УДК 519

АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ ТЕКСТОВЫХ ДОКУМЕНТОВ С ПОМОЩЬЮ АНАЛИЗА СИГНАЛОВ

Лавренов А.О.,
Научный руководитель - доктор физико-математических наук, профессор Садовский В.М.
Сибирский Федеральный Университет

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?

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.

Так же в докладе рассматривается новый способ анализа текстовых документов — перевод документа в сигнал.

Математическое представление текста.

Стандартный подход к моделирование теста — представление в виде вектор.

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.

Предполагается совершенно новый способ представление текста — представление его в виде сигнала. Возможный вариант такого представления — перевод текста в речь, полученную речь можно использовать как сигнал.

Коэффициент схожести (подобия)

Для векторной модели существует большое количество способов определения коэффициентов схожести (подобия), например, евклидово расстояние (представление текста в виде точки), косинус угла между векторами и так далее. На основе этих метрик строятся алгоритмы классификации. Чаще всего алгоритму не важно какая именно модель текстового документа используется. Достаточно найти коэффициент схожести или некоторое расстояние.

Для сигнальной модели текста можно использовать следующие подходы:

1. Коэффициент корреляции
2. Коэффициент кросскорреляции

Математический аппарат обработки сигналов очень развит. Можно попробовать применить различные подходы: сглаживание, избавление от шумов, вейвлет анализ, поиск дубликата и так далее.

Предполагаемые результаты

Уже был проведен грубый эксперимент, проверяющий работоспособность данного подхода.

Для обучающей выборки были взяты различные имена с известным половым соответствием. Алгоритму необходимо обучиться на тестовой выборки и классифицировать (определить пол) для тестовой выборки. За расстояние взято значение корреляции между функциями сигналов. Точность ~65%. По сравнению с классическими методами классификации (~95%) точность низкая, но такой точности уже достаточно для ответа на вопрос «Возможен ли такой подход на практике?».