СИНТЕЗ ТЕКСТОВЫХ СТРУКТУР НА ОСНОВЕ ИХ СИСТЕМООБРАЗУЮЩИХ ХАРАКТЕРИСТИК

Л.С.Ломакина, А.С. Суркова, Д.В. Жевнерчук, И.Д.Чернобаев
(г. Нижний Новгород, Нижегородский государственный технический университет им. Р.Е. Алексеева)
e-mail: llomakina@list.ru, ansurkova@yandex.ru, zhevnerchuk@yandex.ru, ichernobnn@gmail.com

TEXT STRUCTURES SYNTHESIS ON THE BASIS OF THEIR SYSTEM-FORMING CHARACTERISTICS

L.S. Lomakina, A.S. Surkova, D.V. Zhevnerchuk, I.D. Chernobaev
(Nizhny Novgorod, Nizhny Novgorod State Technical University n.a. R.E. Alekseev)

Abstract. In the article the information, parametric and non-parametric synthesis of text structures are described. Algorithmic and software for applied text processing (classification, clustering, identification) are shown.

Key words: text analysis, text data, parametric synthesis, non-parametric synthesis, information synthesis.

Введение. В связи с постоянным увеличением объемов текстовых данных, наблюдаемым в современных информационно-телекоммуникационных системах, возрастает актуальность создания эффективных методов обработки текстовой информации, в первую очередь, в задачах анализа и синтеза текстовых структур. В предложенной авторами [1] методологии формирования системообразующих характеристик, текстовые структуры рассматриваются как совокупности устойчивых связей признаков описания текста, обеспечивающих его целостность, сохранение основных свойств и возможность представления текста как объекта в многомерном пространстве признаков, что привело к совершенствованию методов кластеризации, классификации и идентификации и расширению сферы их практического применения. Рассмотрены параметрический синтез для известной структуры, непараметрический синтез, используемый при неизвестной структуре текстовых данных, а также информационный синтез, в котором для определения структуры и параметров текстов применяется информационная мера.

Параметрический синтез. Параметрический синтез можно охарактеризовать как процесс определения параметров элементов синтезируемого объекта при заранее известной структуре объекта. Методы параметрического синтеза основаны на предположении, что текстовые данные описываются распределениями из параметрических семейств, то есть в предположении о заданном законе распределения, и задача состоит в нахождении неизвестных параметров распределения [2].

К параметрическому синтезу текстовых структур относятся большинство методов и алгоритмов классификации данных, как процедуры обучения с учителем. Предполагается, что общая структура и критерии классификации заранее известны, и в процессе синтеза определяются конкретные параметры объектов каждого класса, и по выявленным данным производится классификация неизвестных текстов. Параметры объектов и структуры классов зависят от задач классификации и классификационных критериев: тематических, языковых, стилистических и т.п. К параметрическим методам можно отнести метод опорных векторов (SVM); метод fuzzy c-means (FCM); метод Kernel Fuzzy Clustering; метод ближайших соседей kNN; методы на основе решающих деревьев.

Непараметрический синтез. Непараметрический синтез - процесс определения структуры объекта и значений параметров составляющих ее элементов. При непараметрическом синтезе для восстановления неизвестной структуры в качестве обучающей выборки используют набор объектов с одинаковой структурой (по некоторым характеристикам). В процессе обучения выделяются параметры, общие для всех объектов, которые позволяют осуществить синтез структуры, необходимой для решения поставленной задачи.

Большинство алгоритмов непараметрического синтеза текстовых структур являются реляционными, основанными на определении взаимного отношения текстов и определении близости (различия) между объектами. Использование известных непараметрических методов, таких как метод знаков или рангов, а также методов, основанных на использовании понятия Колмогоровской сложности, позволяет оценить неизвестную скрытую структуру взаимосвязи объектов. К непараметрическим методам можно отнести методы кластеризации на основе нейронных сетей (GSOM); методы кластеризации на основе нечетких отношений и многие другие.

Информационный синтез. Информационный синтез не является противопоставлением параметрическому и непараметрическому синтезу алгоритмов. Информационные метрики (информационная мера) удобны в применении для выявления общности между структурами объектов с точки зрения количества информации. Использование информационного подхода позволяет упростить сложные вероятностные оценки, в частности, оценивать неопределенность одной величиной.

Алгоритмы параметрического и непараметрического синтеза могут включать элементы информационного синтеза, поскольку использование такой характеристики как количество информации удобно на практике. Алгоритмы информационного синтеза позволяют наиболее полно обрабатывать всю имеющуюся информацию для решения поставленных задач. Для текстовых данных информационная составляющая имеет большое значение, поэтому алгоритмы информационного синтеза были выделены в отдельную группу, при этом методы, основанные на использования информации, могут быть применены для решения любых задач (классификации и кластеризации). Однако именно информационная мера позволяет строить эффективные алгоритмы решения задач идентификации.

Информационный синтез может рассматриваться как обобщающая процедура параметрических и непараметрических методов и в большей степени применяться для решения задач идентификации. Действительно, если определять задачу идентификации, как задачу поиска значимых параметров для исходных данных (текстов) и определения конкретных значений выбранных признаков, то задача процедуры идентификации обладают чертами как кластеризации (определение признаков), так и классификации (вычисление значений признаков). К параметрическим методам идентификации относятся методы анализа текстов на основе информационного подхода; методы идентификации на основе нейронных сетей; методы анализа текстов на основе энтропийных характеристик.

Алгоритмическое и программное обеспечение синтеза текстовых структур. На основе методов параметрического синтеза были разработаны алгоритмы классификации документов по тематическим категориям с использованием предложенной модели текстовых структур в виде спектров N-грамм [3]. В результате было показано увеличение эффективности классификации по сравнению с известными классификаторами.

Методы параметрического синтеза применялись также при создании системы классификации текстовых потоков на основе байесовского подхода [4]. Предложенная модификация наивного байесовского классификатора для текстовых потоков новостных сообщений в реальном времени, использующая метрику tf-idf в качестве оценки вероятности принадлежности терминов классам, показала эффективность выше, чем традиционные классификаторы.

Методы непараметрического синтеза применялись при разработке алгоритмов и программного обеспечения иерархической кластеризации текстов художественных произведений на основе понятия Колмогоровской сложности и расчета близости текстов с использованием алгоритмов сжатия [5]. Предложенная система иерархической кластеризации текстов [6] может быть использована для анализа текстов разного объема большого числа авторов.

Методы непараметрического синтеза также были применены для решения задачи классификации пользователей, входящих в социальное сообщество, по их текстовым сообщениям [7]. Было показано, что при работе с интернет текстам следует основываться на ме-

тодах нечеткой кластеризации и учитывать такие параметры, характеризующие авторов текстовых сообщений, как авторитетность и число последователей.

Информационная мера использовалась при создании алгоритмов идентификации автора художественных текстов, основанных на сравнении их информационных портретов [8]. Алгоритмы информационного синтеза использовались при создании системы идентификации авторства исходных кодов программ [9] с применением энтропийных характеристик символьного разнообразия, рассчитываемых для различных элементов текста.

Заключение. В работе рассмотрены особенности информационного, параметрического (классического) и непараметрического синтеза текстовых структур в основных задачах кластеризации, классификации и идентификации текстов. При практическом применении методов синтеза текстовых структур в прикладных задачах обработки текстов представляется перспективным создание открытой информационной системы анализа текстов, позволяющей реализовать комплекс алгоритмов синтеза текстовых структур для решения задач обработки текстов разных типов (художественных, научных, интернет текстов, текстов программ).

ЛИТЕРАТУРА

1. Ломакин Д.В., Ломакина М.Д., Суркова А.С. Методология формирования системоорганизующих характеристик текстовых данных // Фундаментальные исследования. 2015. № 11 (3). с. 480-483.

2. Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды, М., Наука, 1976. 736 с.

3. Ломакина Л.С., Мордвинов А.В., Суркова А.С. Построение и исследование модели текста для его классификации по предметным категориям. //Системы управления и информационные технологии, 2011, №1(43), с. 16-20.

4. Ломакин Д.В., Ломакина Л.С., Субботин А.Н. Программа классификации потоков текстовых данных на основе байесовского подхода // Свидетельство о государственной регистрации программы для ЭВМ №2017611236 от 31.10.2016.

5. Lomakina L.S., Rodionov V.B., Surkova A.S. Hierarchical Clustering of Text Documents // Automation and Remote Control, 2014. – Vol. 75. – No. 7. – pp. 1309-1315.

6. Ломакина Л.С., Родионов В.Б., Суркова А.С. Система иерархической кластеризации текстов // Свидетельство о государственной регистрации программы для ЭВМ №2013611004 от 09.01.2013.

7. Ломакина Л.С., Суркова А.С. Прикладные аспекты концептуального анализа и моделирования текстовых структур // Фундаментальные исследования. 2015. № 7 (3). с. 540-544.

8. Суркова А.С. Идентификация текстов на основе информационных портретов // Вестник Нижегородского университета им. Н.И. Лобачевского, 2014, № 3 (1), с. 145–149.

9. Ломакина Л.С., Семенцов М.С., Суркова А.С. Система идентификации авторства исходных кодов программ // Свидетельство о государственной регистрации программы для ЭВМ №2016612838 от 12.01.2016.