факторов текстовыми описаниями, построение семантических сетей для этих описаний и их последующую хеш-кластеризацию [7] с получением скалярного семантического фактора. На момент нахождения его требуемого приращения (по прототипу) строим различные варианты семантических подсетей исходной совокупной сети, хеш-значения которых равны требуемому приращению, а затем явно указываем термины и связи между ними, которые необходимо должны присутствовать в описании цели разрешения ситуации (для нахождения самого этого описания могут быть использованы поисковые машины и анализаторы текстов).

1. Филиппович А.Ю., Интеграция систем ситуационного, имитационного и экспертного, Эликс+ (2003).
2. Ткаченко Т.Я., Инструментальная среда системотехнического обслуживания сложных объектов, ГОУ ВПО «УГТУ-УПИ» (2002).
3. Поспелов Д.А., Ситуационное управление: Теория и практика, Наука, (1986).
4. Дудко В.А., Динамическое моделирование ситуационного управления промышленным предприятием, Тамбов (2004).
5. Лисиенко В.Г., Моделирование и разработка системы диагностики технологического процесса для управления качеством продукции (на примере процесса непрерывного литья заготовки): учеб. пособие, ФГОУ ВПО НГТИ (2008).
6. Симанков В.С., Шопин А.В., Ситуационное управление сложным объектом в условиях нечеткой исходной информации, Труды ФОРА (2004).
7. Организация размещения данных и доступа к данным. Хеширование и кластеризация. Особенности СУБД Oracle. [Электронный ресурс]. Режим доступа: http://rema44.ru/resurs/study/dbmat/db_hash_cluster.ppt

ОЦЕНКА ТОЧНОСТИ КЛАССИФИКАЦИИ ТЕКСТОВ В ЗАВИСИМОСТИ ОТ ИХ ЧИСЛА СРЕДСТВАМИ DATA MINING

Бызова А.К.[1*], Гольдштейн С.Л.[1]

[1)] Уральский федеральный университет имени первого Президента России Б.Н. Ельцина, г. Екатеринбург, Россия

*E-mail: anastasia.byzova@gmail.com

ESTIMATION OF ACCURACY OF TEXT CLASSIFICATION DEPENDING ON THE QUANTITY BY DATA MINING

Byzova A.K.[1*], Goldstein S.L.[1]

[1)] Ural Federal University, Yekaterinburg, Russia

Известно, что полнота и точность – меры, противоречащие друг другу в том смысле, что 100%-ую полноту легко достичь, просто поместив все документы в i-ый класс (точность будет мала), и наоборот 100%-ую точность можно обеспечить, помещая в i-ый класс малое число документов (полнота будет мала) [1].

Также известно, что достаточными для классификации по индивидуальному стилю будут тексты объемом в 800 предложений или 6÷9 тысяч слов [2].

2 гипотеза: можно ли за счет тщательной подготовки материала снизить количество текстов для обучения модели классификации высокой точности.

Задача: имеется 4 класса текстов (экономические, медицинские, естественные и технические). Исходные данные для построения модели отобраны в виде 20 научных текстов объемом 3000 слов на русском языке (*.txt) для каждого из классов, а в качестве тестовых данных – 5 технических текстов такого же объема на том же языке (*.txt). Требуется распознать технические тексты средствами Data Mining (программа Weka [3]).

Процесс интеллектуального анализа текстов реализован поэтапно:

- предварительная обработка: отобранные тексты импортируются в формат *.arff (*Attribute-Relation File Format*),
- обучение, начиная с исходных данных, но с одним текстом из естественно-технического класса, затем с двумя и т.д. (в качестве метода выбран алгоритм дерева принятия решений),
- проверка: испытание модели на тестовой выборке.

В итоге получили 5 моделей. На каждой испытали тестовую выборку. Результаты тестирования представлены на рис. 1. Видно, что, начиная с трёх текстов можно добиться точности более 50%.

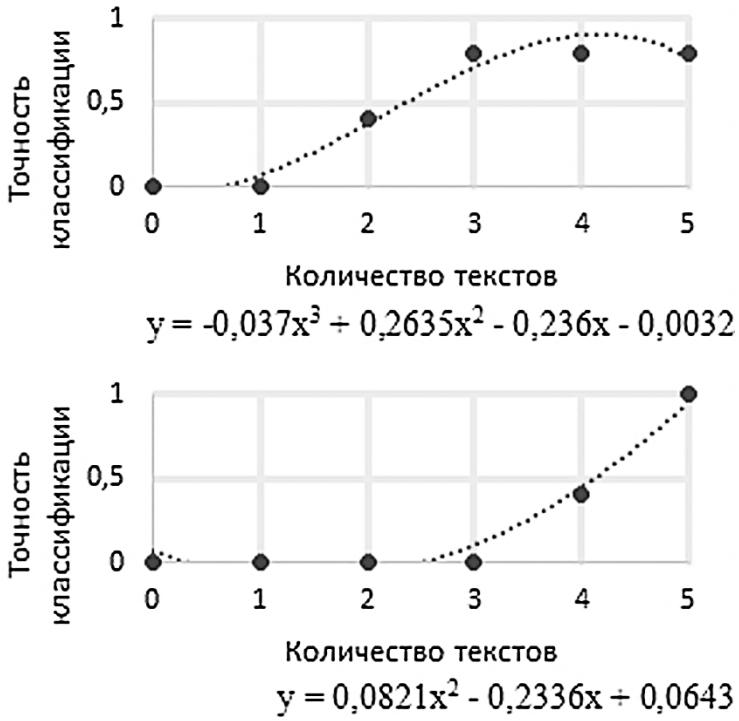

Рис. 1 Зависимость точности классификации от числа текстов для 4-х классов текстов и для 8-ми классов текстов соответственно

Далее увеличили количество классов вдвое: экономические, медицинские, естественные, технические, юридические, физико-математические, искусство-

ведческие, психологические. Требовалось распознать технический текст. Результаты приведены на рис. 1. Здесь наблюдаем смещение границы, определенной ранее: точность более 50% достигаем, начиная с пяти текстов.

Поскольку классов стало больше, то возникает более высокая степень неопределенности, а значит для точности классификации технических текстов более 50% требуется большее число текстов для построения модели классификации.

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Е.И. Большакова, Э.С. Клышинский, Д.В. Ландэ и др. – М.: МИЭМ, 2011. – 272 с.
2. Классификация текстов с помощью деревьев решений и сетей прямого распространения / Шевелев О.Г., Петраков А.В. // Вестник Том. гос. ун-та, - 2006. – № 290. – С. 300 – 307.
3. Computer Science Department, University of Waikato: [сайт]. URL: http://www.cs.waikato.ac.nz

АГЕНТНО-ОРИЕНТИРОВАННАЯ МОДЕЛЬ ВЕРТИКАЛЬНЫХ МИГРАЦИЙ У РАЗЛИЧНЫХ ВИДОВ ЦИАНОБАКТЕРИЙ

Чеснокова О.И.[1*]

[1)] Уральский федеральный университет имени первого Президента России Б.Н. Ельцина, г. Екатеринбург, Россия

*E-mail: choksy@mail.ru

AGENT-BASED MODEL OF VERTICAL MIGRATIONS OF DIFFERENT CYANOBACTERIA

Chesnokova O.I.[1*]

[1)] Ural Federal University, Yekaterinburg, Russia

A 2-dimensional virtual environment, populated with autonomous decision-making agents, was created. Agents evolve moving around the environment under their own autonomous control, gathering both resources necessary for survival and reproduction. 6 groups of possible behavioral strategies were sorted out and compared to real cyanobacterial strategies.

Цианобактерии насчитывают более 2000 видов и обитают практически повсеместно вплоть до шельфовых ледников, горячих источников, арктических и антарктических озёр, а также могут выживать в растениях, лишайниках и животных в качестве эндосимбионтов. В работе [1] токсичные цианобактерии, обитающие в воде, разделены на шесть групп в соответствии со стратегиями их поведения. Однако достаточно хорошо изучены лишь цианобактерии, способные регулировать плавучесть.