С. Н. Карпович, Многозначная классификация текстовых документов с использованием вероятностного тематического моделирования ml-PLSI, *Тр. СПИИРАН*, 2016, выпуск 47, 92–104

DOI: https://doi.org/10.15622/sp.47.5

УДК 004.912 **DOI:** 10.15622/sp.47.5

С.Н. Карпович

МНОГОЗНАЧНАЯ КЛАССИФИКАЦИЯ ТЕКСТОВЫХ ДОКУМЕНТОВ С ИСПОЛЬЗОВАНИЕМ ВЕРОЯТНОСТНОГО ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ ml-PLSI

Карпович С.Н. **Многозначная классификация текстовых документов с использованием вероятностного тематического моделирования ml-PLSI.**
Аннотация. В работе рассмотрен подход к многозначной классификации текстовых документов на основе вероятностного тематического моделирования. На базе корпуса SCTM-ru построена тематическая модель методом обучения с учителем, приведен алгоритм многозначной классификации. Описан состав программного прототипа, реализующего предложенный подход.
Ключевые слова: многозначная (нечеткая) классификация, обучение с учителем, тематическое моделирование, обработка текста на естественном языке.

Karpovich S.N. **Multi-Label Classification of Text Documents using Probabilistic Topic Model ml-PLSI.**
Abstract. In this paper, we describe an approach to multi-label classification of text documents based on probabilistic topic modeling. On the basis of SCTM-ru a topic model has been built with the help of supervised learning. A multi-label classification algorithm is presented. We propose tools for multi-label classification implementing this approach.
Keywords: multi-label classification, supervised learning, topic model, natural language processing.

1. Введение. В настоящее время количество информации, которое создается в текстовом электронном виде колоссально, год от года объем этой информации увеличивается. Лингвисты, маркетологи и аналитики нуждаются в инструментах автоматической обработки текстов на естественном языке. Одним из направлений обработки текстов является категоризация. Категоризация выполняется различными алгоритмами классификации и кластеризации. Задачи классификации волновали исследователей с середины XIX века, в работе [1] дан обзор ряда алгоритмов классификации 1969–1980 гг. Под задачей кластеризации документов понимают задачу разбиения заданной выборки текстовых документов на непересекающиеся подмножества — кластеры, так чтобы каждый кластер состоял из похожих документов, а документы разных кластеров существенно отличались. Под задачей классификации документов понимают задачу отнесения документа к одной из нескольких категорий (классов, тем) на основании содержания документа. Классификация относится к задачам обучения с учителем, когда алгоритм классификации сначала обучается на размеченных документах, а затем классифицирует новые документы.

Большое количество исследований алгоритмов кластеризации и классификации связаны с определением одной категории, к которой

может принадлежать документ. В реальном мире чаще бывает, что один и тот же документ может быть отнесен к нескольким категориям. Например, статья или новость про футбольный матч может быть отнесена к категориям: спорт, футбол, спортивные соревнования, городские мероприятия. Поэтому особенно актуальны методы и алгоритмы многозначной (нечеткой) классификации (multi-label classification) и мягкой кластеризации.

Цель данной работы — предложить подход к многозначной классификации с использованием методик вероятностного тематического моделирования. Создать тематическую модель на размеченных данных и предсказать категории, к которым относится новый документ, фраза или слово.

2. Обзор существующих алгоритмов многозначной классификации. Multi-label classification — не имеет устоявшегося русскоязычного термина, в литературе встречаются многозначная классификация и нечеткая классификация. В этой работе мы используем термин многозначная классификация. В машинном обучении многозначная классификация представляет собой вариант задач классификации, в которой к каждому документу (классифицируемому объекту) должны быть определены несколько меток. Не следует путать многозначную классификацию с многоклассовой классификацией, цель которой определить один класс из более чем двух классов кандидатов. В работе [2] представлен обзор алгоритмов многозначной классификации. Существует два основных метода для решения задач многозначной классификации: методы преобразования проблемы и метод адаптации. Метод преобразования проблемы трансформирует задачу в набор двоичных классификационных задач. Методы адаптации выполняют классификацию множества меток классов, решают задачу в ее полном виде.

Для решения задачи многозначной классификации используют адаптированные версии алгоритмов классификации, такие как: Boosting (AdaBoost), k-ближайших соседей, деревья решений, ядерные методы, SVM, нейронные сети. Метрики оценки качества многозначной классификации отличаются от обычной классификации в силу особенности задачи.

Тематическое моделирование — это способ построения тематической модели коллекции текстовых документов. Тематическая модель задает отношение между темами и документами в корпусе текстов. В обзоре [3] рассмотрены пять основных классов вероятностных тематических моделей: базовые, учитывающие отношения между документами, учитывающие отношения между

словами, темпоральные, обучаемые с учителем. Тематические модели задают мягкую кластеризацию слов и документов по кластерам-темам, означающую, что слово или документ могут быть отнесены сразу к нескольким темам с различными вероятностями. В результате синонимы с большой вероятностью будут отнесены к одной теме, так как часто употребляются в рамках одних и тех же контекстов, а омонимы попадут в разные, так как их контексты различаются. Тематические модели, как правило, основаны на гипотезе «мешка слов» и «мешка документов», т.е. порядок документов в коллекции не имеет значения и порядок слов в документе не имеет значения.

В работе [4] описан алгоритм тематической модели классификации под названием Label-LDA. В основе работы алгоритма лежит базовый алгоритм LDA, векторы документов и векторы тем порождаются распределением Дирихле. Основан он на двух сильных ограничениях, темы отождествляются с классами, предполагается, что для каждого документа точно известно множество всех классов, к которым он относится. Аналогичные ограничения касаются алгоритма Flat LDA, описанного в работе [5]. Для задачи классификации несбалансированных классов в этой работе был предложен алгоритм Prior-LDA, использующий частотную регуляризацию. По утверждению авторов работы [5] Flat-LDA, Prior-LDA являются частными случаями более общего алгоритма Dependency LDA, в котором предлагается моделировать классы документов через распределение тем документов и вводится новая неизвестная матрица класс-тема. В работе [6] предложен подход к многозначной классификации методом LDA с использованием знаний толпы под названием ML-PA-LDA-C. Используется информация не только о присутствующем классе, но и об отсутствующем применяется для построения модели по зашумленным размеченным данным. В работе устранено одно ограничение Label-LDA, предполагается, что точно не известно множество всех классов, к которым принадлежит документ.

3. Многозначная классификация с использованием Вероятностного тематического моделирования (multi-label Probabilistic Latent Semantic Indexing — ml-PLSI). Модели, разработанные на основе латентного размещения Дирихле (LDA), как указано в работе [7], не имеют сильных лингвистических обоснований. При этом классическая модель вероятностного латентно семантического анализа PLSA [8] не связана с какими-либо параметрическими априорными распределениями.

Пусть D — множество текстовых документов, W — словарь терминов. Каждый документ $d \in D$ представляет собой последовательность терминов n_d терминов $(w_1,\ldots,w_{n_d})$ из словаря W.

С учетом гипотезы условной независимости $p(w|d,t) = p(w|t)$ по формуле полной вероятности получаем вероятностную модель порождения документа d:

$$p(w|d) = \sum_{t\in T} p(w|d,t)\, p(t|d) = \sum_{t\in T} p(w|t)\, p(t|d) = \sum_{t\in T} \varphi_{wt}\theta_{td},$$

Для вычисления φ_{wt} и θ_{td} используется EM-алгоритм.

Согласно вероятностному тематическому моделированию, впервые предложенному в работе [8], вероятностная модель появления пары «документ-слово» может быть записана тремя эквивалентными способами:

$$p(d,w) = \sum_{t\in T} p(t)\, p(w|t)\, p(d\,|\,t) = \sum_{t\in T} p(d)\, p(w|t)\, p(t\,|\,d) =$$
$$= \sum_{t\in T} p(w)\, p(t|w)\, p(d\,|\,t),$$

где: T — множество тем;

$p(t)$ — неизвестное априорное распределение тем в коллекции;

$p(d)$ — априорное распределение на множестве документов, эмпирическая оценка $p(d) = n_d / n$, где $n = \sum_d n_d$ — суммарная длинна всех документов, а n_d — длина документа в словах;

$p(w)$ — априорное распределение на множестве слов, эмпирическая оценка $p(w) = n_w / n$, где n_w — число вхождений слова w во все документы.

Если мы отождествим понятие темы тематической модели и категории документа, учтем, что задача построения тематической модели имеет бесконечно много решений [9], то сможем построить один из вариантов тематической модели, обучившись на размеченном корпусе. Построенная на данных предположениях тематическая модель зависит от качества выбранной для обучения коллекции. Например, категории в корпусе SCTM-ru проставлены авторами новостей, перед авторами не стояла задача указать все категории, которые только возможно для каждой новости, поэтому часть

документов не получили полный набор категорий, даже если они этого заслуживали. При этом объем корпуса позволяет предположить, что в своем большинстве авторы использовали наиболее характерные категории. Основываясь на знании толпы, мы можем рассчитывать вероятностную оценку отнесения слова к категории:

$$p(w|d) = \sum_{t \in T} p(w|t)\, p(t \mid d) = \sum_{t \in T} p(w|c)\, p(c|d) = \sum_{t \in T} \varphi_{wt} \theta_{td}.$$

Построенная таким образом тематическая модель может быть одним из множества решений задачи тематического моделирования. В работе [10] рассмотрен алгоритм создания тематической модели методом обучения с учителем.

Для уменьшения размерности векторного пространства рекомендуется все слова в корпусе привести к нормальной словоформе. Если документов в корпусе немного и алгоритм будет выполнен за конечное время, то этого делать необязательно, т.к. словоформа также может стать важной информацией для определения категории документа, слова в конкретной словоформе чаще могут встречаться в документах, принадлежащих одной категории. На основании семиологии, [11] слово в знаковой системе наделено смыслом и является частью языка. Поэтому в данной работе мы не приводим слова в корпусе к нормальной словоформе.

Алгоритм 1. Алгоритм построения тематической модели методом обучения с учителем:

Вход: коллекция документов D с указанием категорий C.

Выход: распределения $p(w|c), p(d|c), p(w \mid d)$:

1. Для всех $d \in D, w \in d$:
   a. Считаем $p(w|d) = {n_{dw}}/{n}$.
2. Для всех $c \in C, d \in D$:
   a. Считаем $p(d|c) = {n_c}/{n}$.
3. Для всех $w \in W, c \in C$:
   a. Считаем $p(w|c) = \varphi_{wt} = {n_{dwc}}/{n_{dw}}$.

Обучить тематическую модель по размеченным данным — это значит рассчитать матрицы «слово-документ», «документ-категория» и «слово-категория» для каждого слова из коллекции документов. На первом шаге рассчитываем матрицу «слово-документ». Значения матрицы — это количество повторений слова в документе. На втором

шаге рассчитываем матрицу «документ-категория». Для этого для каждого документа в корпусе получаем список категорий, к каждой категории документ может быть отнесен не более одного раза, поэтому значения матрицы «документ-категория» — это единицы в том случае, если документ связан с категорией, и нуль, если такой связи нет. На третьем завершающем шаге рассчитываем матрицу «слово-категория». Значения матрицы — это вероятность встретить слово в этой категории. Как мы ранее отмечали, разметка документов категориями, содержит ошибки. Эти ошибки можно разделить на два вида:

1. Слово, которое имеет отношение к определенной категории, редко встречается в корпусе поэтому связь между словом и категорией не установилась.
2. Документ отмечен какой-либо категорией по ошибке.

В данной работе для уменьшения влияния ошибок на результат предлагаем использовать регуляризацию.

В результате обучения тематической модели мы получили три матрицы, по которым мы можем узнать, с какой вероятностью то или иное слово относится к теме, из каких слов формируется тема и к каким темам относится каждый документ. По матрице «слово-документ» можем восстановить каждый документ в формате мешка слов. По матрице «слово-категория» можем оценить с какой вероятностью то или иное слово относится к категории, можем рассчитать вероятность отнесения нескольких слов к категориям, для этого достаточно просуммировать вероятность отнесения каждого слова к категории $\sum_{w \in d} p(w \mid c)$. По матрице «документ-категория» можем получить список всех документов, которые связаны с категорией. Для примера в таблице 1 приведены наиболее характерные и наиболее часто встречаемые слова для трех категорий: спорт, происшествия и политика.

Таблица 1. Пример слов характерных для категорий

Спорт	Происшествия	Политика
Зимнему	Даги	Реймер
Виндсёрфинг	Дагов	Халип
Перепёлкин	Гоа	Гольман
Трофименко	Моди	Минтимер
Заезда	Зингер	Муртаза
Кайтингу	Ксанте	Рахимова
Полумарафона	Ока	Дарькина
Фонак	Реанимация	Спутникам
Гимнаст	Стропила	Хамитов

Стоит заметить, что эти слова однозначно характеризуют категорию, к которой относятся, но не являются часто употребимыми во всем корпусе; зачастую это имена собственные и фамилии либо редкие по написанию слова с буквой «ё».

4. Многозначная классификация. Переходя к следующему шагу, мы имеем тематическую модель, в которой определены отношения между словами и категориями. Для предсказания категорий нового документа, выполняем Алгоритм 2

Алгоритм 2. Многозначная классификация на базе вероятностного тематического моделирования:

Вход: Тематическая модель, новый документ d_{new}.

Выход: Список предсказанных категорий:

1. Для всех $w \in d_{new}$:
   a. $\sum_{w \in d} p(w \mid c)$.

2. Возвращаем список категорий по убыванию суммы вероятностных оценок.

Следует учитывать, что слова, которые есть в новом документе и отсутствуют в корпусе, на котором обучалась тематическая модель, не будут учтены при предсказании категорий.

Для того чтобы отобрать только наиболее релевантные категории используем регуляризацию, а именно:

– Регуляризация по документам. Если слово встречается в большом количестве документов, то оно перестает быть для нас информативным, вероятность отнесения документа к теме по этому слову может быть сведена к нулю.

– Регуляризация по темам. Если слово встречается в большом количестве категорий, то оно перестает быть для нас информативным, вероятность отнесения слова к категории по этому слову может быть сведена к нулю.

5. Эксперимент с корпусом SCTM-ru. В качестве данных для исследования используем корпус SCTM-ru [12], созданный специально для тестирования задач тематического моделирования. Источником данных корпуса является международный новостной сайт «Русские Викиновости». Корпус SCTM-ru состоит из 7000 документов, 185 авторов, почти 12000 уникальных категорий. События, описанные в документах, распределены по более чем 2000 уникальным датам, с

ноября 2005 года по июнь 2014 года. В корпусе SCTM-ru 2400000 словоупотреблений, состоящих только из русских букв. Словарный состав корпуса — 150600 уникальных словоформ, 59000 уникальных лемм.

Каждая новость содержит указанные автором категории. У автора новости не стояла задача перечислить все категории, к которым новость может иметь отношение, тем не менее указанные категории дают весомые основания полагать, что новость сильно связана с темой этих категорий. Обычно новость определена автором к нескольким категориям, поэтому мы рассматривали алгоритмы многозначной классификации. Для каждого документа мы не знаем точного множества всех категорий.

В корпусе SCTM-ru часть категорий авторы используют редко, 19284 категориями размечено менее 50 новостей из корпуса. Так как эти категории не представляют большой ценности для нашей тематической модели и могут создавать дополнительные трудности при прогнозировании категорий новых документов, мы не будем их учитывать при построении тематической модели. Категории, которыми размечены более 50 новостей — 230 штук, далее их называем рабочие категории или просто категории, их мы используем для обучения тематической модели. Всего документов, размеченных рабочими категориями 6428. Для обучения модели используем 5000 новостей. Чтобы отсеять не информативные слова, междометия и предлоги, создаем словарь стоп-слов, в который вошло 151 слово.

Для разработки прототипа программы используется дистрибутив Anaconda, язык разработки Python и модули для машинного обучения pandas, numpy, scikit-learn. Векторизуем документы, создаем матрицу $p(w \mid d)$. Затем создаем вектора категорий и создаем матрицу $p(d \mid c)$. Затем рассчитываем матрицу $p(w|c)$. Таким образом, мы обучили тематическую модель на размеченных данных корпуса SCTM-ru.

6. Оценка качества полученной тематической модели. Наиболее известным критерием является коэффициент неопределенности, используемый для оценки качества различных моделей. Коэффициент неопределенности представляет собой меру несоответствия модели $p(w|d)$ словам w, наблюдаемым в документах коллекции, и определяется через логарифм правдоподобия:

$$Perplexity(D) = \exp(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw} log\ p(w|d)),$$

где n — число всех рассматриваемых слов в текстовой коллекции, D — множество всех документов в коллекции, n_{dw} — частотность слова w в документе d, $p(w|d)$. — вероятность появления слова w в документе d. Чем меньше значение коэффициента неопределенности, тем лучше модель предсказывает появление слов в документах коллекции. Коэффициент неопределенности полученной тематической модели 0,11. В работе [13] отмечено, что коэффициент, вычисленный по той же самой коллекции, дает оптимистически заниженную оценку. Для предсказания тематических категорий используем незатронутые в обучении новости.

Используем следующие метрики оценки качества многозначной классификации:

1. Функция потерь Хэмминга для ошибочных предсказаний.
2. Количество документов, в которых первая предсказанная категория, совпала с любой из указанных автором новости.
3. Процент верных предсказаний из первых 50 предсказанных категорий.

Результаты оценки качества приведены в таблице 2.

Таблица 2. Оценка качества многозначной классификации

Метрика качества	Результат (100 новостей)	Результат (500 новостей)
Функция потерь Хэмминга	0,18	0,17
Первая предсказанная (точность)	82%	84%
Процент верных предсказаний из 50 первых категорий (полнота)	71%	72%

Построенная тематическая модель позволяет определить наиболее вероятные категории не только для текстового документа, но и для отдельной фразы или слова. Например, для фразы «выборы президента России» десять первых категорий отображены на рисунке 1, для слова «футбол» на рисунке 2 показаны десять наиболее вероятных категорий.

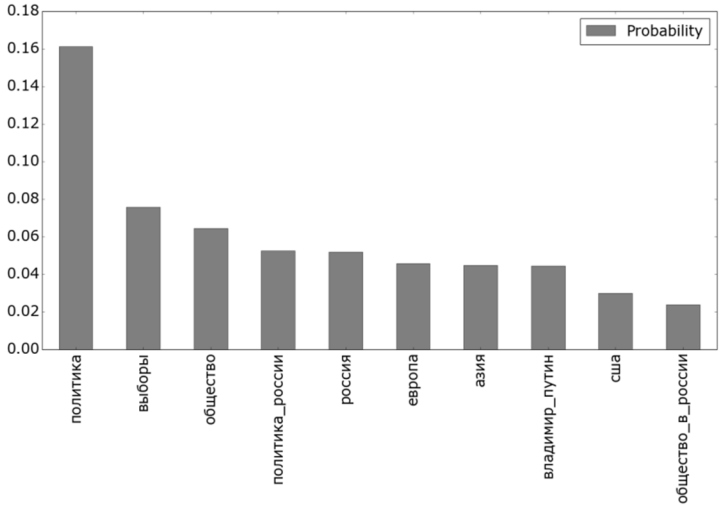

Рис. 1. Наиболее вероятные категории для фразы «выборы президента России»

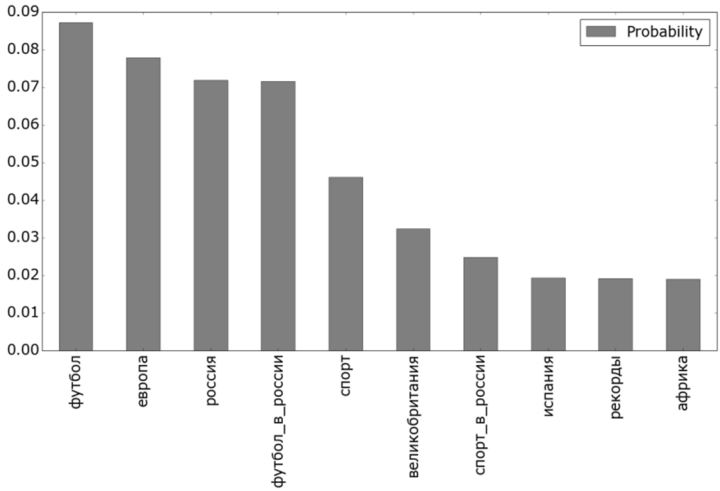

Рис. 2. Наиболее вероятные категории для слова «футбол»

6. Заключение. В результате проделанной работы разработан подход к обучению тематических моделей с учителем (supervised Probabilistic Latent Semantic Indexing — PLSI). Предложен метод многозначной классификации текстовых документов с использованием обученной тематической модели ml-PLSI. Проведенные эксперименты на корпусе SCTM-ru демонстрируют

перспективность использования тематического моделирования в задачах многозначной классификации. Поставленные в работе цели были достигнуты.

Вероятностная тематическая модель может быть использована в задачах ассоциативной классификации [14] в комбинации с другими алгоритмами классификации, а также может быть решателем в алгоритмах коллективного распознавания, описанных в работе [15].

Далее будут продолжены исследования возможностей многозначной классификации методом вероятностного тематического моделирования, будет проверена гипотеза обучения модели новыми словами за счет уже имеющийся информации о связях слов и категорий. Эксперимент с программной частью метода доступен на <https://github.com/cimsweb/mlPLSI>.

Литература

1. *Журавлёв Ю.И., и др.* Задачи распознавания и классификации со стандартной обучающей информацией // Журнал вычислительной математики и математической физики. 1980. Вып. 20. № 5. С. 1294–1309.
2. *Tsoumakas G., Katakis I.* Multi-label classification: an overview // International Journal of Data Warehousing & Mining. 2007. vol. 3(3). pp. 1–13.
3. *Daud A. et al.* Knowledge discovery through directed probabilistic topic models: a survey // Frontiers of computer science in China. 2010. vol. 4. no. 2. pp. 280–301.
4. *Ramage D., Hall D., Nallapati R., Manning C. D.* Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. USA. 2009. vol. 1. pp. 248–256.
5. *Rubin T.N., Chambers A., Smyth P., Steyvers M.* Statistical topic models for multilabel document classification // Machine Learning. 2012. vol. 88. no. 1–2. pp. 157–208.
6. *Padmanabhan D. et al*. Topic Model Based Multi-Label Classification from the Crowd // arXiv preprint arXiv:1604.00783. 2016.
7. *Воронцов К.В., Потапенко А.А.* Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных. 2013. Вып. 1. № 6. С. 657–686.
8. *Hoffman T.* Probabilistic Latent Semantic Indexing // Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval. 1999. pp. 50–57.
9. *Воронцов К.В., Потапенко А.А*. Регуляризация, робастность и разреженность вероятностных тематических моделей // Компьютерные исследования и моделирование. 2012. Вып. 4. № 4. С. 693–706.
10. *Blei D., McAuliffe J.* Supervised topic models //Advances in neural information processing systems. 2008. vol. 20. pp. 121–128.
11. *Плохотнюк В.С.* Аксиоматизация семиологии и научный статус семиотики // Terra economicus. 2010. Вып. 8(4). С. 124–132.
12. *Карпович С.Н.* Русскоязычный корпус текстов SCTM-RU для построения тематических моделей // Труды СПИИРАН. 2015. Вып. 39. С 123–142.
13. *Blei D., Ng A., Jordan M.* Latent Dirichlet Allocation // Journal of Machine Learning Research // MIT Press. 2003. vol. 3(Jan). pp. 993–1002.
14. *Городецкий В.И., Тушканова О.Н.* Ассоциативная классификация: аналитический обзор. Часть 1 // Труды СПИИРАН. 2015. Вып. 1(38). С 183–203.
15. *Городецкий В.И., Тушканова О.Н.* Ассоциативная классификация: аналитический обзор. Часть 2. // Труды СПИИРАН. 2015. Вып. 2(39). С 212–240.

16. *Городецкий В.И., Серебряков С.В.* Методы и алгоритмы коллективного распознавания // Труды СПИИРАН. 2006. №3. С 139–171.

Reference

1. Zhuravlev Yu.I., et. al. [Recognition and classification problems with standard training information]. *Zh.-vychisl.-matem.-i-matem.-fiz. – Comput. Math. Math. Phys.*1980. vol. 20. no. 5. pp. 1294–1309. (In Russ.).
2. Tsoumakas, Grigorios; Katakis, Ioannis. Multi-label classification: an overview. *International Journal of Data Warehousing & Mining.* 2007. vol. 3(3). pp. 1–13.
3. Daud A. et al. Knowledge discovery through directed probabilistic topic models: a survey // Frontiers of computer science in China. 2010. vol. 4. no. 2. pp. 280–301.
4. Ramage D., Hall D., Nallapati R., Manning C. D. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. USA. 2009. vol. 1. pp. 248–256.
5. Rubin T.N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multilabel document classification. *Machine Learning.* 2012. vol. 88. no. 1–2. pp. 157–208.
6. Padmanabhan D. et al. Topic Model Based Multi-Label Classification from the Crowd. arXiv preprint arXiv:1604.00783. 2016.
7. Vorontsov K.V., Potapenko A.A. [EM-like algorithms for probabilistic topic modeling]. *Mashinnoe obuchenie I analiz dannyh – Machine Learning and Data Mining.* 2013. vol. 1. no. 6. pp. 657–686. (In Russ.).
8. *Hoffman T.* Probabilistic Latent Semantic Indexing. Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval. 1999. pp. 50–57.
9. Vorontsov K.V., Potapenko A.A. [Regularization, robustness and sparsity of probabilistic topic models]. *Kompyuternyye issledovaniya i modelirovaniye — Computer research and modeling*. 2012. vol. 4 no. 4. pp. 693–706. (In Russ.).
10. Blei D., McAuliffe. J. Supervised topic models. *Advances in neural information processing systems.* 2008. vol. 20. pp. 121–128.
11. Plokhotnuk V.S. [Axiomatization of semiology and scientific status of semiotics]. Terra economicus. 2010. vol. 8(4). pp. 124–132. (In Russ.).
12. Karpovich S.N. [The Russian language text corpus for testing algorithms of topic model]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2015. vol. 39. pp 123–142. (In Russ.).
13. Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. *Journal of Machine Learning Research*. 2003. vol. 3. pp. 993–1002.
14. Gorodetsky V.I., Tushkanova O.N. [Associative classification: analytical overview. Part 1]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2015. vol. 1(38). pp. 183–203. (In Russ.).
15. Gorodetsky V.I., Tushkanova O.N. [Associative classification: analytical overview. Part 2.]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2015. vol. 2(39). pp 212–240. (In Russ.).
16. Gorodetsky V.I., Serebryakov S.V. [Methods and algorithms of the collective recognition: a survey]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2006. vol. 3. pp. 139–171. (In Russ.).

Карпович Сергей Николаевич — руководитель направления поисковой оптимизации, ООО "Рамблер Интернет Холдинг". Область научных интересов: тематическое моделирование, обработка текстов на естественном языке, data mining. Число научных публикаций — 1. cims@yandex.ru, http://www.cims.ru; Варшавское ш., 9, стр. 1, БЦ «Даниловская мануфактура», корпус «Ряды Солдатенкова», Москва, 117105; р.т.: +7(495)7851700.

Karpovich Sergey Nikolaevich — head of search engine optimization direction, Rambler Internet Holding LLC. Research interests: topic model, natural language processing, classification, clustering, data mining. The number of publications — 1. cims@yandex.ru, http://www.cims.ru; 9, Varshavskoe sh., str. 1, BC «Danilovskaja manufaktura», k. «Rjady Soldatenkova», 117105, Moscow; office phone: +7(495)7851700.

РЕФЕРАТ

Карпович С.Н. **Многозначная классификация текстовых документов с использованием вероятностного тематического моделирования ml-PLSI.**

В статье рассмотрен подход к задаче многозначной классификации текстовых документов с использованием методик вероятностного тематического моделирования. Большое количество исследований алгоритмов классификации и кластеризации, связаны с определением одной категории, к которой может принадлежать документ. Зачастую один документ может быть отнесен к нескольким категориям. Обозначена актуальность задачи. Проведен обзор существующих алгоритмов многозначной классификации.

Описана программная реализация алгоритма многозначной классификации. Методом обучения с учителем построена тематическая модель. Приведены оценки качества классификации, представлен пример предсказания возможных категорий для фразы.

Предложенный подход продемонстрировал свою эффективность. Вероятностные оценки отнесения документа к категории позволяют использовать его в задачах коллективного распознавания и в задачах ассоциативной классификации. Далее будут продолжены исследования возможностей многозначной классификации методом вероятностного тематического моделирования.

SUMMARY

Karpovich S.N. **Multi-Label Classification of Text Documents using Probabilistic Topic Model ml-PLSI.**

The paper considers an approach to multi-label classification of text documents based on probabilistic topic modeling. A large number of studies of clustering and classification algorithms are related to determination of one label to which a document may belong. Often one document can be relevant to many labels. The significance of this task is shown. A comparative analysis of multi-label classification algorithms has been conducted.

The article describes technology tools for multi-label classification. A topic model has been built with the help of supervised learning. Evaluation of classification quality is given, and an example of the prediction of possible categories for the phrase is presented. The developed approach has proven to be efficient. Probabilistic estimations of categorizing a document allow using it for collective recognition and associative classification tasks. We will continue our studies on the opportunities of multi-label classification by the method of probabilistic topic modeling.