документа

Расширенная матрица «документ-термин»

Русские термины *Английские термины*

$$X = \left[\begin{array}{ccc:ccc} x_1^{(1)} & \boxtimes & x_1^{(p)} & x_1^{(p+1)} & \boxtimes & x_1^{(M)} \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ x_N^{(1)} & \boxtimes & x_N^{(p)} & x_N^{(p+1)} & \boxtimes & x_N^{(M)} \end{array}\right] \quad p \neq M/2$$

где $x_j^{(i)}$ – вес термина *i* в документе *j* (*i*=1,…,*M*; *j*=1,…,*N*);
M – общее количество терминов в смешанной выборке;
N – количество документов.

Методы взвешивания, меры близости и методы классификации

Методы взвешивания:

Взвешивание частотой слова (term frequencies)

$$x_j^{(i)} = f_{ij}$$

Tf–idf - взвешивание (term frequencies – inverse document frequencies)

$$x_j^{(i)} = f_{ij} \log\left(\frac{N}{N_i}\right)$$

Меры близости:

Евклидова мера близости

$$d(\vec{X}_j, \vec{X}_l) = \sqrt{\sum_{i=1}^{M} (x_j^{(i)} - x_l^{(i)})^2}$$

Косинусоидальная мера близости

$$d(\vec{X}_j, \vec{X}_l) = \cos(\vec{X}_j, \vec{X}_l) = \frac{\left|\sum_{i=1}^{M} x_j^{(i)} x_l^{(i)}\right|}{\sqrt{\sum_{i=1}^{M} (x_j^{(i)})^2 \sum_{i=1}^{M} (x_j^{(i)})^2}}$$

Классические методы классификации:

Метод k-ближайших соседей

$$d\left(\vec{X}_j^*, \vec{X}_{N+1}\right) = \min d\left(\vec{X}_j, \vec{X}_{N+1}\right),$$

для $j = 1, \ldots, N.$

Метод центроидов

$$\vec{C}_k = \frac{1}{N_k} \sum_{j=1}^{N_k} \vec{X}_j$$

Профильные методы классификации

Профиль – формальный объект, который способен характеризовать все остальные элементы класса и состоит из наиболее информативных слов, определенных специальным образом.

	Принадлежность классу Q_g	Непринадлежность классу Q_g
Наличие признака $x^{(i)}$	A	B
Отсутствие признака $x^{(i)}$	C	D

PO-профиль:

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

Профиль Соукала-Сниса (C-C):

$$SS = \frac{2(A+D)}{2(A+D)+B+C}$$

Нормированный МИ-профиль:

$$I = \frac{A \log \frac{AN}{(A+B)(A+C)}}{(A+B)\log \frac{N}{A+B}}$$

Ошибки классификации

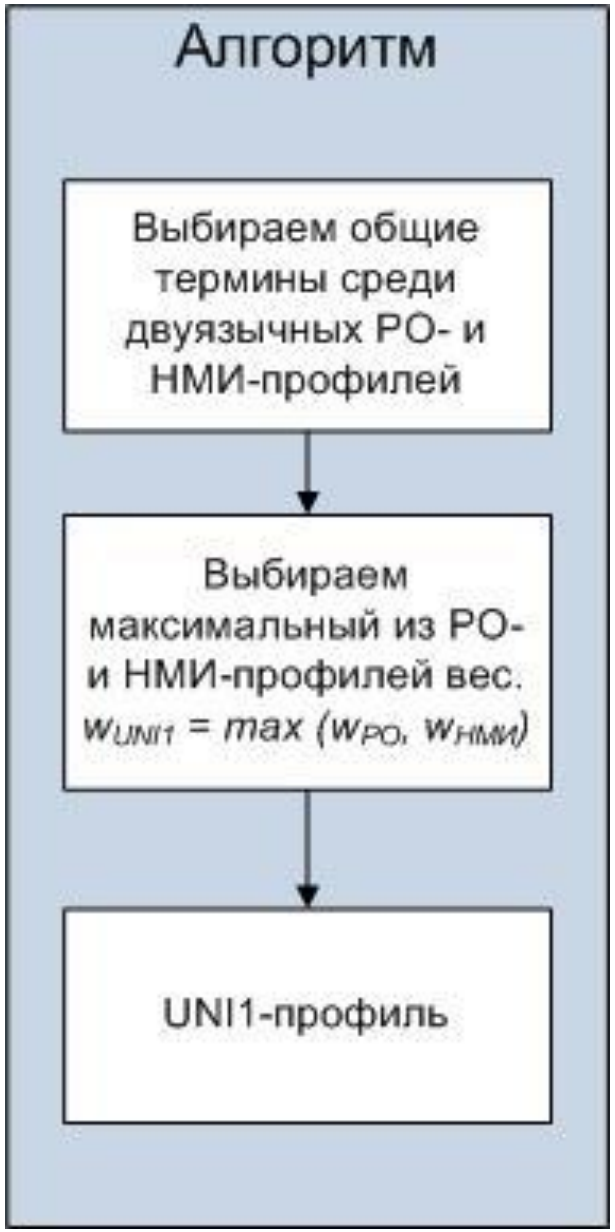
Профильные методы обладают более высокой точностью классификации, чем «классические» к-БС и метод центроидов.

Синтезированные профили. UNI1

Предположение:

Построение смешанного профиля, в который включались бы самые информативные термины обоих языков, рассчитанные по формулам РО- и НМИ-профилей.

Сюда должны попасть частотные слова РО-профиля и достаточно редкие (специфические) термины из НМИ-профиля

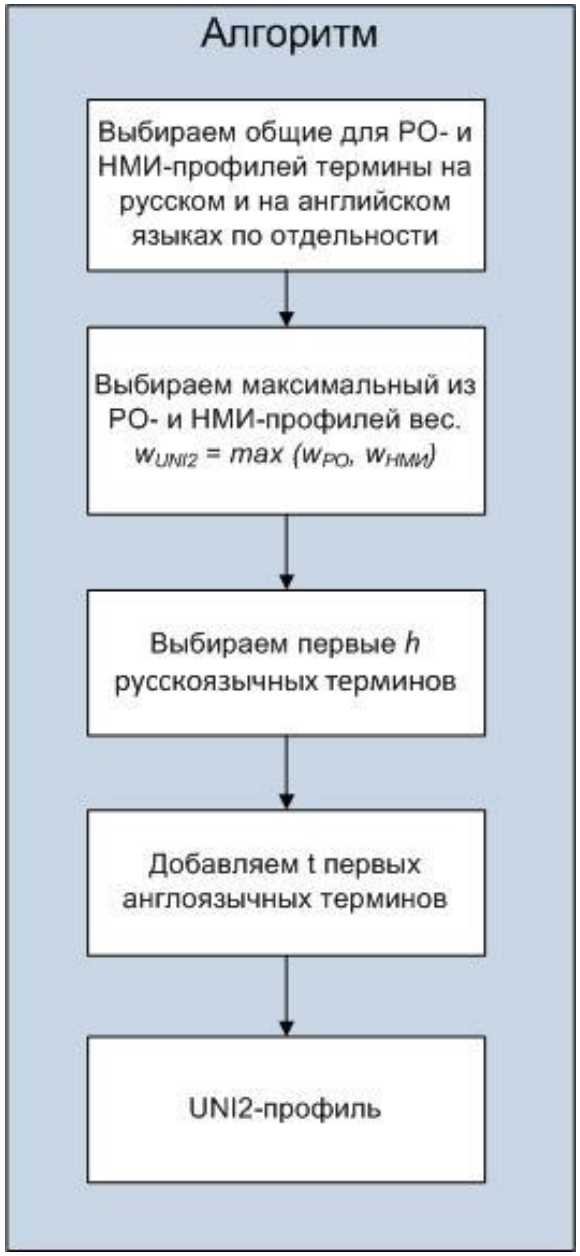


Синтезированные профили. UNI2

Предположение:

Русскоязычные и англоязычные тексты неравнозначны. Поскольку русский язык является «родным» для авторов, изложение на нем материала более квалифицированное и информативное.

В профиль включается h классообразующих русских терминов из РО- и НМИ-профилей, дополненных t наиболее информативными английскими словами.

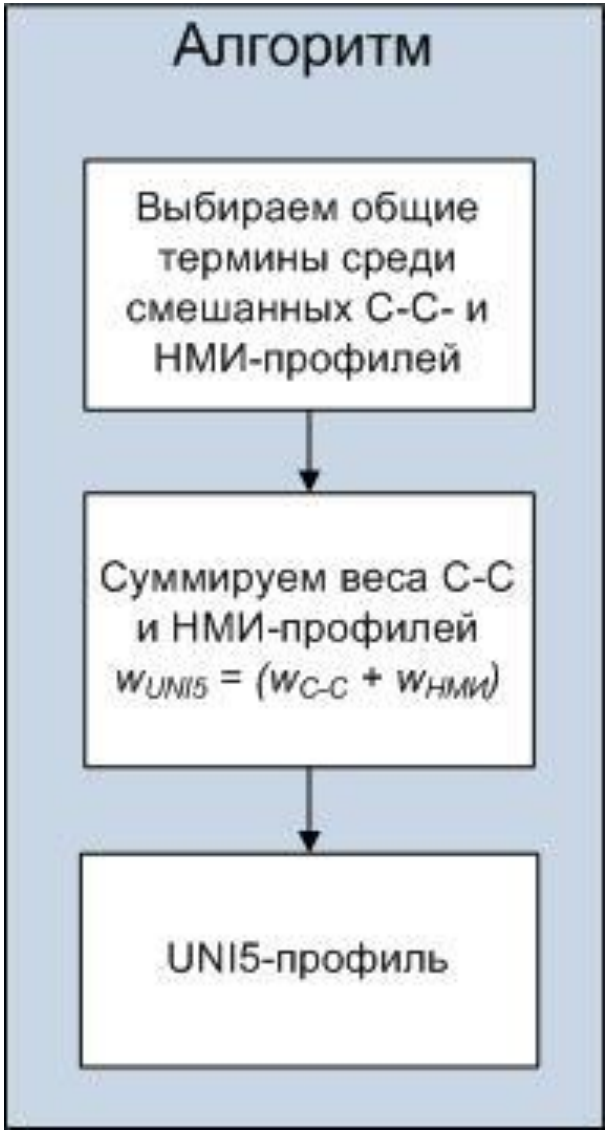


Синтезированные профили. UNI5

Предположение:

Элементы профиля рассчитываются как сумма весов НМИ- и С-С-профилей.

За счет высоких значений С-С-профиля, результирующие веса информативных терминов существенно возрастают (становятся больше 1) и усиливается их влияние на определение класса нового документа.



Результаты экспериментов для профильных методов

Результаты экспериментов, профильные методы:

Классификатор	Средняя ошибка классификации, %
PO	14,97
НМИ	11,21
C-C	13,09
UNI1	13,79
UNI2	12,41
UNI5	12,59

Получили группу приблизительно равноточных методов, основанных на разных подходах к выявлению информативных терминов, способная обучаться на английских, русских и смешанных выборках

Коллективы решающих правил

При объединении в коллектив можно ожидать, что разнородные процедуры будут «исправлять» ошибки друг друга и увеличивать результирующую точность.

- *КРП1* (*РО, НМИ, С-С*) - три наиболее разнородных классификатора: статистический РО-профиль, теоретико-информационный нормированный МИ-профиль и эвристический С-С-профиль.
- *КРП2* (*РО, НМИ, С-С, UNI2, UNI5*) - представляет собой *КРП1*, расширенный за счет включения *UNI2-* и *UNI5-* профилей.
- *КРП3* (*РО, НМИ, С-С, метод центроидов, к-БС*) – представляет собой *КРП1*, расширенный «классическими» методами: методом центроидов и методом к-ближайших соседей.

Результаты экспериментов, КРП:

Классификатор	Средняя ошибка классификации, %
НМИ-профиль	11,21
КРП1(РО, НМИ, С-С)	11,04
КРП2(РО, НМИ, С-С, UNI2, UNI5)	9,84
КРП3(РО, НМИ, С-С, метод центроидов,	10,36

Выводы

- Использование смешанных выборок, которые содержат терминологическую информацию на русском и английском языках, в большинстве случаев обеспечивает более высокую точность классификации по сравнению с одноязычными выборками.
- На основе экспериментальных результатов можно сделать вывод о хороших точностных характеристиках профильных методов. Эти методы, за счет более эффективного выявления информативных терминов позволяют улучшить точность классификации на смешанных выборках по сравнению с известными «классическими» методами.
- Приблизительная равноточность всех профильных методов при их разнородности позволяют объединять эти процедуры в *КРП,* обладающие наиболее высокой точностью классификации двуязычных документов.

Спасибо за внимание