УДК 004.912

ПРИКЛАДНЫЕ 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

Представлены прикладные аспекты анализа и обработки текстовых данных, рассмотренные с позиций их применения в информационно-аналитических системах, информационно-поисковых системах и системах обеспечения информационной безопасности. Функционирование этих систем основано на учете основных принципов: принципа системного представления текстов, принципа нечеткой логики и принципа обучающихся систем. Предложена типология и рассмотрены примеры задач, относящихся к классификации, кластеризации и идентификации текстов. Рассмотрены примеры практических реализаций. Решение задачи нечеткого разделения пользователей, входящих в социальное сообщество, основано на выявлении характерных признаков, извлеченных из их сообщений. Решение задачи категоризации (классификации текстов на естественном языке по тематическим (предметным) категориям) основано на предложенной модели текста в виде спектров N-грамм. Показана возможность использования методов сжатия и Колмогоровской сложности для кластеризации текстов исходных кодов и идентификации их авторов.

Ключевые слова: анализ и обработка текстов, сжатие, нечеткость, категоризация, кибербезопасность, тексты исходных кодов

APPLIED 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

Applied aspects of text data analysis and processing are presented and considered from the standpoint of their application in information-analytical systems, information retrieval systems and information security systems. The functioning of these systems is based on the accounting basic principles: principle of systematic text presentation, fuzzy logic principle and learning systems principle. Typology is proposed and examples of texts classification, clustering and identification tasks are considered. The article reviewed examples of practical implementations. The solution for the problem of users' fuzzy separation in the social community is based on the identification of characteristics, which extracted from their messages. The solution for the categorization problem (natural language texts classification by theme (subject) categories) is based on the proposed text model in the form of the N-grams spectra. We have shown the possibility of using compression methods and Kolmogorov complexity for clustering of source codes and the authorship identification.

Keywords: text analysis and processing, compression, fuzziness, categorize, CyberSecurity, the source code texts

Предложенные теоретические положения и методологические аспекты анализа и моделирования текстов [3] позволяют обобщить основные принципы для решения задач анализа текстов и формализовать выбор моделей и методов обработки текстов при решении конкретных прикладных задач. К активно развивающимся направлениям анализа и обработки текстов относятся задачи, связанные с обеспечением работы информационно-поисковых и информационно-аналитических систем, а также систем обеспечения информационной безопасности (рис. 1).

Во многих прикладных системах, связанных с обработкой информации, находят отражение задачи анализа и обработки текста: задачи кластеризации, классификации и идентификации. Специфика решения конкретной задачи зависит от выбранных методов и рассматриваемых данных, а также направленности системы. Приведем некоторые направления и примеры задач, возникающие в практических информационных системах.

А.I. Кластеризация. Разделение данных на отдельные непересекающиеся группы по заданному параметру. *Предварительная обработка большого объема данных без какой-либо дополнительной информации (возможен случай, когда рассматриваются не только текстовые данные)*.

А.II. Классификация. Отнесение документов к известным группам, при котором каждый текст может принадлежать нескольким классам (нечеткая классификация). *Рассмотрение большого текста, в котором требуется выделить отдельные части, близкие по некоторым признакам (тематике, стилистическим особенностям)*.

Прикладные аспекты анализа и моделирования текстов		Задачи анализа и обработки текстов		
		I Кластеризация	II Классификация	III Идентификация
Направления	A. Информационно-аналитические системы	A.I	A.II	A.III
	B. Информационно-поисковые системы	B.I	B.II	B.III
	C. Системы обеспечения информационной безопасности (CyberSecurity)	C.I	C.II	C.III

Рис. 1. Прикладные аспекты анализа текстов

A.III. Идентификация. Выявление и проверка различных характеристик текста, проверка возможности использования их в качестве идентификационных признаков. *Идентификация авторства и стиля художественных текстов.*

B.I. Кластеризация. Разделение документов по группам в специализированных информационно-поисковых системах. *Кластеризация патентной документации и заявочных материалов, извлекаемых из разных патентных баз.*

B.II. Классификация по предметным тематическим категориям (категоризация) – отнесение неизвестного текста к одной или нескольким тематическим категориям. *Классификация по тематике сообщений в новостной ленте. Определение эмоциональной окрашенности текстов в рекомендательных системах.*

B.III. Идентификация. Определение признаков автоматического перевода на основе выявления особенностей написания текстов на родном языке. *Проблемы переводного плагиата и заимствования. Определение ключевых слов, терминов и фраз с целью автоматического аннотирования и реферирования*

C.I. Кластеризация. Анализ текстов в социальных сетях, рассмотрение потока текстов глобальной сети для выявления авторских характеристик и обнаружения предпосылок к нежелательным действиям. *Определение эмоционального состояния автора. Определение исходного языка текста (переводной или оригинальный текст)*

C.II. Классификация по различным признакам, когда имеет место отнесение текста к некоторым классам по конкретному заранее заданному признаку, такому, как стиль или жанр текста, временные характеристики и т.п. *Задача классификации Интернет-сообщений (комментариев, писем, сообщений). Определение эмоционального состояния автора текстовых сообщений.*

C.III. Идентификация. Определение особенностей написания программ как по исходному, так и по бинарному коду. *Идентификация авторов исходных кодов программ. Проверка истинности авторства программ в учебных целях.*

Примеры практической реализации

Решение задачи нечеткого разделения пользователей, входящих в некоторое социальное сообщество, основано на выявлении характерных признаков, извлеченных из их сообщений [4]. Практическая реализация выполнялась с помощью исследовательского пакета KNIME [1], разработанного в одном из университетов Германии и доступного для свободного использования в учебных и исследовательских целях, данные были получены со страницы Slashdot. Исходное подмножество содержало около 140,000 комментариев к 500 заметкам от приблизительно 24,000 пользователей. Кластеризация осуществлялась по методу, основанному на алгоритме нечеткой кластеризации Fuzzy c-means (FCM).

Результат полученной кластеризации для десяти кластеров приведен на рис. 2. Кластеризация проводилась по параметрам: «авторитетность» пользователя-автора сообщений, отражающая степень соответствия тематике и «ссылочность» (оценка последователей). Также учитывалась положительная/отрицательная оценка каждого пользователя на основе употребляемых им слов. Размер получившегося кластера показывает размер радиуса окружности на диаграмме, насыщенность цвета отражает положительность/отрицательность объектов в кластере.

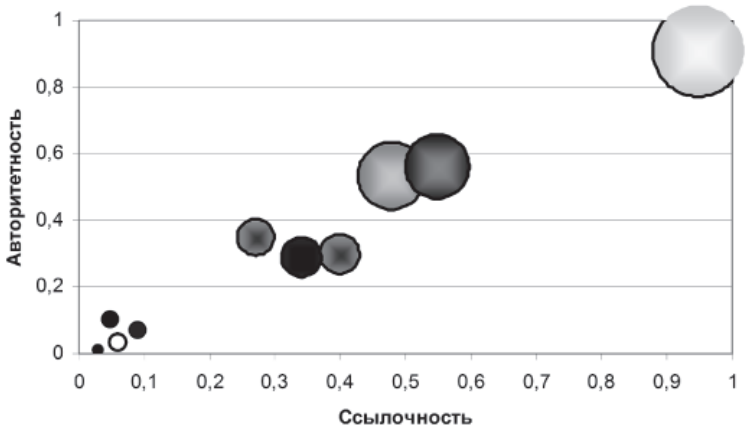

Рис. 2. Результат кластеризации пользователей в социальном сообществе

В результате кластеризации выделяется один кластер с высокой степенью авторитетности и ссылочности, при этом его объекты обладают наибольшей положительностью. Выделяются несколько кластеров активных пользователей в центре схемы с высокой оценкой авторитетности и ссылочности. Среди них встречаются кластеры и с положительными и с отрицательными оценками.

Для задачи категоризации (классификации текстов на естественном языке по тематическим (предметным) категориям из заранее определенного набора) были рассмотрены стандартные методы классификации, применяемые к предложенным моделям текста на основе спектров различной детализации.

Система состоит из модуля обработки текста документа и создания на его основе *N*-граммной древовидной модели документа и построения спектров различной детализации [5]. Модуль оценки эффективности разработанной модели текста и тестирования результатов ее применения с различными классификаторами построена на основе статистического программного пакета KNIME.

Была проведена оценка классификации при рассмотрении спектров *N*-грамм разного уровня, при этом эффективность классификаторов оценивалась для четырех уровней детализации спектров, равных соответственно 2, 3, 4 и 5, результаты приведены в таблице. Эффективность классификатора определялась как процент верно классифицированных документов от их общего числа.

Практическая реализация методов идентификации авторов текстов исходных кодов рассматривалась в рамках систем для обеспечения кибербезопасности и информационной безопасности. В понятие «**кибербезопасность**» (CyberSecurity) входит широкий спектр практических приемов, инструментов и концепций, тесно связанных с технологиями информационной и операционной безопасности.

При распространении текстов в электронном виде с использованием глобальных сетей большое значение приобретают задачи обеспечения безопасности и предотвращения так называемых киберпреступлений. Одним из направлений развития кибербезопасности является подраздел, который занимается задачами обработки объектов, представленных в текстовом виде (документы, письма, сообщения, тексты программ и т.д.). С текстовыми объектами связаны такие кибернарушения, как распространение спама и вирусов, террористические угрозы и многие другие [9, 10].

Результаты категоризации

Классификатор	Уровень детализации			
	2	3	4	5
Support Vector Machines (SVM)	1,00	1,00	0,94	0,83
Probabilistic Neural Network (PNN)	0,97	0,78	0,53	0,50
Multilayer Perceptron (MLP)	0,94	0,97	0,97	1,00
Fuzzy Rules (FR)	0,89	0,75	0,61	0,56
Decision Trees (DT)	0,86	0,94	0,78	0,31
Naive Bayes (NB)	0,25	0,30	0,33	0,33

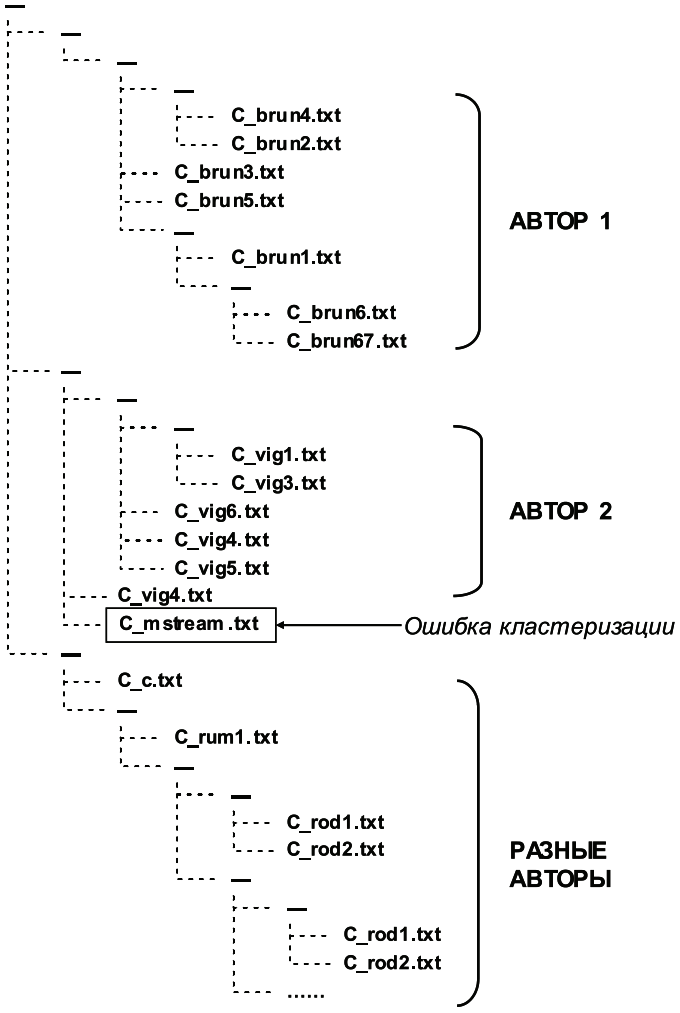

Рис. 3. Фрагмент кластеризации текстов программ

Можно выделить основные задачи анализа текстовых данных в рамках обеспечения кибербезопасности:

- определение авторства литературных и научных текстов;
- выявление характерных признаков интернет-сообщений (письма по электронной почте, записи в блогах, комментарии и т.п.);
- определение авторства программных кодов [7];
- и др.

Была проверена возможность использования методов сжатия и Колмогоровской сложности [8] для кластеризации текстов исходных кодов и идентификации их авторов. Фрагмент построенного иерархического разбиения по классам, соответствующим разным авторам, приведен на рис. 3. Необходимо также отметить, что при кластеризации текстов исходных кодов кластеры, относящиеся к разным языкам программирования, выделяются на самом верхнем уровне и уже в каждом из них происходит кластеризация по авторам программ.

Рассмотрены некоторые прикладные аспекты построения систем анализа и обработки текстовых данных, опирающиеся на предложенную методологию, которая учитывает основные принципы анализа текстов: принцип системного представления текстов, принцип нечеткой логики и принцип обучающихся систем. Однако рассмотренные задачи не ограничивают сферы применения предложенной методики. К другим проблемам информационно-поисковых систем можно отнести такие задачи, как фильтрация и рубрикация документов, автоматическое аннотирование и сегментирование текстов и др. Для успешного решения подобных задач необходимо предусмотреть использование лингвистических и онтологических знаний [2, 6]. Еще одним направлением

развития может служить рассмотрение таких текстовых данных, как тексты патентной документации и заявочных материалов, текстовые данные, создаваемые при использовании различными базами знаний и системами документооборота в организациях и на предприятиях.

Список литературы

1. Аналитическая система KNIME, сайт – URL: www.knime.org (дата обращения 06.06.2015).

2. Кучуганов В.Н. Вербализация реальности и виртуальности. Ассоциативная семантика // Искусственный интеллект и принятие решений. – 2011. – № 1. – С. 55–66.

3. Ломакина Л.С., Суркова А.С. Теоретические аспекты концептуального анализа и моделирования текстовых структур // Фундаментальные исследования. – 2015. – № 2 (часть 17). – С. 3713–3717.

4. Ломакина Л.С., Суркова А.С. Информационные технологии анализа и моделирования текстовых данных: монография. – Воронеж: Изд-во «Научная книга», 2015. – 208 с.

5. 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.

6. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: Изд-во Московского университета, 2011. – 512 с.

7. Маевский Д.А., Чербаджи Ю.П. Определение авторства программного обеспечения по исходному коду программ // Радиоэлектронные и компьютерные системы. – 2014. – № 6. – С. 64–68.

8. Суркова А.С., Родионов В.Б. Алгоритм разбиения неструктурированного множества текстовых объектов // Научно-технический вестник Поволжья. – Казань, 2013. – № 5. – С. 298–300.

9. Kontostathis A., Edwards L., Leatherman A. Text mining and cybercrime // Text Mining. Applications and Theory. Ed. by Berry M.W., Kogan J. – Wiley, 2010. – P. 149–164.

10. Dua S., Du X. Data Mining and Machine Learning in Cybersecurity. – New York, 2011. – 224 p.

References

1. Analytical system KNIME, Available at: www.knime.org (accessed 6 June 2015).

2. Kuchuganov V.N. Verbalizacija realnosti i virtualnosti. Associativnaja semantika. Iskusstvennyj intellekt i prinjatie reshenij, 2011, no. 1, pp. 55–66.

3. Lomakina L.S., Surkova A.S. Teoreticheskie aspekty konceptual nogo analiza i modelirovanija tekstovyh struktur. Fundamental nye issledovanija. 2015. no. 2 (chast 17), pp. 3713–3717.

4. Lomakina L.S., Surkova A.S. Informacionnye tehnologii analiza i modelirovanija tekstovyh dannyh: Monografija. Voronezh: Izdatelstvo «Nauchnaja kniga», 2015, 208 p.

5. 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.

6. Lukashevich N.V. Tezaurusy v zadachah informacionnogo poiska. M.: Izdatel stvo Moskovskogo universiteta, 2011, 512 p.

7. Maevskij D.A., Cherbadzhi Ju.P. Opredelenie avtorstva programmnogo obespechenija po ishodnomu kodu program. Radiojelektronnye i komp juternye sistemy, 2014, no. 6, pp. 64–68.

8. Surkova A.S., Rodionov V.B. Algoritm razbienija nestrukturirovannogo mnozhestva tekstovyh ob ektov. Nauchno-tehnicheskij vestnik Povolzhja, Kazan , 2013, no. 5. pp. 298–300.

9. Kontostathis A., Edwards L., Leatherman A. Text mining and cybercrime. Text Mining. Applications and Theory. Ed. by Berry M.W., Kogan J. Wiley, 2010. pp. 149–164.

10. Dua S., Du X. Data Mining and Machine Learning in Cybersecurity. New York, 2011. 224 p.

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.