

УДК 004.912

МЕТОДОЛОГИЧЕСКИЕ АСПЕКТЫ КОНЦЕПТУАЛЬНОГО АНАЛИЗА И МОДЕЛИРОВАНИЯ ТЕКСТОВЫХ СТРУКТУР

Ломакина Л.С., Суркова А.С.

*Нижегородский государственный технический университет им. Р.Е. Алексеева,
Нижний Новгород, e-mail: ansurkova@yandex.ru, llomakina@list.ru*

Представлена методология решения задач анализа текстов – задачи кластеризации, классификации и идентификации. Сформулированы основные принципы анализа и обработки текстовых данных. Принцип системного представления текстов позволяет учитывать структуру и внутренние закономерности текстов. Принцип нечеткой логики предполагает использование нечетких моделей при представлении текстовых данных с применением качественных, неточных или неопределенных признаков. Принцип обучающихся систем предполагает построение систем, способных работать при недостатке априорной информации и улучшающих свое функционирование в процессе работы. Рассмотренные принципы позволяют представить совокупность обобщенных методов и алгоритмов решения основных задач обработки текстов. Представлена система выбора моделей текстов и методов решения прикладных задач в зависимости от целей анализа и характеристик текстов. Предложенная методология позволяет обозначить прикладные аспекты анализа и моделирования текстов в задачах идентификации авторства, в информационно-поисковых системах и в системах, связанных с обеспечением кибербезопасности.

Ключевые слова: анализ и обработка текстов, модели текстов, классификация, кластеризация, идентификация

METHODOLOGICAL ASPECTS OF CONCEPTUAL ANALYSIS AND MODELING OF TEXT STRUCTURES

Lomakina L.S., Surkova A.S.

*R.E. Alekseev Nizhny Novgorod State Technical University, Nizhny Novgorod,
e-mail: ansurkova@yandex.ru, llomakina@list.ru*

The methodology of the solution of text analysis tasks is presented which consists of clustering, classification and identification. The basic principles of text data analysis and processing are formulated. The principle of the system text representation allows to take into account the internal structure and regularities of texts. The fuzzy logic principle involves the use of fuzzy models for representing text data using qualitative, imprecise or uncertain signs. The learning systems principle involves the construction of systems which can operate with a lack of a priori information and improve its functioning in the process. The principles allow to represent a set of generic methods and algorithms for solving the main text processing tasks. The system selection of text models and methods of solving applied problems are presented that depend on the analysis goals and texts characteristics. The proposed methodology allows to identify the applied aspects of texts analysis and modeling for problems of authorship identification, information retrieval systems and cybersecurity systems.

Keywords: analysis and processing of texts, texts models, classification, clustering, identification

Важность и значимость анализа и обработки текстовых и других слабоструктурированных данных постоянно возрастает в связи с развитием глобальных информационных сетей. Все теоретические исследования последнего времени в области интеллектуального анализа текстов в той или иной степени опираются на системный подход к естественному языку. На основе системного подхода к моделированию и представлению текстовых данных были сформулированы основные теоретические положения построения систем анализа и обработки текстовых структур [6]. К таким основам относятся системное и потоковое представление текстов, теория нечеткости, теория обучающихся систем.

Как и любая модель, модели текста являются его упрощением и отражением только некоторых внутренних свойств. **Принцип**

системного рассмотрения текста предполагает выделение множества различных отдельных характеристик или их совокупности, построение на их основе нескольких моделей текста, описывающих текст с разной степенью детализации с различных точек зрения.

Принцип нечеткой логики предполагает взаимное дополнение и пересечение построенных моделей, возможность одновременного использования нескольких из них.

Построенные системы должны соответствовать **принципу обучающихся систем**, согласно которому система улучшает свое функционирование на основе полученной в результате работы информации и тем самым предусматривать возможность работы в условиях неполной исходной информации. В большей степени это касается задач идентификации, поскольку при идентификации, по определению, исходят

из неполной информации, и требуется по имеющимся данным выявить некоторые идентификационные признаки, по которым в дальнейшем может быть проведена идентификация объектов.

При построении систем анализа и обработки текстов возникают задачи, которые можно разделить на три большие группы: задачи **кластеризации**, то есть разбиение корпуса текстовых данных на отдельные кластеры (группы, классы); задачи **классификации** – отнесение неизвестного текста к одному из заданных классов и задачи **идентификации** – определение значимых признаков, структур и основных параметров текстовых данных. Такое разделение на задачи и принципы носит условный характер для большей наглядности и простоты изложения. Традиционно задачи кластеризации и классификации рассматривают в единой системе, кластеризацию данных также называют автоматической классификацией, поскольку процедура предполагает отнесение каждого объекта к конкретному классу. Методы решения задач во многом совпадают или требуют незначительной модификации, при рассмотрении задач другого типа.

В последнее время в связи с развитием вычислительной техники, увеличением ее быстродействия, появлением суперкомпьютеров, а также возникновением новых подходов к процессам обработки данных (распараллеливание задач и другие), на первый план выходит не проблема расчетов, а проблема моделирования. Причем развитие подхода к моделированию состоит не

просто в усложнении моделей, а в использовании иных приемов. Модель может не быть сложной, но она должна учитывать все особенности объектов с точки зрения рассматриваемых задач. При разработке и проектировании систем анализа и обработки текстовых данных необходимо предусмотреть следующие возможности:

1. Построение совокупности моделей текстовых данных, характеризующих различные параметры текстов и отражающие особенности текстов на разных уровнях иерархической системы.

2. Реализация нескольких алгоритмических подходов для решения задач классификации, кластеризации и идентификации с возможностью их и модификаций с точки зрения быстродействия реализации и представления получаемых результатов.

3. Методика выбора из всей совокупности используемых моделей и алгоритмов необходимого подмножества, наилучшим образом отвечающего поставленным задачам.

4. Модификация отобранных моделей и алгоритмов в зависимости от целей конкретных задач и дополнительных условий (точности, времени выполнения и т.п.).

Подобные системы могут быть отнесены к системам семиотического типа [2], которые можно представить набором моделей, правилами выбора моделей и функциями модификаций отдельных моделей и системы в целом.

Методологию анализа и моделирования текстов, ее основные задачи и принципы можно представить в единой таблице, показанной на рисунке.

Методология анализа и моделирования текстов		Задачи анализа и обработки текстов		
		I Кластеризация	II Классификация	III Идентификация
Основные принципы	A. Принцип системного представления текстов	A.I	A.II	A.III
	B. Принцип нечеткой логики	B.I	B.II	B.III
	C. Принцип обучающихся систем	C.I	C.II	C.III

Обобщенная схема методологии анализа и моделирования текстов

I. Задачи кластеризации

А.И. Для решения задачи кластеризации на основе принципа системного представления текстов целесообразно использовать структурно ориентированные **алгоритмы сжатия** и понятие **Колмогоровской сложности** [1, 5].

В.И. Методы кластеризации на основе нечетких отношений не требуют предварительного выявления признаков текстов, при их применении может быть использован подход на основе алгоритмов сжатия. Для этого необходимо составить **матрицу отношений** между текстами, элементами которой могут служить степени сжатия объединенных текстов. Степень сжатия определяется с применением выбранного заранее контекстно зависимого компрессора [10].

Для решения задачи нечеткой кластеризации предложено большое количество методов, таких как Метод fuzzy c-means (FCM), *Kernel Fuzzy Clustering*, Метод Густафсона-Кесселя и другие [9]. Большинство этих методов предполагает векторное представление данных, поэтому для их применения необходима предварительная обработка рассматриваемых текстов и формирование признакового пространства.

С.И. Задачи кластеризации по своей сути предполагают **обучение без учителя**, решение задачи кластеризации на основе принципа обучающимся систем позволяет обратиться к самообучающимся методам нейросетевых технологий, например, могут быть применены самоорганизующиеся **карты Кохонена** [11].

II. Задачи классификации

А.П. Для решения задачи классификации по предметным областям (задачи категоризации) успешно применяется системное представление текстов, в первую очередь, векторное представление и использование *N*-грамм. При этом повышение эффективности различных методов классификации может быть достигнуто за счет применения расширенных моделей текста, лучше отражающих его глубинную структуру, например, таких как **деревья и спектры N-грамм** [4].

В.П. Существует большое число методов классификации, таких как метод опорных векторов, ближайших соседей, которые могут быть обобщены и адаптированы для случая нечеткого представления. Например, для решения задач нечеткой классификации текстов предложено использовать **модификацию метода *k* ближайших соседей**, которая предполагает вычисление степени принадлежности как отношения проголосовавших соседей за

определенный класс к общему числу голосующих [12].

С.П. По многим аспектам задачи идентификации тесно связаны с задачами классификации: отнесение по некоторым данным неизвестного объекта к тому или иному классу. При этом, в зависимости от выбранного идентификационного признака, объект может быть отнесен к нескольким классам – налицо задача нечеткой классификации. Можно утверждать, что задача идентификации в смысле отнесения к группе объектов по сходству признаков является задачей нечеткой классификации.

III. Задачи идентификации

А.Ш. При решении задач идентификации могут быть использованы разнообразные модели текстовых структур, такие как взаимная информация, **Марковская модель текста**, модели, основанные на *N*-граммах, энтропийные характеристики вычисления **символьного разнообразия** [8]. Выбор модели или их совокупности зависит от целей идентификации, особенностей идентифицируемых текстов и т.п. [7].

В.Ш. и С.Ш. В задачах классификации и идентификации обычно предполагается наличие обучающих выборок, на основе которых обучается классификатор или выявляются идентификационные признаки. Использование одновременно нескольких подходов к обучению позволяет в лучшей мере определить значимые особенности данных, а также решать **задачи при неполной или недостоверной исходной информации**.

В таблице приведены некоторые базовые методы, использующиеся при решении основных задач обработки текстовых данных, на основе рассмотренных принципов.

Однако приведенное обобщение (рисунк) нельзя рассматривать как полное и окончательное. В дальнейшем таблица может быть расширена введением других принципов и рассмотрением новых задач. Так во многих задачах может оказаться перспективным использование знаний о предметной области, то есть применение различных онтологий и тезаурусов [3]. Другим направлением может являться развитие методов решения задач автоматического аннотирования и реферирования, которые по многим параметрам близки к задаче идентификации как к задаче выявления значимых признаков, но большая специфичность используемых приемов и техник позволяет рассматривать аннотирование и реферирование как отдельный тип задач и, соответственно, рассмотреть особенности ее решения на основе указанных принципов.

Совокупность методов анализа и обработки текстовых данных

Задачи. Принципы	I Кластеризация	II Классификация	III Идентификация
А. Принцип системного представления текстов	AI Сжатие для представления текстов. Методы, основанные на Колмогоровской сложности	АII Векторное представление, <i>N</i> -граммы, деревья и спектры <i>N</i> -грамм. Потоковое представление	АIII Взаимная информация, информационные портреты. Энтропийные характеристики символического разнообразия
В. Принцип нечеткой логики	BI Метод <i>fuzzy c-means (FCM)</i> . Метод <i>Kernel Fuzzy Clustering</i> . Метод Густафсона-Кесселя Кластеризация на основе нечетких отношений	ВII Нечеткие модификации метода к ближайших соседей (<i>k-nearest neighbor, k-NN</i>)	ВIII Нечеткие модификации метода к ближайших соседей (<i>k-nearest neighbor, k-NN</i>)
С. Принцип обучающихся систем	CI Нейросетевые технологии. Технологии на основе решающих деревьев	СИ Метод опорных векторов. Нейросетевые технологии	СИИ Метод опорных векторов. Нейросетевые технологии. Взаимная информация

Предложенные методологические аспекты анализа и моделирования текстов позволяют обобщить основные принципы для решения задач анализа текстов и формализовать выбор моделей и методов обработки текстов при решении конкретных прикладных задач. Рассмотренная методология позволяет обозначить прикладные аспекты анализа и моделирования текстовых структур, а именно, в задачах нечеткой идентификации потоковых текстовых сообщений, информационно-поисковых системах и в системах, связанных с обеспечением информационной безопасности.

Список литературы

1. Верещагин Н.К., Успенский В.А., Шень А. Колмогоровская сложность и алгоритмическая случайность. – М.: МЦНМО, 2013.
2. Еремеев А.П. Построение решающих функций на базе тернарной логики в системах принятия решений в условиях неопределенности // Изв. РАН. Теория и системы управления. 1997. – № 5. – С. 138–143.
3. Кучуганов В.Н. Элементы теории ассоциативной семантики // Управление большими системами. Выпуск 40. – 2012. – С. 30–48.
4. 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.
5. Ломакина Л.С., Родионов В.Б., Суркова А.С. Иерархическая кластеризация текстовых документов. // Системы управления и информационные технологии. – 2012. – № 2(48). – С. 39–44.
6. Ломакина Л.С., Суркова А.С. Теоретические аспекты концептуального анализа и моделирования текстовых структур // Фундаментальные исследования. – 2015. – № 2 (часть 17). – С. 3713–3717.

7. Семенцов М.С., Суркова А.С. Энтропийные характеристики символического разнообразия в текстах исходных кодов программ // Системы управления и информационные технологии. – 2015. – № 1.1(59). – С. 173–176.

8. Сметанин Ю.Г., Ульянов М.В. Мера символического разнообразия: подход комбинаторики слов к определению обобщенных характеристик временных рядов // Бизнес-информатика. – 2014. – № 3 (29). – С. 40–48.

9. Advances in Fuzzy Clustering and its Applications. Editor(s): J. Valente de Oliveira, W. Pedrycz. John Wiley & Sons, Ltd. – 2007. – 434 p.

10. Lomakina L.S., Rodionov V.B., Surkova A.S. Hierarchical Clustering of Text Documents // Automation and Remote Control. – 2014. – Vol. 75, № 7. – P. 1309–1316.

11. Surkova A.S., Domnin A.A., Bulatov I.V., Tsarev A.A. Neural networks and decision trees algorithms – the base of automated text classification and clustering // American Journal of Control Systems and Information Technology. Science Book Publishing House, LLC. – 2013. – № 2. – P. 33–35.

12. Surkova A.S., Skorynin S.S. Modified classification algorithm with fuzzy interpretation of clusters // American Journal of Control Systems and Information Technology. – 2014. – Vol. 4, № 2. – P. 27–30.

References

1. Vereshagin N.K., Uspenskiy V.A., Shen A. Kolmogorovskaya slozhnost i algoritmicheskaya sluchajnost. M.: MC-NMO, 2013.
2. Eremeev A.P. Postroenie reshajushhih funkciy na baze ternarnoj logiki v sistemah prinjatija reshenij v uslovijah neopredelennosti // Izv. RAN. Teorija i sistemy upravlenija. 1997. no. 5. pp. 138–143.
3. Kuchuganov V.N. Jelementy teorii associativnoj semantiki // Upravlenie bolshimi sistemami. Vypusk 40. 2012. pp. 30–48.
4. Lomakina L.S., Mordvinov A.V., Surkova A.S. Postroenie i issledovanie modeli teksta dlja ego klassifikacii po predmetnym kategorijam // Sistemy upravlenija i informacionnye tehnologii. 2011. no. 1(43). pp. 16–20.
5. Lomakina L.S., Rodionov V.B., Surkova A.S. Ierarchicheskaja klasterizacija tekstovyh dokumentov. // Sistemy

upravlenija i informacionnye tehnologii. 2012. no. 2(48). pp. 39–44.

6. Lomakina L.S., Surkova A.S. Teoreticheskie aspekty konceptualnogo analiza i modelirovanija tekstovyh struktur // Fundamentalnye issledovanija. 2015. no. 2 (chast 17). pp. 3713–3717.

7. Semencov M.S., Surkova A.S. Jentropijnye harakteristiki simvolnogo raznoobrazija v tekstah ishodnyh kodov programm // Sistemy upravlenija i informacionnye tehnologii. 2015. no. 1.1(59). pp. 173–176.

8. Smetanin Ju.G., Uljanov M.V. Mera simvolnogo raznoobrazija: podhod kombinatoriki slov k opredeleniju obobshhenykh harakteristik vremennyh rjadov // Biznes-informatika. 2014. no. 3 (29). pp. 40–48.

9. Advances in Fuzzy Clustering and its Applications. Editor(s): J. Valente de Oliveira, W. Pedrycz. John Wiley & Sons, Ltd. 2007. 434 p.

10. Lomakina L.S., Rodionov V.B., Surkova A.S. Hierarchical Clustering of Text Documents // Automation and Remote Control. 2014. Vol. 75, no. 7. pp. 1309–1316.

11. Surkova A.S., Domnin A.A., Bulatov I.V., Tsarev A.A. Neural networks and decision trees algorithms the base of au-

tomated text classification and clustering // American Journal of Control Systems and Information Technology. Science Book Publishing House, LLC. 2013. no. 2. pp. 33–35.

12. Surkova A.S., Skorynin S.S. Modified classification algorithm with fuzzy interpretation of clusters // American Journal of Control Systems and Information Technology. 2014. Vol. 4, no. 2. pp. 27–30.

Рецензенты:

Баландин Д.В., д.ф.-м.н., профессор, заведующий кафедрой численного и функционального анализа, Нижегородский государственный университет им. Н.И. Лобачевского, г. Нижний Новгород;

Федосенко Ю.С., д.т.н., профессор, заведующий кафедрой «Информатики, систем управления и телекоммуникаций», Волжский государственный университет водного транспорта, г. Нижний Новгород.