АВТОМАТИЧЕСКАЯ ФИЛЬТРАЦИЯ РУССКОЯЗЫЧНОГО НАУЧНОГО КОНТЕНТА МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ И ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

Воронов С. О. (rdkl.hrd@gmail.com)[2],
Воронцов К. В. (voron@forecsys.ru)[1, 2]

[1]Яндекс, Москва, Россия
[2]Московский Физико-Технический Институт
(государственный университет), Москва, Россия

Предлагается метод жанровой классификации текстов, основанный на статистическом обучении по частично классифицированной выборке документов. Для автоматизации формирования достаточной обучающей выборки используется метод активного обучения. Возникающая при этом смещённость оценок качества классификации компенсируется поправкой Хансена-Гурвица. Для извлечения словарных признаков используются методы тематического моделирования. Для извлечения признаков библиографии строится классификатор строк документа. Эксперименты проводились на выборке общедоступных документов с сайтов российских вузов. Оценивалась важность отдельных признаков для выделения научных документов. Анализ словарных признаков показал, что тематическое моделирование позволяет находить группы слов, различающие научный и образовательный контент.

Ключевые слова: жанровая классификация текстов, извлечение признаков, машинное обучение, вероятностная тематическая модель, регуляризация

AUTOMATIC FILTERING OF RUSSIAN SCIENTIFIC CONTENT USING MACHINE LEARNING AND TOPIC MODELING

Voronov S. O. (rdkl.hrd@gmail.com)[2],
Vorontsov K. V. (voron@forecsys.ru)[1, 2]

[1]Yandex, Moscow, Russia
[2]Moscow Institute of Physics and Technology, Russia

We propose a genre classification technology based on machine learning and topic modeling for the automatic filtering of scientific documents crawled from the Web. We use the uncertainty sampling technique, which is usually used in active learning, in order to accumulate a small but representative training sample. The sampling of new unlabeled data is based on a distribution of margins estimated from the SVM linear classifier. We use Hansen-Hurwitz estimator to compensate for the bias in precision, recall and F1 measure that occurs because of the non-uniform sampling. The feature extraction is based on the Greek and math symbols counters, the bibliography records counters, and the class-specific words counters. We use the Additively Regularized Topic Model to combine the Dependency-LDA topic model for document classification with topic sparsing and decorrelating. Experiments were carried out on a sample of 850K public documents crawled from the sites of Russian universities, which contains no more than 2% of purely scientific content. The analysis of class-specific topics showed that our topic model automatically discovers groups of words that distinguish scientific and educational content.

Key words: genre classification, feature extraction, machine learning, probabilistic topic modeling, regularization

1. Введение

Создание систем агрегирования научной информации требует автоматического пополнения текстовой коллекции общедоступными документами из сети Интернет. Для фильтрации потока документов необходимо определять жанр каждого документа, в частности, отделить научные документы от ненаучных. Согласно ряду исследований, именно жанр научных документов выделяется с наибольшей точностью [Stamatatos 2000; Meyer zu Eißen & Stein 2004; Kanaris & Stamatatos 2009]. Однако вопрос об отнесении документа к жанру научного часто неоднозначен даже для экспертов. На практике само понятие *жанра* может конкретизироваться в контексте целей создаваемого информационного ресурса и предполагаемой целевой аудитории. Например, нас может интересовать «научная информация по фармакологии» или «научная информация за исключением коммерческого и образовательного контента» или «первичные источники научно-технической информации». Автоматизация процессов создания информационно-поисковых ресурсов для профессиональных сообществ требует универсальной и гибкой технологии фильтрации контента, легко настраиваемой на любой жанр и любую тематическую направленность. Целью данной работы является разработка такой технологии.

Для настройки фильтра на определённый жанр и тематику предлагается размечать относительно небольшую часть коллекции на релевантные и нерелевантные документы. Таким образом, понятие «жанра» закладывается на этапе формирования обучающей выборки асессорами путём явного указания положительных и отрицательных примеров. Затем используются методы тематического моделирования с частичным обучением (semi-supervised learning),

позволяющие объединить размеченную выборку относительно небольшого объёма и неразмеченную выборку огромного объёма для выявления жанрово-тематической кластерной структуры коллекции.

Проблема в том, что доля релевантных документов может составлять менее 1% всей коллекции. Случайный отбор документов для асессорской разметки в этом случае не эффективен. Кроме того, несбалансированность обучающей выборки может затруднять применение методов машинного обучения. Поэтому отбор документов производится не случайным образом, а так, чтобы «пограничные» документы с наименее уверенной классификацией чаще попадали в обучающую выборку. Для этого применяются методы активного обучения (active learning), в которых обучающие объекты отбираются самим алгоритмом обучения [Lewis & Gale 1994]. При этом сокращается трудоёмкость разметки, но возникает новая проблема. Вероятностное распределение объектов обучающей выборки становится смещённым, из-за этого оценки качества классификации искажаются. Для получения несмещённых оценок предлагается сэмплировать обучающие объекты из фиксированного вероятностного распределения, существенно повышающего вероятность выбора пограничных объектов, и использовать формулу Хансена-Гурвица [Hansen et al. 1953] для коррекции оценок качества классификации.

Ещё одна проблема заключается в том, что на начальном этапе разработки нет ни обучающей выборки, ни обоснованной системы признаков. Поэтому предлагается циклический процесс разработки. На каждом цикле текущий алгоритм классификации используется для пополнения обучающей выборки, по которой затем обучается алгоритм классификации следующего цикла. Одновременно производится экспертный анализ ошибок классификации и оценивается необходимость разработки новых признаков.

Для выделения групп слов, различающих релевантные и нерелевантные документы, предлагается использовать тематическое моделирование. Широко известные методы вероятностного латентного семантического анализа PLSA [Hofmann 1999] и латентного размещения Дирихле LDA [Blei et al. 2003] слишком универсальны и обычно выделяют темы, соответствующие предметным областям. В данном случае цель тематического моделирования иная — автоматическое выявление характерных терминов того жанра, который был задан размеченной выборкой документов. Поэтому используется регуляризованная тематическая модель, ориентированная на выделение тем, наиболее полезных для классификации документов на релевантные и нерелевантные [Rubin 2012, Vorontsov & Potapenko 2014].

Для экспериментов мы использовали выборку из 850 тысяч общедоступных документов с 2000 сайтов российских вузов. Доля научных документов по предварительным оценкам не превышала 2%. Асессорами было размечено 3000 документов, из них около 40% научных. При разметке асессоры придерживались основного правила, что научные документы прежде всего должны нести знание. К ним относились научные статьи, книги, диссертации и авторефераты, научные отчёты, учебники для вузов, сборники трудов конференций, и т.д. Не считались научными слишком короткие документы (в частности, аннотации

статей), документы рекламного и коммерческого содержания (страницы факультетов, кафедр и преподавателей, краткие описания промышленных разработок), образовательные материалы (программы курсов, учебные задания), списки публикаций, материалы эзотерического и религиозного содержания. В некоторых спорных случаях документ считался научным по формальным признакам, например при наличии фразы «*научное издание*» в метаописании документа. Отметим, что цель данной работы не состояла в том, чтобы предложить критерий научности, а лишь в том, чтобы продемонстрировать применимость технологии, позволяющей задать принцип жанрово-тематической классификации документов путём разметки относительно небольшого числа документов.

2. Формирование признаков и обучающей выборки

Для классификации документов часто используются линейные классификаторы: метод опорных векторов SVM [Joachims 1998, Manevitz & Yousef 2002], метод релевантных векторов RVM [Rafi & Shaikh 2013], логистическая регрессия. Линейный классификатор вычисляет дискриминантную функцию как взвешенную сумму признаков и сравнивает её значение с нулём. Значения выше нуля означают, что алгоритм классифицировал данный документ как релевантный. В качестве числовых признаков документа используются частоты специфичных слов или словосочетаний и различная метаиформация [Stamatatos 2000; Meyer zu Eißen & Stein 2004; Kanaris & Stamatatos 2009].

В данной работе мы использовали следующие числовые признаки (после предварительного отбора различных вариантов).

GL: количество греческих букв в документе. Они часто используются в научных текстах в области точных и естественных наук.

MS: количество математических символов в документе (Δ, ∇, $\subset$, $\subset$, $\leq$, $\geq$, $=$, $\times$, и др.).

BS: логарифм длины документа (в символах), либо 0, если длина текста меньше 15 000 символов. Порог был подобран экспериментальным путём.

DC: количество цифр в документе.

SS: индикатор короткого текста, равный 1, если длина текста меньше 4000 символов.

PG: количество фраз, состоящих из слов, часто употребляемых при упоминании научных грантов, при приблизительном соблюдении порядка слов (*работа, исследование, выполнить, поддержать, финансовая, поддержка, грант, РФФИ, РНГФ, РФТР, РНФ, ФФЛИ, ФПОН, российский, фонда, фундаментальный, исследование*).

Для разметки обучающей выборки была разработана программа, показывающая список заголовков документов, ранжированный по убыванию значений дискриминантной функции линейного классификатора. Асессор мог выбрать любой документ, просмотреть его и поставить метку релевантности, тем самым пополнив обучающую выборку.

На первом цикле, когда обучающей выборки вообще не было, было построено линейное правило классификации с тремя признаками GL, MS, BS и фиксированными коэффициентами. Это уже позволило выделить в ранжированном списке достаточное число научных и ненаучных документов. Первая разметка нескольких сотен документов заняла не более двух часов времени.

На втором цикле по полученной обучающей выборке был построен SVM на шести признаках. На рис. 1 показан график распределения документов по значениям дискриминантной функции.

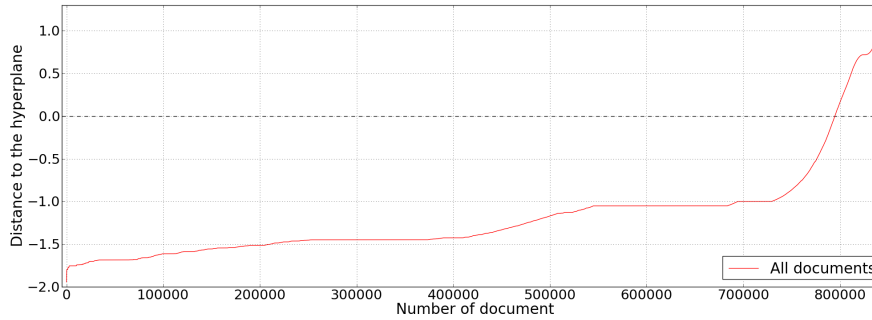

Рис. 1. Распределение документов по расстоянию до разделяющей гиперплоскости для SVM с набором из 6 признаков (GL, MS, BS, DC, SS, PG)

По графику можно заметить, что значения дискриминантной функции сильнее всего различаются в окрестности нуля, то есть в пограничной зоне наименее уверенной классификации. На этом наблюдении основан следующий способ формирования обучающей выборки. Для сэмплирования очередного обучающего объекта выбирается случайное число из равномерного распределения на интервале значений дискриминантной функции. Затем находится документ, соответствующий данному значению дискриминантной функции. Этот способ неравномерного сэмплирования обучающей выборки аналогичен *сэмплированию по неуверенности* (uncertainty sampling), применяемому в активном обучении [Lewis & Gale 1994].

После его применения к нашей выборке доля научных документов в ней выросла с 2% до 40% и была получена сбалансированная выборка, подходящая для построения линейных классификаторов.

Оценки качества классификации, полученные по неравномерно сэмплированной выборке, являются смещёнными. Однако в данном случае распределение $q(d)$, из которого производилось сэмплирование, известно. Поэтому несмещённая выборочная оценка среднего может быть легко получена для любой величины $f(d)$, зависящей от документа, с помощью формулы Хансена-Гурвица. Она равна среднему по выборке значению $f(d)p(d)/q(d)$, где $p(d)$ — равномерное распределение. Таким образом, проблема смещённости оценок качества классификации решается путём введения весовых коэффициентов $w(d) = p(d)/q(d)$ для документов обучающей выборки.

На последующих циклах анализ основных причин ошибочной классификации показал важность введения *словарных* и *библиографических* признаков документов.

Под словарным признаком мы понимаем частоту употребления слов из данной группы слов в данном документе. Слова, характерные для релевантных и нерелевантных документов, будем называть, соответственно, релевантными и нерелевантными.

На третьем цикле было введено пять словарных признаков на группах слов, отобранных вручную:

KW:Struct — слова, образующие структуру научной публикации (*введение, заключение, список литературы, постановка задачи, ISBN, УДК, аннотация, автореферат* и т. д.);

KW:Names — фамилии учёных;

KW:Frequent — слова общенаучной лексики (*анализ, задача, метод, модель, процесс* и т. д.);

KW:Terms — релевантные слова, термины предметных областей;

KW:StopWords — нерелевантные слова, более характерные для образовательного контента (*программа, курс, студент, задание, экзамен, кафедра, доцент* и т. д.).

Использование словарных признаков дало значимый прирост качества классификации на размеченной части коллекции. Однако очевидно, что это не является системным решением проблемы. Поэтому на следующих циклах словарные признаки строились автоматически методами тематического моделирования, и было показано, что они позволяют полностью отказаться от словарных признаков **KW:X**, сформированных вручную.

3. Классификация строк библиографии

Наличие списка литературы и ссылок на литературу является одним из наиболее характерных признаков научных документов. В качестве числового значения признака предлагается использовать максимальное число подряд идущих строк библиографии. Чтобы его вычислить, для каждой строки определяется вероятность того, что это часть библиографического описания. Затем производится сглаживание этих вероятностей по значениям вероятностей соседних строк. Признак **Bib** определяется как максимальное число подряд идущих строк с вероятностью, превышающей 50 %.

Для оценивания строк документа был построен классификатор, определяющий вероятность того, что строка является библиографической записью или её фрагментом. Фрагментирование происходит из-за того, что во многих документах (в частности, полученных из формата PDF) границы строк не совпадают с границами абзацев. Для формирования обучающей выборки было отобрано около 1000 различных строк, которые были размечены асессором на два класса: библиографические и обычные. Далее на этой выборке был обучен отдельный классификатор.

Использовались следующие признаки строк для распознавания библиографических записей или их фрагментов:

1) подходит ли строка под паттерн начала строки библиографической записи (как например [1], 2., [TrD90] или [SDD, 99]): 4 признака (по одному на каждый из паттернов);
2) длина строки;
3) число точек в строке;
4) средняя длина слова;
5) число слов, начинающихся с заглавной буквы;
6) число инициалов в строке (вида С. или К.);
7) число двойных инициалов (вида С.О. или К. В.);
8) число символов пунктуации в строке;
9) число слэшей в строке;
10) число символов, относящихся к странице (р., P., pp., стр., С., page)
11) число символов-разделителей (тире, кавычки и т. д.);
12) число чисел, похожих на год, в строке (1971, 2000, 2014 и т. д.);
13) число цифр в строке.

Наиболее информативными признаками оказались счётчики числа точек, слэшей, годов, слов с прописной буквы, инициалов, а также длина строки и средняя длина слов. При их удалении качество классификации наиболее значительно падает. В совокупности они дают 6,8% ошибки на контроле (без других признаков). Итоговая частота ошибок построенного классификатора составила 5,24%, доля ошибок на строках библиографии составила 10,28% (главным образом из-за фрагментированности), доля ошибок на небиблиографических строках составила 1,84%.

Второй числовой признак документов **Ref**, также связанный с библиографией, определяется как число найденных в документе ссылок на литературу. Для обнаружения ссылок вида [9], [StR02], (Vapnik et al. 1995) и т. д. использовались регулярные выражения.

4. Тематическое моделирование

Тематическая модель коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова образуют каждую тему. Вероятностная тематическая модель описывает каждую тему t дискретным распределением $p(w|t)$ на множестве слов w, каждый документ d — дискретным распределением $p(t|d)$ на множестве тем t. *Тематические модели классификации* учитывают также данные о принадлежности документов классам. На выходе тематической модели классификации оцениваются условные распределения $p(c|d)$ классов c для каждого документа d. Преимущества тематических моделей классификации проявляются на больших текстовых коллекциях с большим числом несбалансированных, пересекающихся, взаимозависимых классов, когда обычные методы классификации и категоризации текстов показывают неудовлетворительные результаты (Rubin 2012).

Большинство известных тематических моделей классификации являются развитием модели латентного размещения Дирихле LDA (Blei et al. 2003) и основаны

на теории графических моделей и байесовского обучения, в частности, MedLDA (Zhu et al. 2009), Labeled LDA (Ramage et al. 2009), Flat-LDA и Dependency LDA (Rubin 2012). В современной литературе область тематического моделирования является скорее полигоном для оттачивания сложных и чрезмерно общих математических методов байесовского вывода. Это создаёт значительные барьеры для исследователей, желающих применять тематические модели для решения прикладных задач анализа текстов. Несмотря на то, что в литературе описаны сотни специализированных моделей с весьма богатой функциональностью, исследователи часто ограничиваются простыми устаревшими моделями PLSA и LDA.

Новый математический аппарат аддитивной регуляризации тематических моделей, ARTM (Vorontsov & Potapenko 2014) упрощает этапы формализации и вывода моделей и позволяет комбинировать модели в любых сочетаниях. На основе ARTM разработана библиотека тематического моделирования с открытым кодом http://bigartm.org, поддерживающая параллельные вычисления, распределённое хранение больших коллекций текстов, комбинирование регуляризаторов и мультимодальность, в частности, возможность решения задач классификации и категоризации.

В данной работе аддитивная регуляризация используется для комбинирования модели классификации, аналогичной Dependency LDA, с регуляризаторами разреживания, сглаживания и декоррелирования, повышающими различность и улучшающими интерпретируемость тем (Vorontsov & Potapenko 2014). Применение данного набора регуляризаторов представляется целесообразным для автоматического построения словарных признаков, хорошо различающих релевантные и нерелевантные документы.

Построение словарных признаков производилось двумя способами.

Teta: в качестве словарных признаков документов принимались вероятности класса научных документов $p(c = +1|t)\ p(t|d)$ для всех 25 тем, найденных тематической моделью классификации.

8TM: производилось несколько запусков тематического моделирования с 25 темами, и из всех полученных тем с вероятностями $p(c|t)\ p(t|d)$, близкими к единице, экспертом отбирались восемь наиболее интерпретируемых.

Для предварительной обработки текстов использовался лемматизатор Lucene Russian Morphology.

5. Результаты экспериментов

В Таблице 1 показаны ранжированные списки наиболее значимых (верхних) слов в некоторых из найденных тем. Хорошо видно, что в качестве первой, «самой ненаучной», выделилась тема, связанная с образовательным контентом. «Самая научная» пятая тема является семантически неоднородной и содержит слова общенаучной лексики. «Менее научная» четвёртая тема собрала терминологию физико-технических наук, что свидетельствует о некоторой смещённости взятой для экспериментов коллекции. Таким образом, тематическая модель настроилась не столько на выделение тем предметных областей, сколько на различение релевантных и нерелевантных документов. Отметим,

что в процессе формирования обучающей выборки асессорам часто приходилось сталкиваться с «пограничными случаями», когда отнесение документа к образовательному или научному контенту вызывало споры.

Таблица 1: Верхние слова из пяти тем, полученных тематической моделью классификации. В первой строке указана вероятность того, что тема научная $p(c = +1 \mid t)$

0,000	0,099	0,203	0,755	1,000
образование	который	страна	прямая	процесс
студент	ряд	Россия	быть	результат
социальный	оценка	производство	точка	модель
учебный	человек	экономика	значение	среда
современный	параметр	который	множество	зависимость
университет	источник	год	движение	различный
российский	система	орган	состояние	структура
научный	помощь	государство	система	позволять
формирование	связь	проблема	время	являться
образовательный	этот	развитие	рис	поверхность
международный	комплекс	период	временить	расчет
организация	наличие	федеральный	тогда	технический
проект	изменение	закон	исследование	обработка
история	труд	хозяйство	свойство	качество
информационный	мир	такой	граница	данный
вуз	знание	власть	вектор	моделирование
кафедра	высокий	стоимость	уровень	сигнал
личность	величина	весь	коэффициент	следующий
субъект	число	условие	представление	основа
стандарт	соответствующий	высокий	теория	учет
место	соответствие	общий	поле	получение
знанье	реальный	человек	фаза	возможность
сфера	район	крупный	класс	концентрация
конференция	весь	единый	выражение	центр
исследование	лицо	государственный	внешний	снижение
учреждение	еще	действие	возможный	метод
литература	данный	отрасль	операция	количество
направление	ток	регулирование	физический	вид
являться	должен	национальный	вероятность	эффект
мир	частота	поддержка	работа	давление
практический	философия	мера	три	оптимальный
уровень	отдельный	требование	очень	степень

В Таблице 2 сведены результаты сравнения моделей классификации, полученных при постепенном увеличении числа признаков и объёма размеченной выборки. Словарные признаки, полученные с помощью тематической модели классификации, улучшают качество классификации. Кроме того, они позволяют полностью исключить словарные признаки **KW**, отобранные вручную по малой размеченной выборке. Также хорошо виден эффект пессимистичного смещения оценок качества классификации. Поскольку для разметки выбирались в основном документы из «пограничной зоны» неуверенной

классификации, именно на них качество несколько хуже (F1-мера 90,85 %). Однако пересчёт на действительное распределение объектов в полной выборке показывает, что F1-мера на самом деле существенно выше и составляет 97,33 %. При этом относительные ранжировки моделей классификации при обоих способах оценивания практически совпадают.

В Таблице 3 показаны оценки важности основных числовых признаков документов. Оценкой важности является разность F1-меры при использовании данного признака и при исключении данного признака. Наиболее важными оказались признаки библиографии, греческих и математических символов.

Таблица 2: Сравнение моделей классификации, построенных при разных наборах признаков. Первые три колонки содержат значения F1-меры, полноты и точности при перевзвешивании объектов по формуле Хансена-Гурвица. Правые три колонки получены при равных весах объектов. Все признаки: GL, MS, BS, DC, SS, PG + Bib + Ref + KW. Theta и 8TM вместе не дают улучшения.

Признаки	F1, %	Recall, %	Prec, %	F1, %	Recall, %	Prec, %
GL, MS, BS	76,57	62,08	99,90	77,40	71,48	84,39
GL, MS, BS + Theta	93,27	91,54	95,07	77,38	72,35	83,13
GL, MS, BS, DC, SS, PG	93,18	90,63	95,87	81,86	83,33	80,06
GL, MS, BS, DC, SS, PG + Theta	92,62	90,13	95,24	81,95	80,21	83,76
GL, MS, BS, DC, SS, PG + 8TM	95,12	95,52	94,72	82,24	78,54	86,31
Все признаки	96,24	95,92	96,57	90,37	91,00	89,75
Все признаки + Theta	96,50	96,22	96,78	90,51	93,21	87,95
Все признаки + 8TM	97,33	97,47	97,2	90,85	93,42	88,41

Таблица 3. Признаки в порядке убывания их важности. Оценкой важности является разность F1-меры при использовании данного признака и при исключении данного признака.

Feature	ΔF1
Bib	1,18
MS	0,67
GL	0,48
Ref	0,46
KW:Struct	0,26
KW:Names	0,17
BS	0,06

Feature	ΔF1
KW:StopWords	0,06
DC	0,02
SS	0,05
PG	0,02
KW:Frequent	0,01
KW:Terms	0,01

6. Заключение

Предложена технология жанрово-тематической фильтрации текстового контента, которая может быть использована для автоматического наполнения информационно-поисковых ресурсов в интересах профессиональных сообществ. Технология основана на методах машинного обучения и тематического моделирования. Участие асессоров сведено к минимуму: разметка необходимой выборки релевантных и нерелевантных документов под новый тип контента может быть выполнена за несколько человеко-дней.

Работа выполнена при финансовой поддержке РФФИ, гранты 14-07-00847, 14-07-00908, 14-07-31176 и программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Литература

1. *Blei D. M., Ng A. Y., and Jordan M. I.* (2003), Latent Dirichlet allocation, Journal of Machine Learning Research, Vol. 3, pp. 993–1022.
2. *Hansen M. H., Hurwitz W. N., and Madow W. G.* (1953), Sampling survey methods and theory. John Wiley and Son Inc., New York.
3. *Hofmann T.* (1999), Probabilistic latent semantic indexing, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA: ACM, pp. 50–57.
4. *Joachims T.* (1998), Text categorization with support vector machines: Learning with many relevant features. Proceedings of the 10th European Conference on Machine Learning, pp. 137–142.
5. *Kanaris I. and Stamatatos S.* (2009), Learning to recognize webpage genres. Information Processing and Management. pp. 499–512.
6. *Lewis D. D. and Gale W. A.* (1994), A sequential algorithm for training text classifiers. Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 3–12.
7. *Manevitz L. M. and Yousef M.* (2002), One-class SVMs for document classification. The Journal of Machine Learning research, Vol. 2, pp. 139–154.
8. *Meyer zu Eißen S. and Stein B.* (2004), Genre classification of web pages. Advances in Artificial Intelligence, pp. 256–269.
9. *Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., and Duchesnay E.* (2011), Scikit-learn: Machine learning in Python, Journal of Machine Learning Research, Vol. 12, 2825–2830.
10. *Rafi M. and Shaikh M. S.* (2013), A comparison of SVM and RVM for document classification, available at: arxiv.org/abs/1301.2785.
11. *Ramage D., Hall D., Nallapati R., and Manning C. D.* (2009), Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora, In Proceedings of the 2009 conference on empirical methods in natural language processing, pp. 248–256.

12. *Rubin T. N., Chambers A., Smyth P., and Steyvers M.* (2012), Statistical topic models for multi-label document classification, Machine Learning, Vol. 88, No. 1–2, pp. 157–208.
13. *Stamatatos E., Kokkinakis G., and Fakotakis N.* (2000), Automatic text categorization in terms of genre and author. Computational Linguistics, Vol. 26, No. 4, pp. 471–495.
14. *Zhu J., Ahmed A., and Xing E. P.* (2009), MedLDA: Maximum margin supervised topic models for regression and classification. In Proceedings of the 26th annual international conference on machine learning, pp. 1257–1264.
15. *Vorontsov K. V., and Potapenko A. A.* (2014), Tutorial on probabilistic topic modeling: additive regularization for stochastic matrix factorization. AIST'2014, Analysis of Images, Social networks and Texts. Communications in Computer and Information Science (CCIS), Springer International Publishing Switzerland. Vol. 436. pp. 29–46.