УДК 004.912

ТЕОРЕТИЧЕСКИЕ АСПЕКТЫ КОНЦЕПТУАЛЬНОГО АНАЛИЗА И МОДЕЛИРОВАНИЯ ТЕКСТОВЫХ СТРУКТУР

Ломакина Л.С., Суркова А.С.

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

Рассмотрены основные теоретические подходы, которые необходимы при построении эффективных систем анализа и обработки текстовых структур. К таким аспектам относится системное представление текста, в котором структура текста рассматривается с единых системных позиций и выделяется на нескольких структурных уровней. Принцип системного представления текстов определяет формализацию и представление глубинных свойств текста в явной форме, отражающей внутренние закономерности текстов. Потоковое представление предполагает рассмотрение текстов как непрерывный поток информации, среди которой могут быть выделены отдельные структурные элементы. Теория сжатия использует методы компрессии данных, основанные на выявлении внутренних закономерностей и структуры сжимаемых объектов (текстов). Теория нечеткости предполагает использование нечетких моделей при представлении текстовых данных с применением качественных, неточных или неопределенных признаков. На основе рассмотренных теоретических аспектов сформулированы требования к моделям текстовых структур в системах их обработки.

Ключевые слова: анализ и обработка текстов, сжатие, Колмогоровская сложность, нечеткость

THEORETICAL 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

We have considered the main theoretical approaches which are needed for building effective systems of analysis and processing of text structures. These aspects include a systematic text presentation in which text structure is treated with a single system point of view and several structural levels are discriminated. The principle of systematic text presentation determines the formalization and representation of the text underlying properties in an explicit form, this form reflects internal regularities of the texts. The stream presentation implies the consideration of the texts as a continuous information flow with structural elements inside. The theory of compression uses the methods of data compression, based on the identification of the internal regularities and the structure of compressible objects (texts). Fuzzy theory involves the use of fuzzy models in the presentation of text data using qualitative, imprecise or uncertain features. The requirements to models of text structures in their processing systems are formulated based on the theoretical aspects.

Keywords: text analysis and processing, compression, Kolmogorov complexity, fuzziness

Важность и значимость анализа и обработки текстовых и других слабоструктурированных данных постоянно возрастает. В связи с широким распространением систем электронного документооборота, социальных сетей, сетевых информационных порталов, персональных сайтов это становится особенно важным и как техническая задача, и как значимая часть взаимодействия людей в современном информационном мире. При работе с большими объемами текстовой информации постоянно возникают новые проблемы, которые требуют своего решения. В настоящее время ведутся активные исследования в данной области, однако существует ряд нерешённых проблем, связанных с созданием общесистемных подходов к представлению текстов.

Все исследования последнего времени в области интеллектуального анализа текстов в той или иной степени опираются на системный подход к естественному языку. Все более распространяющаяся тенденция – рассматривать как некую системную целостность текст полностью или даже корпус текстов. Применение системного подхода оправдано, поскольку язык обладает всеми свойствами и характеристиками, присущими сложным системам. Однако он обладает и своими особенностями. Систему языка можно определить как некоторое основополагающее свойство языка, обусловленное его сложным составом и сложными функциональными задачами.

Можно выделить следующие основные свойства и фундаментальные качества естественного языка [3]:

– принципиальная нечеткость значения языковых выражений;

– динамичность языковой системы;

– образность номинации, основанная на метафоричности;

– семантическая мощь словаря, позволяющая выражать любую информацию с помощью конечного инвентаря элементов;

– гибкость в передаче эксплицитной и имплицитной информации;

– разнообразие функций, включающее коммуникативную, когнитивную, планирующую, управляющую, обучающую, эстетическую и другие;

– специфическая системность и разделение языка на уровни и подсистемы.

Текст является одним из наиболее четких и значимых выражений естественного языка, поэтому на основании перечисленных свойств и качеств языка можно определить основные теоретические положения, опираясь на которые можно строить системы анализа и обработки текстовых структур. К таким основам относятся системное и потоковое представление текстов, теория сжатия и теория нечеткости.

Системное представление текстовых структур

Системное представление текстов предполагает формализацию и представление в явной структуре глубинных свойств и характеристик текста и построение его моделей. Достаточно очевидной является иерархическая организация текста, схема которой представлена на рис. 1, причем на каждом уровне иерархии текст структурирован не только под влиянием законов этого уровня, но и законами вышележащих уровней.

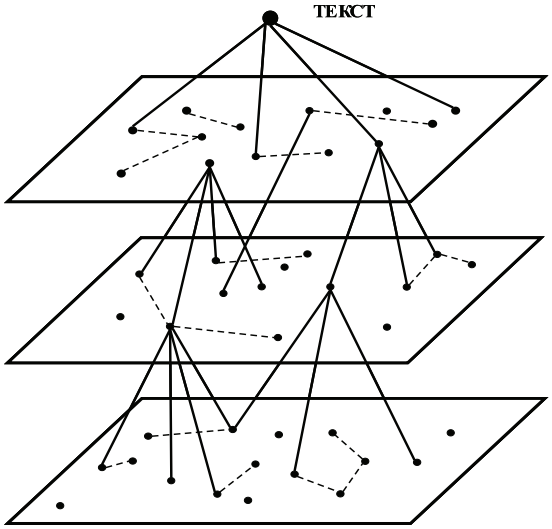

Рис. 1. Иерархическое представление текста

Иерархическая модель может отражать разные параметры текстовой структуры. Одно из представлений – выделение уровней букв, слов и предложений [6]. Структура целого текста определяет не только связи между предложениями, но и связи между словами, а также в некоторой степени связи между элементами слова (буквами и слогами).

Аналогично в [9] построена система смыслового содержания текста. Нижним уровнем также является уровень знаков (буквы, цифры и т.п.), следующий уровень – уровень отдельных слов без учета их значимости в тексте, третий уровень – уровень терминов и последний – уровень понятий и онтологий. В структурно-иерархической словарной модели текста все уровни представляют собой значимые структуры. В иерархической системе смыслового содержания нижние уровни имеют гораздо меньшее значение и используются в основном как вспомогательные элементы для составления объектов более высокого уровня.

При такой модели элементы высшего уровня могут содержать тематически связанные слова и термины, прямо не встречающиеся в текстах или содержащиеся не во всех рассматриваемых текстах, относящихся к данной содержательной области. Именно уровень понятий позволяет решать вопросы, связанные с синонимией и полисемией терминов в текстах. Например, предложена плекс-грамматика [4], позволяющая уменьшить неоднозначность семантических моделей высказываний. С учетом контекстов употребления выражений использование лингвистических отнологий позволяет извлекать сущности и тематические цепочки при рассмотрении разного рода текстов [1].

Потоковое представление текстовых структур

Потоковое представление текстовых данных широко распространено при описании и анализе динамически изменяющихся массивов текстов в Интернете. Текстовые потоки – коллекции документов или сообщений, которые постоянно генерируются и распространяются. Подход, основанный на потоковом принципе представления текстовых данных (stream-based), может быть использован и при анализе больших текстовых объектов, таких как художественные тексты или научные публикации. Весь текст представляют как непрерывный поток текстовой информации, среди которых могут быть выделены его отдельные структурные элементы.

Текст **X** можно рассматривать как последовательность (поток) из n элементов $x_1, x_2, ..., x_n$ некоторого алфавита Q, при этом длина текстовой строки (текста) $|X| = n$ [7]. Элементом текста x_n может быть как одиночный текстовый символ, так и слово, грамматический класс, любая группировка или подстрока символов текста. Схематично потоковое представление текстов изображено на рис. 2.

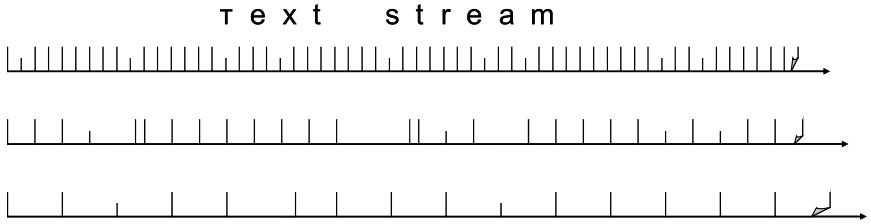

Рис. 2. Потоковое представление текста для разных элементов

При рассмотрении текстовых потоков на передний план выходят задачи кластеризации и классификации, а также прогнозирования и **поиска закономерностей** для определения возможных изменений в тематике текстовых потоков и направлений развития текстов. Также необходимо выявление и отслеживание появления различных текстов или их частей с отличной от прогнозируемой тематикой.

Теория сжатия и Колмогоровская сложность

Задача сжатия (компрессии) информации возникла с появлением первых трудностей при передаче большого объема информации и во многом связана с ограничениями на возможности хранения информации, требованиями к быстрому доступу и передаче в условиях ограниченной пропускной способности. В последнее время возможности хранения и передачи информации возросли, но и количество информации продолжает увеличиваться, поэтому методы и подходы к сжатию данных постоянно развиваются и совершенствуются.

Говоря о текстовых данных и текстовых структурах, необходимо учитывать их особенности, которые позволяют использовать алгоритмы сжатия для дальнейшего анализа текстов. Для информации в текстовом виде характерны следующие особенности, на основе которых проектируются эффективные алгоритмы сжатия:

– последовательное представление;

– избыточность языка и текста.

Последовательное представление данных связано с природой самого текста, который представляется в виде некоторого потока элементов. Причем, опираясь на иерархическую модель текста, можно говорить о потоке (последовательности) данных разных уровней: последовательность букв, последовательность слогов, слов, словосочетаний и т.п. Таким образом, основная идея рассмотрения текстов как потока данных – рассмотрение зависимости появления очередного элемента текста от предшествующих.

Избыточность текста. Избыточность присуща любым текстам и языкам. Избыточность языка в целом объясняется главной функцией языка – коммуникативной, то есть язык избыточен по своей природе, но позволяет легко общаться. Известно, что избыточность текста в большей степени вызвана гласными буквами: текст, написанный без гласных, может быть понят. Если рассматривать лингвистические, языковые причины избыточности, то также можно указать различные функции и категории падежей, морфонологические (связанные с производным словообразованием) явления и т.п. Избыточность художественных текстов также связана с коммуникативностью в обществе, то есть желанием автора некоторым образом повлиять на читателя, вызвать те или иные эмоции, донести некоторую мысль. Но именно избыточность информации, как текстовой, так и любой другой, позволяет эффективно применять различные алгоритмы сжатия, путем избавления от избыточности различными способами.

Как правило, соседние или близко расположенные элементы изображения близки по цвету и интенсивности, а следующие друг за другом кадры отличаются только некоторыми элементами. Текстовые объекты также необходимо рассматривать с позиций контекстного моделирования (context modeling): с одной стороны, любой текст представляет собой поток данных, с другой – для текстов на естественном языке появление любого элемента во многом зависит от предшествующих элементов.

Основу возможности применения алгоритмов сжатия для оценки близости двух объектов составляет понятие Колмогоровской сложности, которую также иногда называют описательной сложностью. Формальное определение Колмогоровской сложности задается следующим образом: сложностью строки x является длина минимальной бинарной программы, выводящей строку x. **Колмогоровская сложность** x при заданном y – это длина наикратчайшей бинарной программы, которая,

получая на вход y, выводит x (это обозначается $K(x|y)$). Колмогоровская сложность x как длина наикратчайшей бинарной программы без входных данных, выводящая x, обозначается $K(x) = K(x|\lambda)$, где λ – отсутствие данных на входе. Данная величина является невычислимой, но ее можно аппроксимировать как длину максимально сжатой версии входной строки [2].

В работе [10] была показана необходимость **модификации** введенного в [8] **нормализованного расстояния сжатия** текстов при рассмотрении текстов разной длины. Предложенная модификация позволяет учесть различный объем рассматриваемых текстов; действительно, в случае, когда один из рассматриваемых текстов существенно меньше другого, то он привносит в обобщенный объект существенно меньше информации. Поэтому производится предварительное разделение больших объектов на отдельные части в зависимости от длины наименьшего рассматриваемого текста.

Теория нечеткости

Принцип нечеткой логики использует нечеткие модели при представлении текстовых данных с применением качественных, неточных или неопределенных признаков. Теория нечеткости возникла в связи с попытками смоделировать способность человека принимать решения в условиях неполной информации, а также способности человека к обобщению различных характеристик и свойств.

Принцип нечеткости позволяет учесть сразу два важных момента, которые возникают в задачах анализа текстов, как, впрочем, и при решении многих других проблем:

- естественная неоднозначность при рассмотрении текстовых объектов;
- принципиальная невозможность учесть все возможные факторы и параметры.

Последняя проблема связана со сложностью моделирования текстовых объектов, необходимостью рассмотрения множества глубинных параметров текста. С другой стороны, любой текст сам по себе является моделью некоторой ситуации или возникшего у автора образа, тем самым моделируя текст, производится опосредованная модель процесса человеческого мышления и других особенностей и свойств человеческого мозга.

Любая построенная модель всегда является некоторым приближением реального объекта, в котором выделены лишь некоторые признаки, важные для решения конкретной задачи. Для других задач необходимо рассматривать другую или измененную модель, в которой учтены другие характеристики. При этом никакая, даже самая сложная модель не может абсолютно точно отразить реальную систему, даже простую. Введение некоторых допущений, не важных для решения конкретных задач, помогает составлять и использовать модели реальных объектов. Но при этом при переходе к модели возникает естественная неопределенность (нечеткость), связанная с влиянием «отброшенных» факторов на рассматриваемые в модели.

Неопределенность модели может быть также связана с самим моделируемым объектом – нечеткость в определении (измерении). Неопределенность-случайность отражает распределение объектов по каким-то параметрам, выраженное количественно, отражает случайные воздействия внешних факторов на рассматриваемый объект.

Естественная (природная) нечеткость текстовых данных обусловлена субъективностью использования терминов, синонимией и многозначностью слов, стилистической и жанровой неоднозначностью, а также эволюцией языков, например, проникновением терминов одной научной области в другие.

Использование **нечеткого отношения** как бинарной функции, определяющей степень выполнения отношения для каждой пары объектов, позволяет формализовать многие реальные явления и задачи при обработке и анализе текстов. Если рассматривать некоторое множество текстов, то для каждой пары может быть вычислена **степень близости**, например, на основе понятия Колмогоровской сложности и определения степени сжатия объединенных объектов. Тогда полученную **матрицу расстояний** можно рассматривать как нечеткое бинарное отношение, заданное на множестве текстовых объектов. На рис. 3 приведен фрагмент таблицы, построенной для художественных текстов русских авторов, элементы которой могут быть интерпретированы как значения нечеткого отношения близости объектов. В работе [5] рассмотрена задача нечеткого разделения пользователей социального сообщества в сети Интернет путем выявления характерных признаков оставленных ими сообщений.

	Idiot.txt	Junost.txt	Otro4estvo.tx	PrestuplNaka	Detstvo.txt
Idiot.txt	0,999337542	0,999425419	0,999467668	0,999351453	0,999386550
Junost.txt	0,999369651	0,997488088	0,997482327	0,999359811	0,997770390
Otro4estvo.tx	0,999437249	0,997753106	0,995670229	0,999411628	0,996290044
PrestuplNaka	0,999391569	0,999394912	0,999414971	0,999426671	0,999336410
Detstvo.txt	0,999530195	0,997793435	0,996939963	0,999436700	0,996750403

Рис. 3. Пример матрицы близости текстовых объектов

Обобщение с единых позиций теоретических аспектов моделирования текстов позволяет эффективно решать основные задачи их обработки: задачи кластеризации, классификации и идентификации. При разработке и проектировании систем анализа текстовых структур целесообразно использовать совокупность моделей текстов, характеризующую различные параметры и особенности текстов на разных уровнях иерархии и учитывающую природную нечеткость. При решении конкретных задач из всей совокупности выбирается только определенный спектр моделей, наилучшим образом отвечающих поставленным задачам. Рассмотренные теоретические аспекты могут найти применение в задачах развития и совершенствования информационно-поисковых систем, а также систем информационной безопасности, в частности при решении задач идентификации Интернет-сообщений и определении авторства исходных кодов программ.

Список литературы

1. Алексеев А.А., Лукашевич Н.В. Автоматическое извлечение сущностей на основе структуры новостного кластера // Искусственный интеллект и принятие решений. – 2011. – № 4. – С. 95–103.

2. Верещагин Н.К., Успенский В.А., Шень А. Колмогоровская сложность и алгоритмическая случайность. – М.: МЦНМО, 2013.

3. Городецкий Б.Ю. Компьютерная лингвистика: моделирование языкового общения // Новое в зарубежной лингвистике. – Вып. 24. – М., 1989.

4. Кучуганов В.Н. Элементы теории ассоциативной семантики // Управление большими системами. – 2012. – Вып. 40. – С. 30–48.

5. Ломакина Л.С., Суркова А.С., Буденков С.С. Кластеризация текстовых данных на основе нечеткой логики // Системы управления и информационные технологии. – 2014. – № 1(55). – С. 73–77.

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

7. Шевелёв О.Г. Методы автоматической классификации текстов на естественном языке: Учебное пособие. – Томск: ТМЛ-Пресс, 2007. – 144 с.

8. Bennett C.H., Gacs P., Li M., Vitanyi P.M.B., Zurek W. Information Distance // IEEE Transactions on Information Theory. – 44:4(1998). – P. 1407–1423.

9. Feldman R., Sanger J. The text mining handbook. Advanced Approaches in Analyzing Unstructured Data // Cambridge University Press. – 2007. – 410 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. – P. 1309–1316.

References

1. Alekseev A.A., Lukashevich N.V. Avtomaticheskoe izvlechenie sushchnostei na osnove struktury novostnogo klastera, Iskusstvennyi intellekt i prinyatie reshenii, 2011, no. 4, pp. 95–103.

2. Vereshchagin N.K., Uspenskii V.A., Shen A. Kolmogorovskaya slozhnost' i algoritmicheskaya sluchainost, Moscow, 2013.

3. Gorodetskii B.Yu. Komp'yuternaya lingvistika: modelirovanie yazykovogo obshcheniya, Novoe v zarubezhnoi lingvistike, Vol. 24. Moscow, 1989.

4. Kuchuganov V.N. Elementy teorii assotsiativnoi semantiki, Upravlenie bol'shimi sistemami, Vol. 40, 2012, pp. 30–48.

5. Lomakina L.S., Surkova A.S., Budenkov S.S. Klasterizatsiya tekstovykh dannykh na osnove nechetkoi logiki, Sistemy upravleniya i informatsionnye tekhnologii, no. 1(55), 2014, pp. 73–77.

6. Surkova A.S. Identifikatsiya tekstov na osnove informatsionnykh portretov, Vestnik Nizhegorodskogo universiteta im. N.I. Lobachevskogo, 2014, no. 3(1), pp. 145–149.

7. Shevelev O.G. Metody avtomaticheskoi klassifikatsii tekstov na estestvennom yazyke, Tomsk: TML-Press, 2007, 144 p.

8. Bennett C.H., Gacs P., Li M., Vitanyi P.M.B., Zurek W. Information Distance, IEEE Transactions on Information Theory, 44:4 (1998), pp. 1407–1423

9. Feldman R., Sanger J. The text mining handbook. Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, 2007, 410 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.

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 «Информатика, системы управления и телекоммуникаций», Волжский государственный университет водного транспорта, г. Нижний Новгород.