Обзор методов классификации. Профильные методы классификации

Курс «Основы анализа текстовых данных»

Кафедра управления и интеллектуальных технологий

НИУ «МЭИ»

Весна 2023 г.

Логистическая регрессия

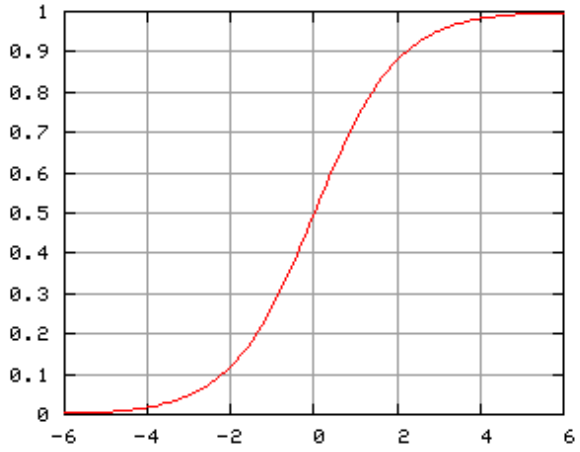
Логистическая регрессия (Logistic regression) — метод построения линейного классификатора, позволяющий оценивать апостериорные вероятности принадлежности объектов классам.

Уравнение линейной регрессии: $y(x,w) = w_0 + x_1 w_1 + x_2 w_2 + \ldots + x_n w_n$ $\qquad \hat{w} = \arg\min(Q(x,w))$

Мы решаем задачу бинарной классификации и хотим оценить вероятность принадлежности к классу “+” и “-”: $p \in [0;1]$
Проблема в том, что правая часть уравнения $\in [-\infty;+\infty]$

Выход из ситуации – использование логистической функции:

$$f(z) = \frac{1}{1+e^{-z}} \Rightarrow P_+ = \frac{1}{1+e^{-w^T x}}$$



Задача обучения линейного классификатора заключается в том, чтобы по обучающей выборке настроить вектор весов *w*. В логистической регрессии для этого решается задача минимизации эмпирического риска с функцией потерь специального вида:

$$\hat{w} = \operatorname{argmin}\left(\sum \ln(1 + e^{-y_i w^T x}) + \lambda|w|\right)$$

Margin = M => exp(-M)

Метод опорных векторов (SVM, Support Vector Machine)

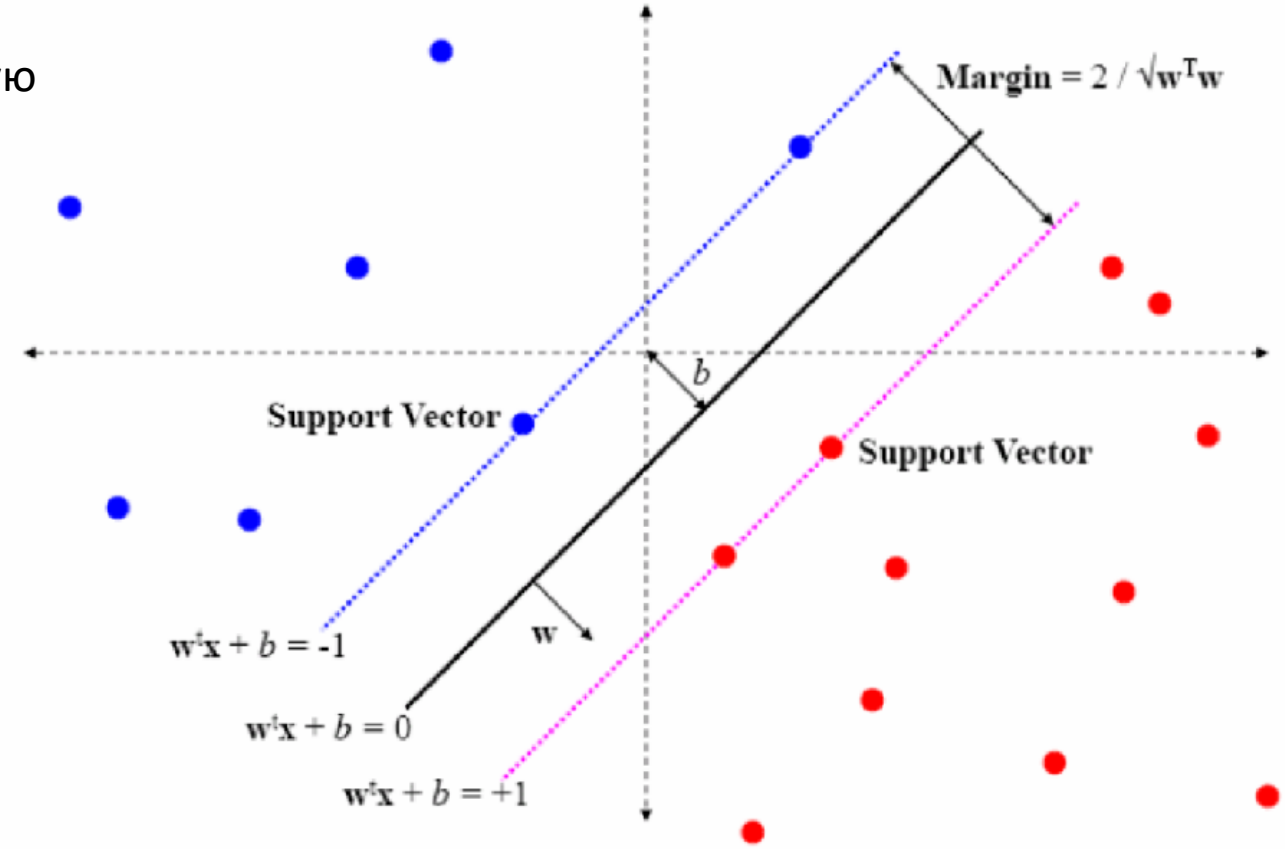
Алгоритм предложен в 1963 году Владимиром Вапником и Алексеем Червоненкисом. Основная идея метода — перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы. Разделяющей гиперплоскостью будет гиперплоскость, максимизирующая расстояние до двух параллельных гиперплоскостей.

Метод опорных векторов строит классифицирующую функцию $\hat{a}(x)$ в виде:

$$\hat{a}(x) = sign(w^T x)$$

Далее выбираются такие w$_i$

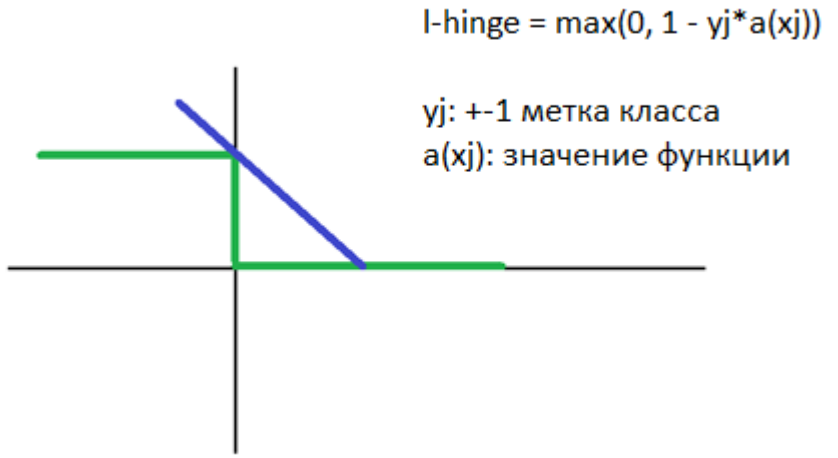
максимизируют расстояние $\frac{1}{\|w\|}$ до каждого класса



Метод опорных векторов (SVM, Support Vector Machine)

$$\widehat{w} = \sum_{j=1}^{N} l_{hinge}\left(y_j, \hat{a}(x_j)\right) + \lambda|w|$$

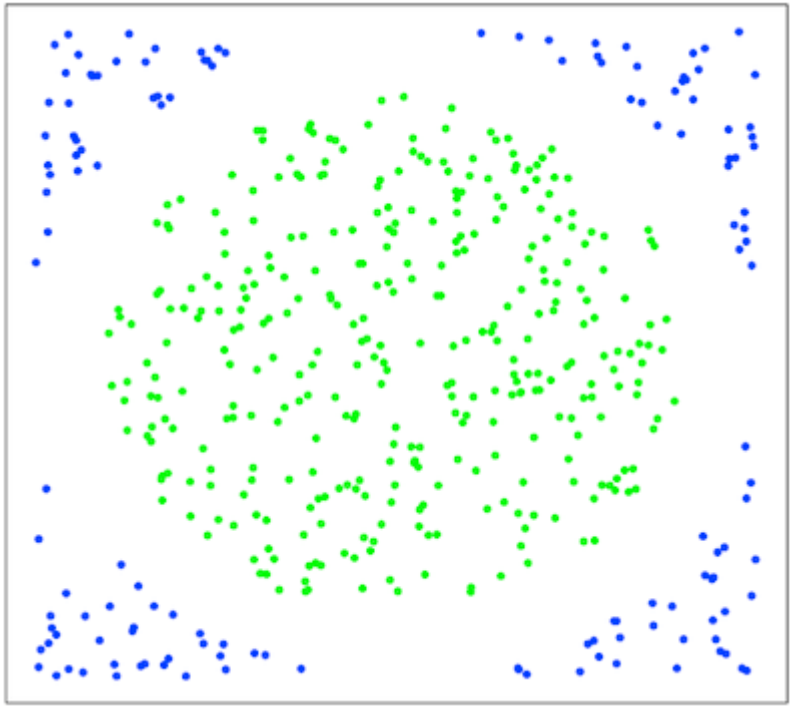
Здесь $y_j \in \{-1; 1\}$ - метка класса, принимающая одно из двух возможных значений.
Первое слагаемое соответствует ошибки классификатора на обучающей выборке, которая измеряется с помощью функции потерь.

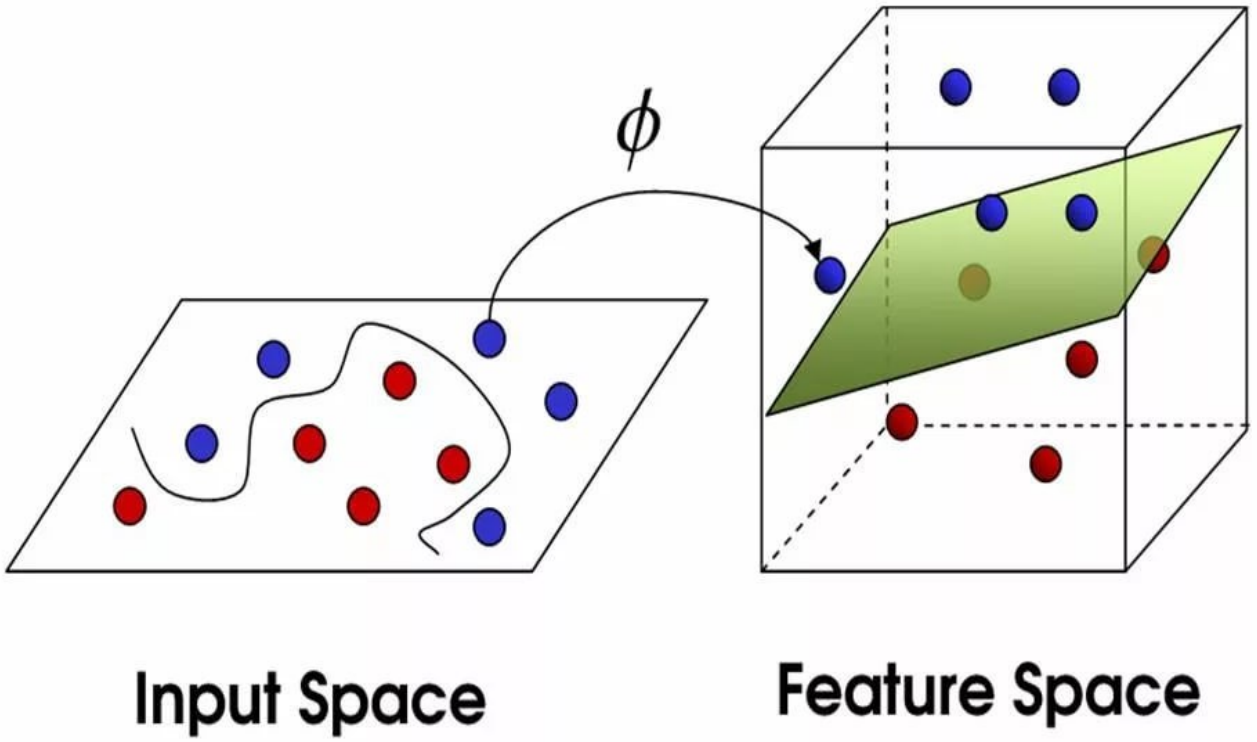


Второе слагаемое – регуляризатор (см. далее)

Метод опорных векторов. Линейная неразделимость

Если данные линейно неразделимы, то все элементы обучающей выборки вкладываются в пространство **X** более высокой размерности с помощью специального отображения[1] $\phi : R^n \to X$





[1]**К. В. Воронцов.** *Лекции по методу опорных векторов.* http://www.ccas.ru/voron/download/SVM.pdf

Регуляризация

Регуляризация — метод добавления некоторой дополнительной информации к условию с целью решить некорректно поставленную задачу или предотвратить переобучение. Эта информация часто имеет вид штрафа за сложность модели.
Переобучение в большинстве случаев проявляется в том, что в получающихся многочленах слишком большие коэффициенты. Соответственно, и бороться с этим можно довольно естественным способом: нужно просто добавить в целевую функцию штраф, который бы наказывал модель за слишком большие коэффициенты.

$$a(x, w) = w_0 + x_1 w_1 + x_2 w_2 + \ldots + x_n w_n$$

В большинстве случаев коэффициенты должны быть небольшими числами, идеально — многие малозначимые коэффициенты должны быть нулями

Хорошо:

$$a(x, w) = 0.5 + 0.69 x_1 + 0.38 x_2 + 0.84 x_3$$

Подозрительно:

$$a(x, w) = 0.5 + 365723 x_1 + 0.38 x_2 + 0.00000084 x_3$$

Регуляризация (2)

L1 – регуляризация (лассо, lasso)

$$\hat{w} = \arg\min\left(\sum \ln(1+\exp(-y_i < x_i, w >)) + \lambda \mid w \mid\ \right)$$

L2 – регуляризация (гребневая, ridge)

$$\hat{w} = \arg\min\left(\sum \ln(1+\exp(-y_i < x_i, w >)) + \lambda \mid w \mid^2\right)$$

$C = \frac{1}{\lambda}$ - Обратный коэффициент регуляризации.

- Чем С больше, тем более сложные зависимости в данных может восстанавливать модель.
- Если регуляризация слишком сильная (малые значения С), то решением задачи минимизации логистической функции потерь может оказаться тот случай, когда многие веса занулились или стали слишком малыми
- Если С слишком большая, то модель может переобучиться
- При равной точности следует выбирать более простую модель.

Профильные методы классификации

Профиль класса – это формальный объект, способный охарактеризовать все остальные элементы класса.

Например – центроид

Профили классов могут быть разделены на следующие категории:

1) Логический профиль, который состоит из признаков, представленных в классе: (вес признака в таком профиле равен *“0”* или *“1”*).

2) *Рассчитываемый профиль*, в этом случае вес признаков рассчитывается на основе какого-либо правила. Примером может служить центроид класса

3) *Экспертный профиль*, задаваемый пользователем на основе собственных знаний и опыта.

Что такое профиль класса?

Взвешивание и отбор признаков может проводиться на основе известных процедур выявления информативных терминов:

- Частотный,
- Статистический,
- Теоретико-информационный,
- Эвристический,
- …

Q_k / X	Принадлежность классу Q_k	Непринадлежность классу Q_k	$\sum$
Наличие признака $x^{(i)}$	A	B	$A+B$
Отсутствие признака $x^{(i)}$	C	D	$C+D$
$\sum$	$A+C$	$B+D$	N

Статистический подход выявления информативных терминов

Хи-квадрат - профиль:

$$\chi^2(x^{(i)}, Q_g) = N \cdot \frac{(AD - BC)^2}{(A+B)(C+D)(A+C)(B+D)}$$

РО-профиль:

$$\rho(x^{(i)}, Q_k) = \frac{(AD - CB)}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

Профиль Юла:

$$Q(x^{(i)}, Q_g) = \frac{AD - BC}{AD + BC}$$

Теоретико-информационный подход выявления информативных терминов

МИ-профиль (Mutual information):

$$MI(x^{(i)}, Q_g) = \log_2 \frac{A \cdot N}{(A+B) \cdot (A+C)}$$

Нормированный МИ-профиль:

$$NMI(x^{(i)}, Q_k) = \frac{A \log_2 \frac{AN}{(A+B)(A+C)}}{(A+B) \log_2 \frac{N}{A+B}}$$

Эвристический подход выявления информативных терминов

Простой коэффициент совстречаемости

$$S = \frac{A + D}{A + B + C + D}$$

Первый коэффициент несогласия

$$SN1 = \frac{C + B}{A + B + C + D}$$

Коэффициент Рассела-Рао

$$RR = \frac{A}{A + B + C + D}$$

Коэффициент Роджерса-Танимото

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

Первый коэффициент Сокала-Сниса

$$SS2 = \frac{2(A + D)}{2(A + D) + B + C}$$

Коэффициент Хаммана

$$H = \frac{(A + D) - (B + C)}{A + B + C + D}$$

Эвристический подход выявления информативных терминов (2)

Коэффициент Джаккарда (Жаккара)

$$J = \frac{A}{A + B + C}$$

Второй коэффициент несогласия

$$SN2 = \frac{C + B}{A + B + C}$$

Коэффициент Dice

$$Dice = \frac{2A}{2A + B + C}$$

Второй коэффициент Сокала-Сниса

$$SS2 = \frac{A}{A + 2(B + C)}$$

Первый коэффициент Кульчинского

$$K1 = \frac{A}{B + C}$$

Профильные методы

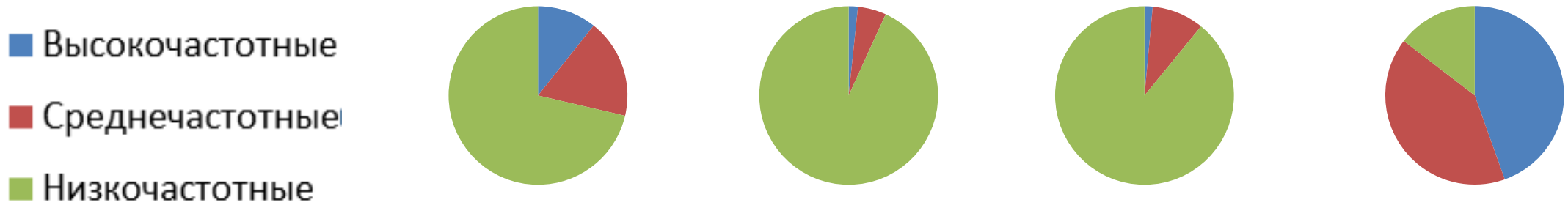
$$W_k = \sum_{i=1}^{Mk} tf_i \cdot \Pr of(x^{(i)}, Q_k) \qquad W_k = \max \text{ (для } \forall k, k = 1,\ldots, K)$$

Профили класса «Базы данных»

	РО-профиль		НМИ-профиль		J-профиль		С-С-профиль	
	databas	0,750	субд	0,689	databas	0,644	запрос	0,952
	субд	0,615	databas	0,652	баз	0,459	databas	0,949
	запрос	0,588	sql	0,623	запрос	0,420	субд	0,949
	баз	0,579	dbm	0,613	субд	0,415	реляцион	0,944
	sql	0,525	запрос	0,608	queri	0,352	queri	0,942
	реляцион	0,522	реляцион	0,599	реляцион	0,318	sql	0,939
	queri	0,512	sql-запрос	0,523	sql	0,308	dbm	0,937
	dbm	0,511	queri	0,520	модел	0,295	sql-запрос	0,933
	structur	0,409	ms	0,510	dbm	0,292	ms	0,932
	relat	0,385	mysql	0,466	relat	0,269	mysql	0,929
	sql-запрос	0,385	таблиц	0,466	структур	0,255	oracl	0,929
	ms	0,367	oracl	0,450	structur	0,234	postgresql	0,928
	структур	0,314	postgresql	0,450	data	0,184	relat	0,928
	mysql	0,306	баз	0,443	sql-запрос	0,169	sql-queri	0,928
	таблиц	0,306	sql-queri	0,431	model	0,161	таблиц	0,928
	oracl	0,283	inject	0,411	manag	0,160	end-us	0,927
	postgresql	0,283	pargresql	0,411	ms	0,154	laplas-stilt	0,927
	tabl	0,279	клиентск	0,411	выполнен	0,153	lst	0,927
	выполнен	0,271	tabl	0,392	access	0,153	secur	0,927
	access	0,271	db	0,388	base	0,153	бд	0,927

Сравнение профильных методов классификации

	PO	НМИ	C-C	J
Какие термины выделяет	Высокочастотные, среднечастотные	Среднечастотные, низкочастотные	Среднечастотные	Высокочастотные
Интервал значений весов	[-1;1]	[0;1]	[0;1]	[0;1]
Коэффициент убывания весов терминов λ	Средний 4.8	Низкий 2.2	Очень низкий 1.05	Высокий 7.4
Используемые значения из таблицы сопряженности	A, B, C, D	A, B, C	A, B, C, D	A, B, C



UNI-профили

Разнообразие профильных методов позволяет использовать их для увеличения точности классификации:

- Комбинирование профилей для создания более сильных классификаторов
- Использование знаний о структуре документа для увеличения точность классификации

«Union» - объединение разных подходов к выявлению информативных терминов – статистического, теоретико-информационного, эвристического.

Для их построения используются различные комбинации PO-, НМИ-, J- и C-C-профилей, которые, за счет различных принципов определения наиболее информативных понятий, позволят скомпенсировать слабые стороны каждого из подходов.

Алгоритм построения профиля UNI6

- *Входными данными алгоритма* являются: обучающая выборка документов, PO- , НМИ- и J-профили
- *Выходные данные:* профили классов, представленные в виде вектора терминов и упорядоченные по убыванию веса.
- *Шаг 1*. Задается параметр метода - L.
- *Шаг 2*. Суммируем веса профилей для каждого общего термина и исходных профилей:

  $w_{uni6} = (w_{PO} + w_{НМИ} + w_J)$, , здесь w_{uni6} – вес термина в UNI6-профиле, w_{PO}, $w_{НМИ,}$ w_J– вес термина в PO-, НМИ-, J- профиле соответственно.
- *Шаг 3*. Полученные термины упорядочиваются по убыванию веса.

Как использовать структуру документа?

- Слова в разных частях документа (названия, аннотации, ключевые слова) неравнозначны
- Как учесть эту неравнозначность при классификации?

Семантическая интерпретация в системах компьютерного анализа текста

Описывается подход к построению семантического компонента в системах компьютерного анализа текста на естественном языке. Подход основан на применении специальных шаблонов к сети синтактико-семантических отношений между словами текста, которая строится синтаксическим анализатором. Шаблоны определяют способ интерпретации фрагментов сети в заданные фреймы с идентификацией участников ситуаций и их ролей.
Ключевые слова: компьютерный анализ текста, семантическая интерпретация, семантическая сеть, синтаксический анализ, фреймы.

Как использовать структуру документа? (2)

Подход №1. Использовать разные способы выявления информативных терминов для разных частей документа

$$W_k = \sum_{i=1}^{Mk} tf_T^{(i)} \Pr of_T(x^{(i)}, Q_k) + tf_A^{(i)} \Pr of_A(x^{(i)}, Q_k) + tf_K^{(i)} \Pr of_K(x^{(i)}, Q_k)$$

Подход №2. Использовать специальные веса для терминов из разных частей документа

$$W_k = \sum_{i=1}^{Mk} \alpha * tf_T^{(i)} \Pr of\ (x^{(i)}, Q_k) + \beta * tf_A^{(i)} \Pr of\ (x^{(i)}, Q_k) + \gamma * tf_K^{(i)} \Pr of\ (x^{(i)}, Q_k)$$

Подход №1

Использовать разные способы выявления информативных терминов для разных частей документа

$$W_k = \sum_{i=1}^{Mk} tf_T^{(i)} \operatorname{Prof}_T(x^{(i)}, Q_k) + tf_A^{(i)} \operatorname{Prof}_A(x^{(i)}, Q_k) + tf_K^{(i)} \operatorname{Prof}_K(x^{(i)}, Q_k)$$

Основная проблема – выбор методов, которыми будут оцениваться термины из разных частей документов.

Решение:

1. Анализ встречаемости низко-, средне- и высокочастотных терминов в различных разделах БО с целью выбора таких методов, которые предпочитают отбирать в профиль и присваивать более высокие веса той категории терминов, которая наиболее часто появляется в разделе.
2. Полный перебор всех возможных вариантов путем комбинирования разных профилей.

Подход №2

Использовать специальные веса для терминов из разных частей документа

$$W_k = \sum_{i=1}^{Mk} \alpha * tf_T^{(i)} \operatorname{Prof}(x^{(i)}, Q_k) + \beta * tf_A^{(i)} \operatorname{Prof}(x^{(i)}, Q_k) + \gamma * tf_K^{(i)} \operatorname{Prof}(x^{(i)}, Q_k)$$

Веса α, β, γ будем настраивать по методу Фишберна*.
В качестве $\operatorname{Prof}(x^{(i)}, Q_k)$ будем использовать наиболее точный профиль.

*веса Фишберна - это рациональные дроби, в знаменателе которых стоит сумма арифметической прогрессии N первых членов натурального ряда с шагом 1, а в числителе - убывающие на 1 элементы натурального ряда от N до 1 (например, 3/6, 2/6, 1/6 в сумме дают единицу)

Сравнение методов на разных типах выборок