УДК 519

РЕАЛИЗАЦИЯ ПОДХОДОВ К АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ДОКУМЕНТОВ ДЛЯ ОБЕСПЕЧЕНИЯ ИХ ТЕМАТИЧЕСКОГО ПОИСКА В БОЛЬШИХ МАССИВАХ ДАННЫХ

Лавренов А. О.

Научный руководитель - кандидат философских наук, доцент Олейников Б. В.

Актуальность задачи.

При работе с большим количеством данных остро встает проблема поиска и классификации информации. Во многих библиотечных и поисковых системах используется полнотекстовый поиск, поиск по автору, дате и т. п. Иногда этого бывает недостаточно, если мы, например, хотим видеть документы в классифицированных категориях по кодам УДК и осуществлять поиск по ним.

Такое представление документов могут предоставить некоторые электронные каталоги в виде УДК-навигаторов. Информация, к какой категории принадлежит документ, уже заранее известна и определена человеком. Но как быть если имеется большой массив данных с неизвестными кодами УДК? Необходимо построить некоторый классификатор и решить задачу классификации данных. Далее мы рассмотрим задачу классификации по кодам УДК. Все изложенные материалы ниже можно применять для классификации по любым другим категориям, не ограничиваясь кодами УДК.

Так же в связи с ростом информационных потоков обрабатываемых человеком и информационными системами всё чаще обсуждаются вопросы связанные с обработкой текста - его математическое представление, интерпретация и управление. Перед рассмотрением способов классификации, определим что есть анализ текста и как его осуществить.

Математическое представление текста.

Текст в векторной модели рассматривается как множество слов — термов, имеющих некоторый вес.

Разными способами можно определить вес слова в документе. Например можно посчитать частоту встречаемости слова в тексте. Чем чаще слово встречается, тем больше его «вес». Если терм не встречается в документе, то его вес равен нулю.

Все слова, которые встречаются в документах, то есть весь словарь русского языка, можно упорядочить некоторым образом, например по алфавиту. Теперь для каждого документа можно выписать весь набор весов слов соответственно словарю. Если некоторого терма нет в документе, то его вес будет равен нулю. То есть вектор будет иметь вид:

$$d_i = (w_1, w_2, \dots, w_n)$$

где d_i — векторное представление j-го документа, w_i — вес i-го терма в документе,

n — общее количество различных термов во всех документах коллекции, то есть размер всего словаря.

Имея такое представление текста мы может применить различные

математические операции над текстом. Например при нахождении «расстояния» между двумя векторами мы можем судить об их похожести. В данном случае за расстояние можно взять различные метрики — Евклидово расстояние, коэффициенты подобия и так далее.

Методы взвешивания термов.

Взвешивать термы в тексте можно различными способами, от которых зависит «качество» представления текста и может существенно повлиять на классификацию текста.

Для каждого терма в документе мы может определить некоторые числовые показатели:

1. частота встречаемости терма в документе или «tf»;
2. частота встречаемости терма в других документах или «df». Например если слово встречается в каждом четвёртом документа коллекции, то df=1/4;
3. длина слова;
4. показатель важности термов, с которыми используется данный терм. Например если некоторое существительное в тексте используется с прилагательными имеющими большой вес, можно так же говорить об его важности.

Из этих числовых показателей можно получить функции веса термов:

- Булево значение веса. $w = sign(tf)$
, то есть 1 — если слово встретилось в документе, 0 — иначе;
- $w = tf$ - стандартная частота слова;
- $w = tf / df$. Такой коэффициент часто называют «tf-idf», то есть произведение частоты слова (tf), на величину, обратную величине частоты встречаемости слова во всех документах коллекции (inverse df). Часто применяется «сглаженная» вариация этот формулы - $w = tf * \log_{10}(1/df)$.

При употреблении формул tf-idf решается проблема общеупотребительных слов - когда слова не несущие смысловой нагрузки имеют большой вес. Например для художественной литературы такими словами могут быть: говорить, думать, человек и так далее.

- другие функции из суперпозиций показателей слова.

Уменьшение размерности общего словаря.

Так как размер всего словаря очень велик — векторное представление документа будет состоять в основном из нулевых значений. Чтобы упростить работу по обработке данных следует по возможности сократить множество словаря. Можно воспользоваться следующими методами:

- удаление вспомогательных частей речи: предлоги, союзы, местоимения;
- удаление слов, которые встречаются слишком редко относительно всех документов. Например в 100000 документов слово встретилось только 1-2 раза;
- объедение слов в словосочетание. Можно использовать, если несколько слов встречаются вместе чаще, чем по отдельности.

Создание обучающей выборки.

Прежде чем приступить к этапу классификации документа необходимо

составить большую базу документов с известными классами классификации. Её можно использовать как для машинного обучения, так и для сбора различный статистических данных.

Обычно код УДК указывают в начале книги, поэтому проблема нахождения документов была решена двумя способами:

1. Для документов, которые уже представлены в сети Интернет в виде полного текста можно найти по запросу вида «УДК 517». Если данный запрос мы укажем в кавычках, то поисковые систему будут искать точное соответствие. Из всего найденного содержимого можно можно выделить файлы, которые, например, имеют расширение *.pdf. Далее из имеющегося pdf файла можно получить текст документа. Но количество таких книг в открытых источниках не велико, и часто это оказываются лишь справочники со ссылками, поэтому также лучше применить второй способ.
2. Для документов, которые не имеют полнотекстового представления, например отсканированные копии книг, можно сделать следующее: пользуясь любым УДК-навигатором электронных каталогов университетов можно получить список книг - их автора и название, соответствующих некоторому коду УДК. Зная название книги, можно попытаться найти копию книги в открытых источниках Интернета. Если такая книга найдена, её необходимо обработать любой OCR системой — системой распознавания символов. В результате можно получить «черновой» текст книги, без рисунков, формул и возможно с некоторыми неточностями определения текста. Такая копия текста книги не пригодна для прямого использования и изучения, но полезна для составления обучающей выборки, так как позволяет получить общую тенденцию употребления слов.

Анализ обучающей выборки, проверка гипотезы похожести документов.

Прежде чем приступить к классификации документов необходимо проанализировать обучающую выборку и проверить гипотезы похожести текстов.

Если два текста принадлежат одной классификационный категории, мы утверждаем что они похожие. Так же мы предполагаем, что векторы, представляющие тексты будут «близки», в смысле некоторой метрики. Основной вопрос: не окажется ли так, что по набору слов текстов нельзя судить об их похожести?

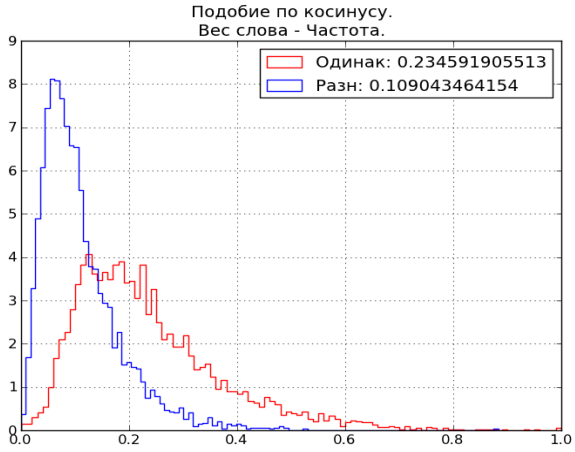
Для проверки гипотезы проведём следующий эксперимент. В качестве «расстояния» возьмём угол между двумя векторами. В данном случаем считаем что тексты тем похожей, чем меньше угол между их векторами. Для измерения угла будем использовать значение косинуса:

$$\cos(a)=\frac{\sum(w_i * q_i)}{(\sqrt{\sum(w_i^2)} * \sqrt{\sum(q_i^2)})}$$, где a - угол между векторами $\vec{w}$ и $\vec{q}$, w_i - значение i-того элемента вектора , - рначение i-того элемента вектора □.

Значение косинуса назовём «коэффициентом подобия». Получаем, чем ближе к единице коэффициент подобия, тем документы похожей. То есть гипотеза сводится к следующей: будет ли у двух случайных документов из одной категории значение коэффициента подобия больше, чем у двух документов из заведомо разных категорий?

Возьмём два документа из одной категории, посчитаем значение косинуса. Будем считать, что значение косинуса - некоторая случайная величина. Для большого набора случайных документов мы получим некоторую выборку из распределения значения косинуса. Так же нужно получить выборку для значения косинуса между векторами документов из разный категорий.

По двум выборкам можно построить гистограмму и посчитать некоторые статистические данные. Результат показан на рисунке:



Выборка для документов с одинаковыми кодами УДК.

Выборочное среднее: 0.234591905513

Выборочная дисперсия: 0.018290763754

Выборка для документов с разными кодами УДК.

Выборочное среднее: 0.109043464154

Выборочная дисперсия: 0.00548606929596

По этим данным можно сказать, что действительно, для векторов документов из одного классификационного класса в **среднем значение косинуса больше**, чем векторов документов из разных классов, то есть они являются «более похожими». Можно считать, что гипотеза является верной.

Классификация документов.

Будем считать что документ классифицирован правильно, если все три основных знака кода определены правильно.

Для классификации документов воспользуемся двумя алгоритмами классификации: «наивный баесовский алгоритм» и «алгоритм k-ближайших соседей». Возьмём тестовую выборку в размере 100 документов, которая в обязательном порядке не входит в состав обучающей выборки. Точность классификатора определяется следующим образом $t = V_r / V_{all}$, где V_r - количество правильно классифицированных документов, V_{all} - количество всех документов в тестовой выборке.

Для наивного баесовского классификатора нам понадобится некоторые «особенности» текста, возьмём 20 самых важных, то есть имеющий наибольший вес, термов. Число 20 получено экспериментальным путём, точность классификации при выборе только 20 слов максимальна.

Для алгоритма k-ближайших соседей, необходимо выбрать число k. Так же экспериментально получено, что при k=6, точность максимальна.

Результаты классификации представлены в таблице:

Название алгоритма	Параметры классификации	Точность
Наивный баесовский алгоритм	• размер обучающей выборки: ~ 12 000 документов; • размер тестовой выборки: ~ 100 документов; • количество «особенностей»	0.323232323232

	текста: 20	
Алгоритм k-ближайших соседей	• размер обучающей выборки: ~ 12 000 документов; • размер тестовой выборки: ~ 100 документов; • количество соседей текста: 6	0.30303030303

Результаты.

Точность классификации ~30% является низкой, но проблема классификации по кодам УДК в том, что количество кодов, по которым можно классифицировать текст очень велико. На данный момент используется примерно 450 основный кодов УДК. Решением данной проблемы является существенное увеличение тестовой выборки.

Так же можно классифицировать не по всем трём значениям кода, а, например, только по одному. При этом точность будет значительно выше. Но в данном случаем мы определим лишь общую тему документа.