Автоматическая классификация документов на основе латентно-семантического анализа*

Игорь Куралёнок, Игорь Некрестьянов

Санкт-Петербургский Государственный Университет

emails: ik@oasis.apmath.spbu.ru, igor@meta.math.spbu.ru

Abstract

В работе рассматривается задача автоматической классификации документов по множеству заданных тематик. Предлагаемый подход использует метод латентно-семантического анализа для извлечения семантических зависимостей между термами. На основе этих зависимостей и происходит классификация документов.

Эксперименты на базе стандартных тестовых данных Text REtrieval Conference продемонстрировали перспективность предложенного подхода. Вычислительная трудоемкость метода на этапе классификации относительно невелика, что позволяет применять предлагаемый подход при классификации потоков документов.

1 Введение

В связи с бурным развитием Интернет все более актуальными становятся проблемы организации эффективного доступа к информации. В частности, в последние годы много внимания привлекает проблема *автоматической классификации* документов по определенному множеству тематических интересов.

В близком классе задач *фильтрации* документов [13, 12, 3] главной целью является обнаружение потенциально интересующих пользователя документов, по описанию множества тематических интересов пользователя, за счет отсева прочих документов.

Особенностью задачи классификации является предположение, что классифицируемое множество документов не содержит "мусора", т.е. каждый из документов соответствует какой-нибудь из заданных тематик. В силу этой особенности, методы применяемые в задачах фильтрации показывают не лучшие результаты в области классификации. Поэтому проблеме классификации разнообразной динамической информации за последнии несколько лет было посвящено много научных работ [4, 18, 8, 16, 19, 5, 17].

*Эта работа выполнялась в рамках международного европейского проекта OASIS (номер контракта PL 96 1116)

Первая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
19 - 21 октября 1999 г., Санкт-Петербург

Большинство предложенных методов классификации [18, 13] основываются на использовании простой векторной модели описания документов (Vector Space Model) — классической модели в области поиска информации. В рамках этой модели документ описывается вектором, в котором каждому используемому в документе терму сопоставляется его значимость (вес) в рамках этого документа. Значимость терма основывается на статистической информации о встречаемости термов в рамках этого и возможно других документов. Описание тематики также представляется вектором и для оценки близости документа и тематики используется скалярное произведение векторов описания тематики и вектора документа.

В последние годы в задачах организации доступа к информации, в том числе и в области автоматической классификации, все больше внимания привлекают более сложные подходы, обеспечивающие лучшее качество [12, 5, 17].

Одним из перспективных направлений является применение *латентно-семантического анализа* (LSA) [15] для выявления структуры семантических взаимосвязей между используемыми словами, за счет статистического анализа большой группы документов. Благодаря этому становится возможным автоматически отличать различные смысловые оттенки одного и того же слова в зависимости от контекста его использования. Отметим, что выявление семантической структуры при помощи латентно-семантического анализа происходит полностью автоматически и не требует ручного составления словарей, и т. п.

Мы занимались исследованием вопросов классификации документов в рамках проекта OASIS[1].

Описываемый в этой работе метод основывается на применении латентно-семантического анализа [15]. Хотя предлагаемый подход и требует много вычислительных ресурсов на подготовительном этапе, на этапе классификации вычислительные затраты невелики, что позволяет использовать этот метод в системах автоматической классификации потоков документов.

Еще одной отличительной особенностью нашей работы является использование обширного тестового набора данных. Во многих опубликованных работах по классификации для экспериментов использовались маленькие

[1] Проект OASIS (Open Architecture Server for Information Server and Delivery) занимается разработкой архитектуры распределенной поисковой системы на базе тематических коллекций. Дополнительная информация о проекте общедоступна в Интернет по адресу www.oasis-europe.org .

наборы данных (до 1000 документов), и/или зачастую большая часть доступных данных использовалась на этапе настройки системы. Поэтому зачастую качество работы таких методов в реальных условиях оказывается значительно хуже качества продемонстрированого во время экспериментов [13].

В наших экспериментах использовалось более 25000 документов относящихся к 104 тематикам. При этом многие тематики сильно перекрываются, т.е. один и тот же документ может относится сразу к нескольким тематикам.

Несмотря на сложность экспериментальной базы проведенные эксперименты показывают хорошее качество классификации с помощью описываемого метода и подтверждают перспективность предлагаемого подхода.

Далее статья организована следующим образом. В следующем разделе кратко излагаются основы латентно-семантического анализа. В разделе 3 описывается базовая теоретическая основа метода, экспериментальная база и результаты экспериментов представлены в разделе 4. В последнем разделе обсуждаются вычислительные характеристики метода и возможности его дальнейшего улучшения.

2 Латентно-семантический анализ (LSA)

Латентно-семантический анализ (LSA[2]) — это теория и метод для извлечения контекстно-зависимых значений слов при помощи статистической обработки больших наборов текстовых данных [15]. В течении нескольких последних лет этот метод не раз использовался как в области поиска информации [1, 10], так и в задачах фильтрования и классификации [12].

Латентно-семантический анализ основывается на идее, что совокупность всех контекстов, в которых встречается и не встречается данное слово, задает множество обоюдных ограничений, которые в значительной степени позволяют определить похожесть смысловых значений слов и множеств слов между собой.

Наиболее распространенный вариант LSA основан на использовании разложения матрицы по сингулярным значениям (SVD[3]) [7]. Огромная исходная матрица *термы-на-документы*, которая описывает используемый набор данных, разлагается во множество из k, обычно от 70 до 200, ортогональных матриц, линейная комбинация которых является неплохим приближением исходной матрицы.

Более формально, любая прямоугольная матрица X, например матрица *термы-на-документы* размерности $t \times d$, может быть разложена в произведение трех других матриц:

$$X = U\Sigma V^T \tag{1}$$

таких, что матрицы U и V состоят из ортонормированных колонок, Σ — диагональная матрица, а r — это ранг матрицы X. Такое разложение и называется разложением по сингулярным значениям.

Если в Σ оставить только k наибольших сингулярных значений и выбрать только соответствующие колонки в матрицах U и V, то произведение получившихся матриц Σ_{lsa}, U_{lsa} и V_{lsa} будет наилучшим приближением исходной матрицы X матрицей ранга k:

$$X \simeq \hat{X} = U_{lsa}\Sigma_{lsa}V_{lsa}$$

Основная идея латентно-семантического анализа в том, что матрица $\hat{X}$, содержащая только k первых линейно независимых компонент X, отражает основную структуру ассоциативных зависимостей присутствующих в исходной матрице, в то же время не содержит шума.

Таким образом каждый терм и документ представляются при помощи векторов в общем пространстве размерности k (так называемом *пространстве гипотез*). Близость между любой комбинацией термов и/или документов может быть легко вычислена при помощи скалярного произведения векторов.

Выбор наилучшей размерности k для LSA — открытая исследовательская проблема. В идеале, k должно быть достаточно велико для отображения всей реально существующей структуры данных, но в то же время достаточно мало чтобы не захватить случайные и маловажные зависимости. Если выбранное k слишком велико, то метод теряет свою мощь и приближается по характеристикам к стандартным векторным методам. Слишком маленькое k не позволяет улавливать различия между похожими словами или документами. Исследования показывают, что с ростом k качество сначала возрастает, а потом начинает падать [9].

3 Классификация с учетом семантической близости слов

В рамках этой работы мы рассматриваем классическую задачу классификации документов по заданному набору тематик Ω [5, 19, 16, 14, 17]. Задача состоит в определении для каждого поступающего в систему документа одной (или нескольких) тематик к которым этот документ относится. Отметим, что в отличие от задачи фильтрации документов [12], здесь подразумевается что, в системы не поступает "мусор", т.е. что каждый из рассматриваемых документов в действительности относится хотя бы к одной из заданных тематик.

Все методы классификации используют один и тот же обобщенный алгоритм, который состоит из следующих этапов:

- задания/построения описаний для всех тематик
- построения описания рассматриваемого документа
- вычисления оценок близости между описаниями тематик и описанием документа и выбора наиболее близких тематик

Методы реализации этих этапов и отличают один метод классификации от другого.

3.1 Описания тематик и документов

В рамках этой работы мы основывались на предположении, что тематика документа определяется его словарным запасом. При этом естественно предположить, что различные синтаксические формы одного и того же слова не отражаются на общей тематике документа и следовательно могут представляться единой базовой словоформой или *термом*. Кроме этого мы не рассматривали так называемые *стоп-слова*, наиболее употребительные слова, которые могут использоваться в документах любой

[2] LSA — Latent Semantic Analisys
[3] SVD — Singular-value decomposition

тематики. Примерами таких слов являются такие слова как: yes, have, and[4].

В качестве описания документа используется все множество встречающихся в документе термов, за исключением общеупотребительных.

Тематики также представляются в системе наборами термов, однако эти наборы содержат не все употребляющиеся в данной тематике слова, а только небольшое, автоматически выбранное их подмножество.

3.1.1 Построение описаний тематик

Вообще, тематика в рамках рассматриваемого подхода задается относительно небольшим множеством относящихся к ней документов. Внутренее описание тематики в виде набора термов автоматически строится по результатам анализа этого множества документов, а также множества документов задающих остальные тематики системы.

Целью этого анализа является выявление отличий этой тематики по сравнению с другими и выбору термов, наилучшим образом подчеркивающих особенности этой тематики.

Выбор слов для описания каждой из тематик производится при помощи следующего алгоритма:

Построение общего словаря термов W: В этот словарь включаются все термы, которые используются хотя бы в одном из документов задающих тематики.

Вычисление вероятностных оценок: Для каждого терма $w \in W$ вычисляется оценка вероятности его использования в документах данной тематики C:

$$TermProb(w, C) = \frac{|\{d : d \in C, d \supset w\}|}{|C|}$$

Построение "тематических" словарей: Для каждой тематики C строится "тематический" словарь. В этот словарь попадают термы, вероятность использования которых в этой тематике превосходит вероятность их использования в любой другой тематике $C_i \in \Omega$, т.е.

$$TermProb(w, C) \geq \frac{\sum\limits_{i \in \Omega} TermProb(w, C_i)}{|\Omega|}$$

Для каждого из отобранных термов вычисляется его *значимость* в рамках данной тематики:

$$\mathrm{TermValue}(w, C) = \frac{\mathrm{TermProb}^3(w, C)}{\sum\limits_{C_i \in \Omega} \mathrm{TermProb}(w, C_i)^2}$$

Отбор термов для описания: Значимость термов, полученная на предыдущем этапе, задает отношение порядка на каждом из "тематических" словарей. Используя это отношение из "тематического" словаря тематики, выбирается несколько термов для использования в качестве описания этой тематики.

[4]Поскольку эксперименты проводились с англоязычными документами, то и приведены английских стоп-слов. Русскими стоп-словами являются, например: *да, как, мы, или.*

Число тематических коллекций	104
Среднее число документов в коллекции	394
Общий размер коллекций (в документах)	40970
Число различных документов в коллекциях	25181
Среднее число коллекций, содержащих один и тот же документ	1.6

Table 1: Характеристики данных построенных на основе коллекции LA Times TREC-5

Оптимальное количество термов для включения в описание зависит от конкретной задачи. Наши эксперименты показали, что с ростом числа термов качество классификации вначале улучшается, а потом начинаете ухудшаться. При этом оптимум достигается при небольшом размере описания — от 10 до 30 термов.

3.2 Вычисление оценок близости

Как уже было сказано выше, описываемый подход основывается на предположении, что тематика документа определяется его словарным запасом.

В рамках этой работы мы определяем функцию FSR, которая сопоставляет каждой паре термов оценку их тематической близости, т.е. вероятность их использования в документах одной тематики. Оценка тематической близости документа и тематики определяется тематической близостью термов входящих в их описания.

В наших экспериментах мы рассмотрели несколько вариантов вычисления оценок близости документа и тематики. Наиболее эффективным оказалось вычисление оценки, как среднего арифметического попарных оценок тематической близости термов из описаний документа d и тематики $C \in \Omega$:

$$Goodness(d, C) = \frac{\sum\limits_{w_i^d \in d} \sum\limits_{w_j^C \in C} (FSR(w_i^d, w_j^C))}{|C| \cdot |D|}$$

После того, как мы оценили документ с точки зрения всех коллекций мы можем выбрать одну или несколько наиболее высоких оценок и классифицировать документ в одну или несколько коллекций.

3.3 Тематическая близость термов

Вычисления тематической близости пары термов представляет собой вычисление вероятности использования этой пары термов в документах одной тематики. Эта вероятность оценивается по результатам анализа использования термов в множестве документов, которыми описываются тематики.

Один из возможных вариантов оценки этой вероятности состоит в следующем. По набору документов строится матрица *термы-на-документы* X, строки которой отражают распределение термов по документам. В качестве оценки тематической близости двух термов используется скалярное произведение соответствующих строк этой матрицы. Таким образом для вычисления оценок близости между всеми парами термов достаточно вычислить матрицу XX^T

Такой подход аналогичен классическим методам поиска информации основанных на векторном представлении описания документа. Поэтому ему присущи те же недостатки:

- метод не обнаруживает зависимости между термами, которые часто используются в документах одной и той же тематики, но редко встречаются вместе
- случайные зависимости и ошибки правописания оказывают существенное влияние на получаемые оценки и негативно сказываются на точности метода
- размер матрицы *термы-на-документы* очень велик даже для небольшого (с точки зрения статистики) числа документов и поэтому использование этой матрицы весьма ресурсоемко

Дальнейшим развитием такого подхода является использование латентно-семантического анализа.

По матрице XX^T строится ее аппроксимация $\hat{X}\hat{X}^T$, где $\hat{X}$ — это аппроксимация X полученная методом латентно-семантического анализа на базе разложения по сингулярным значениям (подробности в разделе 2). Таким образом:

$$\hat{X}\hat{X}^T = U_{lsa}^T \Sigma_{lsa} V_{lsa} (U_{lsa}^T \Sigma_{lsa} V_{lsa})^T \Leftrightarrow$$

$$\hat{X}\hat{X}^T = U_{lsa}^T \Sigma_{lsa} V_{lsa} V_{lsa}^T \Sigma_{lsa} U_{lsa}$$

и в силу ортонормированности матрицы V_{lsa}

$$\hat{X}\hat{X}^T = U_{lsa}^T \Sigma_{lsa}^2 U_{lsa}$$

Отметим, что диагональная матрица Σ имеет размерность k, где k — это выбранная при аппроксимации желаемая размерность пространства гипотез. Таким образом при таком подходе трудоемкость вычисления тематической близости двух термов при вычисленных матрицах U_{lsa} и Σ_{lsa} составляет $O(k)$, т.е. она не зависит от количества анализируемых документов и размера общего словаря.

Функция
тематической близости двух термов $FSR(w_1, w_2)$ однозначно задается матрицей $\hat{X}\hat{X}^T$:

$$FSR(w_1, w_2) = \hat{X}\hat{X}^T[w_1, w_2]$$

4 Проверка эффективности метода

Для практической проверки эффективности описываемого метода мы провели ряд экспериментов, используя большой стандартный набор данных в качестве экспериментальной базы.

4.1 Экспериментальная база

В качестве основной экспериментальной базы мы воспользовались набором данных, предоставляемых Text REtrieval Conference (TREC) [11]. Эти наборы являются наиболее известными стандартными наборами тестовых данных в области поиска информации (information retrieval).

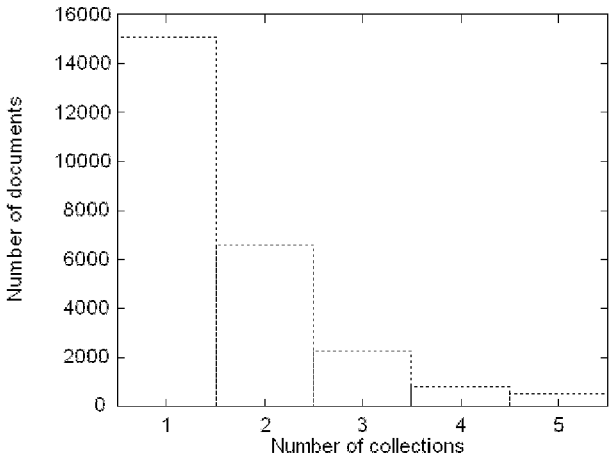

Рис. 1: Распределение документов по количеству соответствующих им тематик

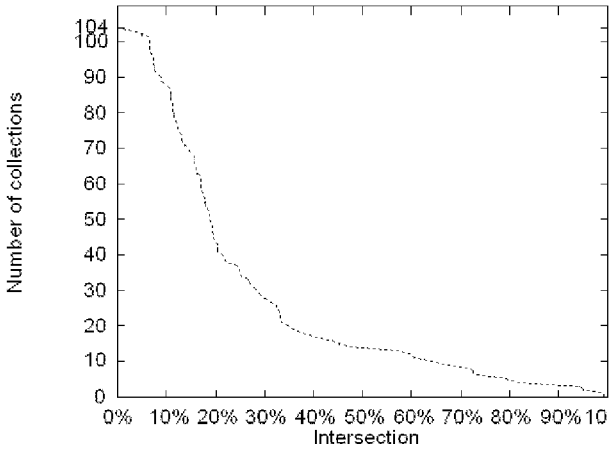

Рис. 2: Процент максимального пересечения тематик

В рамках этой работы мы использовали подмножество документов из коллекции *Los-Angeles Times* с диска TREC-5. Коллекции TREC не разбиты явным образом на тематические группы, но для каждого документа из коллекции *Los-Angeles Times* экспертами указано одна или несколько тем к которым этот документ относится. Среди всех встречающихся тем мы отобрали те, которые упоминаются в не менее чем 200 документам. Таким образом мы получили 104 тематические группы. Более подробная информация об используемом наборе тестовых данных представлена в таблице 1.

Отметим что один и тот же документ может входить сразу в несколько групп согласно экспертным оценкам. Соответствующее распределение изображено на рисунке 1. Тем самым многие из получившихся групп имеют большой процент общих документов. На рисунке 2 проиллюстрирована сильная пересекаемость составов коллекций. В частности для каждой из 15 тематик (из 140) существует "поглощающая" тематика, которая содержит не менее 70% входящих в эту тематику документов. Максимальный процент поглощения достигает 99%, например, для тематики посвященной военным действиям в Ираке и тематики посвященной военным действиям во

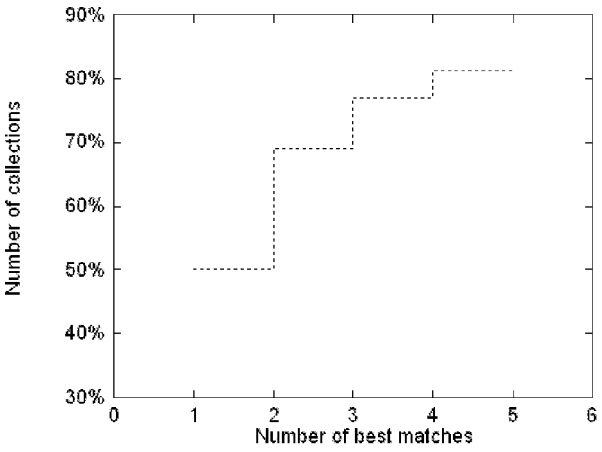

Рис. 3: Вероятность появления указанной экспертами тематик среди первых N выбранных автоматически лучших тематик

Класс (N)	Число кандидатов				Всего
	1	2	3	4	в классе
1	50.1%	69.0%	77.1%	81.3%	25704
2	—	29.4%	46.8%	57.5%	10648
3	—	—	21.0%	36.4%	8382
4	—	—	—	14.4%	7613

Table 2: Вероятность попадания N помеченных экспертами тематик в несколько лучших при автоматической классификации документов класса N.

всем мире.

В качестве группы задающих тематику документов, мы использовали 50 случайным образом выбранных документов из соответствующего тематического набора. На базе объединенного множества документов, использовавшихся для задания тематик, строилась функция семантической близости термов (как это описано в разделе 3.3).

Отметим, что в наших экспериментах для описания каждой тематики использовалось в среднем[5] 12% относящихся к ней документов из используемой экспериментальной базы, в отличие, например, от экспериментов описанных в работе [8], где для описания тематик использовалось более 50% доступных документов.

4.2 Эксперименты

В ходе проводимых экспериментов для каждого из 25181 рассматриваемых документов вычислены оценки его близости каждой из рассматриваемых тематик. По результатам этих оценок для каждого из документов был построен упорядоченный список тематик в порядке их оценок близости. Далее проводился анализ полученной информации.

[5]Процент документов, использовавшихся для задания тематики, изменялся от 7% до 25% в зависимости от общего числа доступных документов на данную тему.

Результаты практических экспериментов показали, что наилучшая[6] тематика, выбранная при помощи описанного подхода, попадает в множество указанных экспертами для данного документа ("искомых") тематик в более 50% случаев. А вероятность попадания указанной экспертами тематики в автоматически отобранную тройку лучших тематик превысила 75%. Для такого сложного набора тестовых данных эти результаты выглядят весьма прилично. Зависимость вероятности нахождения "искомой" тематики среди первых N отобранных лучших тематик проиллюстрирована на графике 3.

В используемом нами экспериментальном наборе данных для многих документов эксперты указали более одной тематики, к которой по их мнению относится этот документ. Естественно, что пользователи системы автоматической классификации заинтересованы в том чтобы система относила документ ко *всем* релевантным тематикам, а не только к какой-нибудь одной из них.

Для того, чтобы оценить насколько хорошо предложенный подход справляется с этой задачей мы разделили все множество используемых документов на классы по количеству указанных экспертами для данного документа тем. В первый класс вошли все доступные документы, во второй — только те документы для которых экспертами было указано не менее двух тематик, в третий — не менее трех, и т.д.

Для каждого класса k мы вычислили процент документов из этого класса для которых среди первых N лучших тематик встречалось не менее k указанных экспертами. Полученные результаты проиллюстрированы в таблице 2.

Как и следовало ожидать, для документов класса k вероятность угадывания k указанных тематик из k автоматически выбранных резко падает. Стоит отметить устойчивое значительное (более 40%) увеличение этой вероятности для случая "k из $k+1$".

4.3 Ошибки классификации

Для того, чтобы лучше понять как можно улучшить качество классификации мы исследовали некоторые из случаев в которых описываемый в данной работе метод дает сбои. Мы рассмотрим два основных класса ошибок классификации: *ложный выбор* и *промахи*. *Ложный выбор* — это ситуация, когда метод слишком высоко оценивает тематики, не указанные экспертами как релевантные для рассматриваемого документа. *Промахи* — это случаи, в которых указанные экспертами коллекции получали слишком низкие оценки.

4.3.1 Ложный выбор

Существует несколько разных причин по которым наилучшие оценки близости данному документу получают тематики, которые не были отмечены экспертами:

- **неточные описания тематик**

  Описание тематики может быть слишком "широким" и тогда документы из других тематик получают неоправданно высокие оценки релевантности.

  Качество описания тематики зависит от множества параметров — наборов документов используемых для описания *всех* тематик, алгоритма выбора слов

[6]Первая тематика в соответствующем этому документу упорядоченном списке.

для описания коллекции, и т.п. Построить идеальное начальное описание очень сложно, однако последовательное изменение описаний в процессе работы может помочь улучшить его качество.

- **поглощение тематик**

  Из-за того, что для многих тематик в нашей экспериментальной базе существовали "поглощающие" тематики (рис. 2), то соответствующие тематики зачастую портят друг другу результаты классификации.

- **неполнота экспертных данных**

  Эксперты TREC указывали одну или несколько подходящих тематик для каждого документа, но TREC не гарантирует, что ими были указаны все подходящие тематики. На самом деле это зачастую не так. Конечно эта проблема проявляется исключительно при экспериментах с TREC и не является актуальной для работы с реальными данными.

4.3.2 Промахи

Хотя в большинстве случаев указанные экспертами тематики получали довольно высокие оценки при выполнении классификации, но для некоторых документов происходили заметные промахи.

Основной причиной промахов при классификации является некоторое различие тематики рассматриваемого документа и документов из группы, которая использовалась для задания тематики. Так, например, документ `LA010190-0069`, относящийся к тематике финансовых судебных разбирательств, содержит только небольшое описание махинаций золотодобывающей компании и информацию о размере выплат инвесторам. Поскольку использовавшиеся при описании данной тематики документы были в основном посвящены махинациям с акциями, то построенное описание плохо соответствовало этому документу.

По-видимому часть подобных проблем может быть исправлена во время работы системы за счет постепенного уточнения описаний тематик и функции тематической близости термов.

5 Обсуждение

Хотя проведенные эксперименты продемонстрировали перспективность предлагаемого подхода, они выявили ряд вопросов, требующих отдельного обсуждения.

5.1 Вычислительная трудоемкость

Применимость метода классификации к реальным задачам сильно зависит от его производительности, которая определяется его вычислительной трудоемкостью.

Вычислительные ресурсы затрачиваемые при использовании данного метода классификации делятся на два класса — ресурсы, необходимые для единовременного проведения подготовительной работы, и ресурсы, необходимые для классификации отдельного документа.

Первая группа вычислительных затрат состоит из следующих компонент:

- Построение общего словаря W
- Построение описаний тематик
  - Построение "тематических" словарей
  - Выбор оптимального описания
- Построение функции близости термов
  - Построение матрицы термы на документы
  - Нахождение k наибольших сингулярных значений матрицы *термы-на-документы*

Отметим, что в общем случае вычисление сингулярных значений является очень трудоемкой задачей, однако сильная разреженность матрицы *термы-на-документы* позволяет использовать весьма эффективный алгоритм Ланкоша[7] [6].

Трудоемкость операции классификации одного документа складывается из затрат на:

- Вычисления оценки близости документа данной тематике (для каждой тематики)
  - Поиск требуемой информации про каждый терм из описания тематики и документа
  - Вычисление оценки близости двух термов
  - Вычисление общей оценки близости документа тематике
- Выбора наиболее близкой тематики

Общая трудоемкость классификации одного документа составляет порядка $O(|\Omega||D|_{avr}|C||W|k)$ операций, где $|\Omega|$ — общее число тематик, D_{avr} — среднее количество термов в документе, $|C|$ — среднее количество термов в описании тематики, $|W|$ — число термов в общем словаре, k — размерность пространства гипотез (число используемых сингулярных значений матрицы *термы-на-документы*).

Описанный нами подход требует значительных вычислительных ресурсов на подготовительном этапе, однако собственно классификация требует значительно меньших ресурсов. Так, во время проведения наших экспериментов, на компьютере с процессором PII–350 и 128 Mb оперативной памяти под управлением ОС Linux подготовительный этап занимал несколько часов машинного времени, а производительность системы достигала 32 документов в секунду. Такая скорость классификации показывает возможность применения разработанного метода для работы с потоковой информацией.

5.2 Настройка метода

В рамках описанного выше базового подхода для получения лучшего качества классификации в конкретных приложениях полезной оказывается дополнительная настройка метода.

- **Выбор функии оценки близости документа и тематики**

  Для вычисления общей оценки близости документа тематики мы опробовали несколько схем, из которых, в рамках наших экспериментов наилучшие результаты показала представленная в разделе 3.2 схема.

[7] В рамках наших экспериментов мы использовали реализацию алгоритма Ланкоша из распространяемого свободно пакета SVDPACK.

Однако при работе со слишком маленькими или слишком большими документами (содержащими менее 20 или более 1000 различных термов), эта схема дает сбои. В таких ситуациях лучше работает следующая оценка:

$$\text{Goodness}(C,d) = \frac{\sum_{q_i \in Q} \max_{c_j \in C} (\text{FSR}(q_i, c_j))}{|Q|} \quad (2)$$

Указанных документов достаточно мало и они не несут большого объема информации, поэтому, мы выкинули их из рассмотрения для упрощения эксперимента, однако сам этот факт говорит о проблеме поиска универсальной формы.

- **Выбор размера описания коллекции**

  Как уже отмечалось в п. 3.1.1 нам необходимо всего несколько термов для описания коллекции, однако оптимальное количество так и не выяснено в экспериментах мы использовали $n = 10$, и даже такие маленькие описания дали достаточно хороший результат. Нами рассматривались и другие схемы получения описаний, однако, учитывая специфику задачи, описанный метод оказался оптимальным.

- **Выбор документов для задания коллекции**

  Набор документов используемых для задания тематики коллекции в значительной степени определяет набор слов, которые будут использоваться в качестве описания данной коллекции, а также косвенно влияет на описания других коллекций.

  Кроме этого весь набор документов, используемых для задания тематик, также определяет общий словарь и задает функцию тематической близости.

  В общем случае довольно сложно собрать достаточный набор документов для получения хорошего начального описания для всех тематик. Однако в процессе работы возможно расширение наборов документов, описывающих тематики, для того чтобы уточнить описание тематик, и, как следствие, улучшить общее качество классификации.

5.3 Дальнейшее улучшение качества

Для дальнейшего улучшения качества классификации мы планируем исследовать ряд идей:

- **Многоуровневая классификация**

  Много трудностей при классификации вызвано тем фактом, что некоторые тематики значительно ближе друг к другу, чем в среднем. Как следствие, в общем тематическом пространстве описания таких тематик слишком похожи друг на друга, что ухудшает результаты классификации.

  Для решения этой проблемы мы предполагаем использовать многоуровневый подход:

  1. обнаруженные группы очень близких тематик объединяются в мегатематики
  2. производится классификация по полученному множеству мегатематик
  3. для каждой мегатематики производится дополнительная классификация попавших в нее документов

  Предварительные эксперименты показывают, что такой подход позволяет значительно повысить точность классификации на этапе классификации по мегатематикам.

  Однако, поскольку описываемый метод классификации требует использования нескольких тысяч документов для построения функции семантической близости термов, то применение его для классификации в рамках небольшой мегатематики затруднено. Использование же функции семантической близости термов построенной на этапе классификации мегатематик не дает должного эффекта в силу резкого сужения общего тематического поля.

  Возможно, что хорошие результаты покажет гибридный подход — использование для классификации в рамках мегатематики другого метода классификации.

- **Учет обратной связи**

  Перспективным методом улучшения качества классификации является учет комментариев пользователей системы для уточнения описаний тематик и функции тематической близости. Такой подход называется *механизмом обратной связи* (*relevance feedback*) и привлекает много внимания в научной литературе в течении нескольких лет [2, 3].

- **Лучшие методы построения описаний тематик**

  В результатах наших экспериментов выяснилось, что, при используемом нами способе вычисления общей оценки близости документа и тематики, наилучшие результаты получаются при относительно небольшом размере описания. И выбор этих нескольких термов оказывает значительное влияние на общее качество классификации.

  На данном этапе наших исследований мы использовали достаточно простой алгоритм для отбора термов в описание коллекции, и весьма вероятно, что полученные таким способом описания не являются оптимальными. В качестве теста на качество описаний можно использовать результаты тестовой классификации документов, используемых для задания тематик, по построенным описаниям.

6 Заключение

В этой работе рассматривается задача классификации множества документов по заданным тематикам, каждая из которых задается некоторым набором относящихся к данной тематике документов.

Предложеный метод классификации опирается на использование латентно-семантического анализа для выделения семантических зависимостей между термами.

Для экспериментальной проверки предлагаемого метода использовалась обширная экспериментальная база, построенная на основе стандартных наборов данных и экспертных оценок предоставляемых Text REtrieval Conference. Классификация более чем 25000 документов проводилась по 104 тематикам.

В отличие от большинства известных работ, используемые в наших экспериментах тематики зачастую являются довольно близкими и рассматриваемые документы могут относится сразу к нескольким из них. Несмотря на

это, результаты экспериментов показывают высокое качество классификации и подтверждают эффективность предложенного подхода.

При относительно высоких вычислительных затратах на подготовительном этапе предлагаемый метод не требует значительных ресурсов для самой классификации. Таким образом метод применим для классификации потоков информации.

Библиография

[1] Latent Semantic Indexing and TREC-2. In D.Harman, editor, *The Second Text REtrieval Conference*, 1994.

[2] J. Allan. Incremental relevance feedback. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval (SIGIR '96)*, pages 298–306, Apr. 1996.

[3] J. Allan. Learning while filtering documents. In *Proc. of SIGIR'98*, Melbourne, Australia, 1998.

[4] Andrei Z. Brooder, Steven C Glassman, Mark S. Manasse. Syntactic Clustering of the Web. In *Proc. of Sixth International World Wide Web Conference (WWW-6)*, 1996.

[5] L. D. Baker and A. K. McCallum. Distributional clustering of words for text classification. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Categorisation, pages 96–103, 1998.

[6] M. Berry. Large scale singular value computations. *International Journal of Supercomputer Applications*, 6(1):13–49, 1992.

[7] J. Cullum and R. Willougby. *Lanczos algorithms for large symmetric eigenvalue computations*, volume 1, chapter "Real rectangular matrix". Brikhauser, Boston, 1985.

[8] Daphen, Koller and Mehran, Sahami. Hierarchically classifying documents using very few words.

[9] S. Dumais. Improving the retrieval of information from external sources. 23:229–236, 1991.

[10] S. Dumais. Latent Semantic Indexing: TREC-3 Report. In *The Third Text REtrieval Conference*, 1995.

[11] Ellen M. Voorhees, Donna Harman. Overview of the Sixth Text REtrieval Conference (TREC-6). In *Text REtrieval Conference*, 1998.

[12] P. Foltz. Using Latent Semantic Indexing for information filtering. In R. Allen, editor, *ACM Conference on Office Information Systems (COIS)*, pages 40–47, Cambridge, 1990.

[13] F. Ilander, J. Palm, and E. Fahraus. The private filtering news agent. Feb. 1997.

[14] T. Joachims. A probabilistic analysis of the rochio algorithm with TFIDF for text categorization. In *Proc. of the International Conference on Machine Learning (ICML)*, 1997.

[15] T. Landauer, P. Foltz, and D. Laham. *Discourse Processes*, volume 25, chapter "An introduction to Latent Semantic Analysys", pages 259–284. 1998.

[16] D. Lewis and M. Ringuette. A comparison of two learning algorithms for text categorisation. In *Proc. of the Third Annual Symposium on Document Analysys and Information Retrieval*, pages 81–93, 1994.

[17] R. Papka and J. Allan. Document classification using multiword features. In G. Gardarin, J. French, N. Pissinou, K. Makki, and L. Bouganim, editors, *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM-98)*, pages 124–131, New York, Nov. 1998. ACM Press.

[18] Scott A. Weiss, Simon Kasif, Eric Brill. Text Classification in USENET Newsgroups: A Progress Report.

[19] Y. Yang and J. Pederson. Feature selection in statistical learning of text categorization. In *Proc. of the ICML'97*, pages 412–420, 1997.