

УДК 58+51-76

ЭЛЕКТРОННАЯ КЛАССИФИКАЦИЯ ТЕКСТОВ В ОБЛАСТИ БОТАНИКИ

Г.В. Андрианов, А.П. Ситников

Аннотация

На базе интегрированной среды разработки C++ Builder 6 Enterprise создан оригинальный автоматический классификатор текстов в области ботаники. Программа позволяет распределять оцифрованные тексты по классам, соответствующим УДК, вести поиск информации по тексту, фамилии автора и названию книги. Точность отнесения к определенному классу пока составляет чуть более 50%, но это значение планируется повысить за счет увеличения объема обучающей выборки.

Ключевые слова: автоматическая классификация, электронная библиотека, ботаника.

Введение

Существует два подхода к задаче автоматической классификации текстов [1]. Первый состоит в написании правил, по которым можно отнести текст к той или иной категории. Например, если текст содержит слова «интегумент» и «зародышевый мешок», то его следует отнести к классу «эмбриология растений». Специалист, знакомый с предметной областью и владеющий навыком написания регулярных выражений, может составить ряд правил, по которым будет определяться тема текстового документа. Однако создание и поддержание правил в актуальном состоянии требует постоянных усилий человека.

Другой подход основывается на машинном обучении. Набор правил для принятия решения текстового классификатора вычисляется автоматически из обучающих данных – некоторого количества документов из каждого класса. В машинном обучении сохраняется необходимость ручной разметки, но по сравнению с первым подходом здесь усилия специалиста направлены не на построение правил, а на составление алгоритма построения классификатора. Таким образом, в случае изменения набора категорий или применения системы в совершенно другой области, потребуется лишь вновь провести автоматическое обучение классификатора на новом обучающем множестве. При этом не требуется вмешательства ни экспертов в предметной области, ни в программировании.

Текстовые классификаторы, основанные на машинном обучении, можно разделять в зависимости от способа представления описаний классов, а также от организации процедуры классификации. В настоящее время практическое применение получили следующие группы.

1. Статистические классификаторы, на основе вероятностных методов. Наиболее известным в данной группе является семейство байесовых моделей [2].

Классы описываются как набор частот встречаемых слов, а процедура классификации характеризуется вычислением вероятностей встречаемых слов.

2. Классификаторы, использующие методы на основе искусственных нейронных сетей. Описание класса представляет собой многомерный вектор чисел, а процедура классификации характеризуется способом преобразования анализируемого текста к аналогичному вектору, видом функции активации нейронов, а также топологией сети.

Сейчас область автоматической классификации текстов активно развивается. Ведутся исследования, запускаются проекты и конкурсы на выявление лучших по точности алгоритмов. Тем не менее качество алгоритмов автоматической классификации оставляют желать лучшего. В международном проекте Large Scale Hierarchical Text classification Challenge, проводившемся в 2013 г., точность наилучшего алгоритма по доле правильно классифицированных документов составила около 73%.

Цель настоящей работы – создание и тестирование статистического классификатора, способного различать тексты по ботанической тематике.

1. Материалы и методы

1.1. Байесовая классификация. В основе наивной баесовой классификации лежит теорема Байеса, которая позволяет определить вероятность какого-либо события при условии, что произошло другое статистически взаимозависимое с ним событие:

$$P(c | d) = P(d | c)P(c)P(d), \quad (1)$$

где $P(c|d)$ – вероятность принадлежности документа d классу c , $P(d|c)$ – вероятность встречи документа d среди документов класса c , $P(c)$ – вероятность встретить документ класса c среди всех текстов, $P(d)$ – вероятность встречи документа d среди всех текстов.

В ходе такого подхода к классификации определяются зависимости между всеми переменными, исходя из предположения, что все они являются одинаково важными и статистически независимыми, то есть значение одной переменной ничего не говорит о значении другой [2].

Классификатор рассчитывает вероятность принадлежности текста по каждому классу и выбирает тот, который обладает максимальным значением вероятности. $P(d)$ является константой и никак не может повлиять на конечную оценку, поэтому ею можно пренебречь:

$$c = \arg \max_{c \in C} P(d | c)P(c). \quad (2)$$

Вероятность встречи документа класса c рассчитывается по следующей формуле:

$$P(c) = D_c / D, \quad (3)$$

где D_c – количество документов, принадлежащих классу c , D – общее количество документов в обучающей выборке.

Байесовый классификатор представляет документ как набор слов, поэтому вероятность документа вычисляется как произведение вероятностей всех слов входящих в документ:

$$P(d | c) = \prod_{i=1}^n P(w_i | c), \quad (4)$$

где $P(w_i | c)$ – вероятность встречи слова w в классе c среди всех текстов.

Если в ходе расчета вероятности встретится слово, которое отсутствовало на этапе обучения, то значение W_{ic} , а следовательно, и $P(w_i | c)$ будут равны нулю. Это приведет к тому, что документ с этим словом нельзя будет классифицировать, так как он будет иметь нулевую вероятность по всем классам, поэтому необходимо ввести сглаживание, то есть прибавить единицу в числителе. Этот подход смещает оценку вероятностей в сторону менее вероятных исходов и такая смещённая оценка никогда не бывает нулевой, что защищает от проблемы неизвестных слов:

$$P(w_i | c) = W_{ic} / (V + L_c), \quad (5)$$

где W_{ic} – встречаемость слова в документах класса c , V – количество уникальных слов всей совокупности документов, L_c – количество слов в документах класса c .

Подставляя (3) и (5) в формулу (2) и используя свойство произведения логарифма, получаем окончательную модель, по которой происходит классификация:

$$c = \arg \max_{c \in C} \left[\ln \frac{D_c}{D} + \sum_{i=1}^n \ln \frac{W_{ic} + 1}{V + L} \right]. \quad (6)$$

Для обучения классификатора необходима тренировочная выборка, в которой имеются соответствия между текстом и его классом. Далее из нее извлекается информация о количестве документов, принадлежащих определенному классу, количестве слов в документах каждого класса, количестве уникальных слов.

В качестве обучающей выборки были использованы 1463 фрагмента длиной около 100 строк из 95 книг, которые принадлежат следующим классам: альгология, эмбриология, фитоценология, микология, лишенология, экология растений, бриология, палеоботаника, анатомия, фитопатология, география растений, морфология.

1.2. Поиск информации. Поиск информации осуществляется при помощи поисковой машины программы, которая предназначена для поиска информации по запросу и ее ранжирования по значимости.

Поисковая машина анализирует текст из тренировочной выборки, строит инвертированный индекс. Он представляет собой список слов, с каждым из которых ассоциированы множество документов, где это слово встречается, с оценкой частоты вхождения слова в документ [2]. Это необходимо для того, чтобы определить местоположение документов, содержащих слова из запроса.

В ходе поиска алгоритм возвращает только те документы, в которых встречаются все слова, которые были в запросе. Далее поисковая машина ранжирует возвращаемые документы по частоте встречаемости слов.

1.3. Оценка качества алгоритма классификации. В простейшем случае оценкой качества работы алгоритма является величина *Accuracy*. Она представляет собой долю документов, по которым классификатор принял правильное решение:

$$Accuracy = P / N, \quad (9)$$

где P – количество документов, по которым классификатор принял правильное решение, N – размер обучающей выборки. *Accuracy* присваивает всем документам одинаковый вес, что может быть не совсем корректно, когда распределение документов в тренировочной выборке сильно смещено в сторону каких-либо классов [2].

F-мера ($F1$) представляет собой гармоническое среднее между долей документов, принадлежащих данному классу относительно всех документов и долей, найденных классификатором документов, принадлежащих классу относительно всех документов этого класса:

$$F1 = 2TP / (2TP + FN + FP), \quad (10)$$

где TP – истинно-положительное решение, FP – ложно-положительное решение, FN – ложно-отрицательное решение.

2. Результаты и их обсуждение

2.1. Структура базы данных. Тренировочная выборка и связанная с ней статистическая информация хранятся в базе данных. В качестве системы управления базой данных (СУБД) использована программа Microsoft SQL Server 2008 R2. Подключение к серверу проходило по технологии ActiveX Data Objects (ADO). Структура базы данных представлена в табл. 1.

2.2. Структура программы. Программа реализована на языке программирования C++. Его главным преимуществом является вычислительная производительность – язык дает максимальный контроль над всеми аспектами структуры и порядка исполнения программы, позволяя работать с памятью на низком уровне [3]. В качестве интегрированной среды разработки была использована C++ Builder 6 Enterprise. Она объединяет в себе комплекс объектных библиотек, компилятор, отладчик, редактор кода и содержит инструменты, которые при помощи технологии drag-and-drop делают разработку визуальной [4].

2.2.1. Блок ввода новых данных. Модуль предназначен для включения новых записей об авторах, книгах в базу данных. В поле «Количество авторов» (рис. 1, а) вводится целое число. После нажатия клавиши Enter динамически создаются поля для ввода данных об авторах. В поле «Название книги» записывается название книги, в поле «Название книги по ГОСТу» вводится название книги по ГОСТу (рис. 1, б). В выпадающем меню «Класс» (рис. 1, в) необходимо выбрать категорию, которой принадлежит книга. В поле «Путь» (рис. 1, г) прописывается путь местоположения книги. В поле «Текст» (рис. 1, д) вводится часть текста из книги.

Табл. 1

Структура базы данных тренировочной выборки

Объект	Функция	Атрибуты
 author	информация об авторе	id_author name_author
 book	информация о книге	id_book name_book gost_name path
author_has_book	связи между автором и книгой	id_author id_book
 document	текстовых фрагментах из книг	id_document id_book id_class text verify
 word	уникальных слова присутствующих в обучающей выборке	id_word word_name
 class	информация о классах	id_class udk class_name
count_words	связи между словами и текстовым фрагментом	id_document id_word count

 – первичный ключ

Выберите количество авторов и введите их ФИО

2

Дудка И.А.

Вассер С.П.

а)

Введите название книги

Грибы

Введите название книги по ГОСТу

Дудка И.А., Вассер С.П. Грибы. Киев: Наукова думка, 1987. 545 с

Выберите класс

Микология

Введите путь книги

C:\book\mycology\griby_dudka.djvu

б)

Введите текст

Грибы были объектом внимания человека с незапамятных времен. Однако многообразие грибов столь велико, что процесс их познания затянулся, до сих пор еще не завершён, и так же, как и прежде, их исследователей ждут многочисленные сюрпризы. Грибы до сих пор не открыли многие свои тайны, в частности, только недавно найден ответ на такой, казалось бы, простой вопрос, каково их место в системе органического мира. В связи с этим вполне уместно вспомнить слова французского ботаника А. Вейана, сказанные им еще в 1727 г.: "Грибы - это изобретение дьявола, придуманное им для того, чтобы нарушать гармонию остальной природы, смущать и приводить в отчаяние исследователей-ботаников".

Впечатляет не только огромное количество грибов в природе, среди которых преобладают не общеизвестные макроскопические, образующие видимые невооруженным глазом плодовые тела, а микроскопические, заметные лишь с помощью оптических приборов организмы. Поражает и многообразие ролей, которые эти организмы выполняют в природе и в хозяйственной деятельности человека. Грибы осуществляют преобразование мертвого органического вещества, выступая в качестве деструкторов, превращая сложные компоненты в исходные минеральные соли. Грибы (не только съедобные макромицеты, но и микромицеты) способны интенсивно синтезировать пищевой и кормовой белок, и в XX веке, который называют "эрой новых технологий", они рассматриваются как организмы-продуценты для разработки перспективных микробиологических методов получения белка. Многие грибы являются продуцентами других биологически активных веществ, в том числе антибиотиков, ферментов, витаминов и т. д. Но грибы играют не только положительную роль. Они могут явиться причиной болезней человека, животных, культурных растений, древесных пород, агентами биокоррозии различной промышленной и пищевой продукции и т. д.

Для того, чтобы ориентироваться во всем многообразии грибоного мира, четко представлять задачи, направления и аппарат науки о грибах - микологии, необходимы справочные пособия и брошюры по различным разделам микологии. К сожалению, до сих пор количество подобных публикаций в отечественной микологической литературе крайне ограничено. Круг вопросов, освещенных в немногочисленных зарубежных справочных работах, значительно уже, они не включают целый ряд микологических

в)

Рис. 1. Блок ввода новых данных: а) поле ввода авторов, б) поле ввода информации о книге, в) поле ввода фрагмента текста

2.2.2. Блок тренировки классификатора. Модуль используется для обработки текстовой информации, подсчета частоты слов и выявления наиболее характерных слов для класса (рис. 2).

В базе данных всего документов: 3
 Необученных документов: 3
 Обученных документов: 0
 Количество уникальных слов: 0
 ТРЕНИРОВКА УСПЕШНО
 ЗАВЕРШЕНА!
 Добавлено уникальных слов: 8
 Добавлено новых слов: 8
 Тренировать

Рис. 2. Блок тренировки классификатора

2.2.3. Блок классификации. Модуль используется для определения категории введенного текста. В поле «Введите текст для классификации» записывается анализируемый текст, в поле «Результат» при нажатии на кнопку «Выполнить» появляется результат классификации (рис. 3).

Введите текст для классификации	Результат
Воздушные гифы обычно тоньше, клетки их длиннее, чем у субстратных. Ветвление гиф мицелия в колонии у разных видов грибов может быть дихотомическое, симподиальное, моноподиальное (чаще), под разным углом по отношению к главной гифе и поверхности субстрата.	Просчет вероятностей: Для класса Фитопатология вероятность -- 39,4649391174316 Для класса Лихенология вероятность -- 7,62948989868164 Для класса Альгология вероятность -- 7,59915113449097 Для класса Микология вероятность -- 42,6303825378418 Наибольшая вероятность, что текст принадлежит классу Микология

Выполнить

Рис. 3. Блок классификации

2.2.4. Блок поиска информации. Осуществляет поиск информации по тексту, фамилии автора и названию книги. В поле «Введите фразу для поиска» вводится интересующий запрос. Если необходимо отфильтровать данные по автору или названию книги, то ставится флажок в поле «Поиск по автору» и «Поиск по названию книги» соответственно (рис. 4, а). После нажатия кнопки «Поиск» появляется ранжированный по релевантности список, который содержит идентификационный номер записи, к какой книге она относится и ее значимость. Для того чтобы открыть документ необходимо щелкнуть левой кнопкой мыши по интересующей записи (рис. 4, б).

Введите фразу для поиска

зародышевый мешок

☒ Поиск по тексту ☒ Поиск по автору ☐ Поиск по названию книги

Барыкина

Поиск

ID статьи	Название книги	Вес статьи
B60D99FE-EE7F-4DBE-A4EA-9F08B011A69D	Справочник по ботанической микротехнике. Основы и методы.	100

а)

Справочник по ботанической микротехнике. Основы и методы.

Р.П. Барыкина и др. Справочник по ботанической микротехнике. Основы и методы. М.: Изд-во МГУ, 2004. 312 с.

ИЗУЧЕНИЕ РАЗВИТИЯ СЕМЯЗАЧАТКА И СЕМЕНИ

Приготовление тотальных эмбриологических препаратов эндосперма по Петровой (1970)

Метод пригоден для изучения эндосперма на ранних стадиях его развития. Отдельные семязачатки, извлеченные из завязи, фиксируют в течение 24 ч по Навашину или в других фиксаторах, содержащих формалин. Промытые после фиксации семязачатки помещают в дистиллированную воду в чашку Петри.

Пользуясь бинокулярной лупой, в отраженном свете тонкими заточенными наподобие обоюдоострого скальпеля плоскими хирургическими или глазными иглами удаляют покровы семязачатка. При этом одной иглой придерживают семязачаток за халазальную часть, а второй подрезают и отворачивают оба интегумента одновременно. Осторожно извлекают эндосперм с зародышем, т.е. зародышевый мешок, и помещают их в бюкс с водой. Дальнейшая

б)

Рис. 4. Блок поиска информации: а) поле поисковой строки, б) результат поиска

2.3. Оценки, полученные в ходе работы классификатора. В ходе тестирования работы классификатора до внедрения алгоритма взаимной информации значение *Accuracy* было около 61%, *F1* – 58%. Наименее точное определение у классов «анатомия» и «морфология», «фитоценология» и «география растений». Это можно объяснить тем, что категории достаточно близки по тематике и алгоритм не может точно выявить наиболее характерные слова для классов.

Заключение

Программа, основанная на баесовой классификации, показала достаточно хороший результат при анализе ботанических текстов, точность составила около 51%. Однако осталась проблема с классификацией близких разделов ботаники, где текст имеет множество одинаковых терминов. Эту проблему можно решить, увеличив тренировочную выборку. В дальнейшем планируется применить данный подход и на другие разделы ботаники.

Литература

1. Manning C., Raghavan P., Schütze H. Introduction to Information Retrieval. – Cambridge: Cambridge Univ. Press, 2008. – 496 p.
2. Сегаран Т. Программируем коллективный разум. – СПб.: Символ-Плюс, 2008. – 368 с.

3. Холлингвэрт Д., Баттерфилд Д., Сворт Б. C++ Builder 5. Руководство разработчика. Т. 1: Основы. – М.: Вильямс, 2001. – 865 с.
4. Холлингвэрт Д., Баттерфилд Д., Сворт Б. C++ Builder 5. Руководство разработчика. Т. 2: Сложные вопросы программирования. – М.: Вильямс, 2001. – 817 с.

Поступила в редакцию
10.03.14

Андрянов Григорий Васильевич – студент кафедры ботаники, Казанский (Приволжский) федеральный университет, г. Казань, Россия.

E-mail: *esquelroche@hotmail.com*

Ситников Андрей Петрович – кандидат биологических наук, доцент кафедры ботаники, Казанский (Приволжский) федеральный университет, г. Казань, Россия.

E-mail: *andrew.sitnykov@kpfu.ru*

* * *

ELECTRONIC TEXT CLASSIFICATION IN BOTANY

G.V. Andrianov, A.P. Sitnikov

Abstract

Based on the integrated development environment C++ Builder 6 Enterprise an original automatic classifier of texts in the field of botany has been created. The program allows one to classify digitized texts according to UDC, and to perform information search by text, author's name and book title. Precision of the classification is still slightly over 50%, but this value is expected to improve by the increase of training sample volume.

Keywords: automatic classification, e-library, botany.

References

1. Manning C., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge, Cambridge Univ. Press, 2008. 496 p.
2. Segaran T. Developing for collective intelligence. St. Petersburg, Symbol-Plus, 2008. 368 p. (In Russian)
3. Holingvert D., Butterfield D., Swart B. C++ Builder 5. Developer's Guide. Vol. 1. Basics. Moscow, Williams, 2001. 865 p. (In Russian)
4. Holingvert D., Butterfield D., Swart B. C++ Builder 5. Developer's Guide. Vol. 2. Sophisticated programming issues. Moscow, Williams, 2001. 817 p. (In Russian)

Received
March 10, 2014

Andrianov Grigori Vasilevich – Student, Department of Botany, Kazan Federal University, Kazan, Russia.

E-mail: *esquelroche@hotmail.com*

Sitnikov Andrei Petrovich – PhD in Biology, Associate Professor, Department of Botany, Kazan Federal University, Kazan, Russia.

E-mail: *andrew.sitnykov@kpfu.ru*