УДК 681.3

Е. А. Энгель

НЕЧЕТКИЕ ГИБРИДНЫЕ СИСТЕМЫ ДЛЯ ОБРАБОТКИ ИНФОРМАЦИИ

Описывается разработанная гибридная система обработки информации. Рассматриваются выявленные на базе кластерного анализа особенности научных текстов, которые используются при разработке нечетких нейронных сетей для анализа текстов. Описывается нечеткая гибридная система анализа научных текстов и результаты ее испытаний.

Ключевые слова: кластерный анализ, нечеткая нейронная сеть, гибридная система.

В настоящее время применение методов анализа данных часто бывает затруднено тем, что для конкретной задачи из какой-либо предметной области выборка характеризуется не числовым характером атрибутов. В частности, в макроэкономических, социологических, маркетинговых, медицинских базах данных широко используется лингвистическая форма представления данных. Для оперирования подобными данными необходимо организовать среду хранения нечетких атрибутов. В настоящее время существует ряд теоретических и практических разработок, позволяющих создавать хранилища данных с нечеткими атрибутами. Это отражено в работах Дидьера Дюбуа [1], Мичинори Наката, Генри Прада, в которых исследуется нахождение элементов, с достаточно важными характеристиками; Ж. К. Куберо, Ф. Куенца, И. Бланко, М. А. Вила, где рассматриваются неполные функциональные зависимости в сравнении с нахождением знаний в базах данных; исследования Греда Вагнера связаны с логической перестройкой нечеткого обращения в базах данных и логических программах. В работах М. А. Вила, Ж. К. Куберо, О. Понс, Дж. М. Медина исследуются запросы в объектно-ориентированных нечетких базах данных; Анны Изабель Агуилера Фарако и Леонида Жозе Тинео Родригаса разработана нечеткая дедуктивная модель и доведена до реального воплощения; в работе Е. А. Горбоконенко [2] предложена и реализована нечеткая реляционная модель данных.

Несмотря на значительные результаты, достигнутые в интеллектуальном анализе данных, остается ряд нерешенных задач. Так, использование не числовых атрибутов, в том числе нечетких, не поддержано разработанными методами анализа.

Гибридная система обработки информации. Разработана гибридная система обработки информации, которая обладает следующими особенностями:

– для извлечения знаний использует статистические данные, которые интерпретирует как обучающие выборки для нечетких нейронных сетей;

– представляет знания в виде лингвистических переменных (функций принадлежности), нечетких продукций и обученных нейронных сетей; редукция множества нечетких продукций, настройка функций принадлежности и базы правил выполняется с помощью генетических алгоритмов.

Базовая задача анализа информации – формирование индексов в соответствии с рубриками. Рубрики представляют собой классы объектов проблемной области, причем трудно вручную составить исходную достаточно полную классификацию. Следовательно, необходимо решить задачу кластеризации – порождение системы нечетких классов для данных. Затем каждый новый объект необходимо относить к нечетким классам.

Вход: множество данных $X = \{x_1, x_2, ..., x_N\}$, max num – максимальное количество кластеров.

Выход: оптимальное множество кластеров $C = \{C_1, C_2, ..., C_c\}$.

Шаг 1. $c_{\text{opt}} = \max\text{num}$, $c = \max\text{num}$, $i = 1$. Случайным образом выбирается объект $x \in X$ в качестве точки старта p. Выполняется многошаговый max min алгоритм с параметрами X, c, i, p для поиска оптимального множества кластеров $C = \{C_1, C_2, ..., C_c\}$ для c. Вычисляется функция оценки SC для C.

Шаг 2. Выполняется алгоритм слияния для получения множества кластеров $C' = \{C'_1, C'_2, ..., C_c\}$, выбирается центр C'_1 в качестве точки старта p. $c = c - 1$, $i = 2$. Выполняется многошаговый maxmin алгоритм с параметрами X, c, i, p для поиска оптимального множества кластеров $C^* = \{C^*_1, C^*_2, ..., C^*_c\}$ для c. Вычисляется функция оценки SC для C^* и принимается как SC^*. Если $SC^* > SC$, тогда $SC = SC^*$, $C = C^*$, $c_{\text{opt}} = c$. Повтор шага 2, пока $c \le 2$.

Шаг 3. Вывод: $C = \{C_1, C_2, ..., C_{\text{opt}}\}$ – оптимальное множество кластеров.

Данный алгоритм имеет ряд преимуществ перед другими алгоритмами кластеризации (см. табл. 1).

Реализация нечеткой гибридной системы для анализа научно-технических текстов. Разработанная нечеткая гибридная система была применена для анализа научно-технических текстов. Первоначально была решена задача определения жанра текста. Очевидно, что такую задачу можно решить только на основе аннотаций текстов, которые должны присутствовать в каждом ресурсе. Текст можно отнести к нечеткому классу на основе встречаемости ключевых слов. Использовался описанный выше гибридный алгоритм. Далее был проведен детальный анализ ключевых слов, характеризующих научно-технический жанр.

Основная цель научного произведения – сообщение о результатах проведенного исследования и объяснение способа их получения, формулировка новых идей и их обоснование. Соответственно, научный дискурс представляет собой логически взаимосвязанную последовательность речевых (дискурсивных) действий, соответствующих операциям научного мышления [3–5]. К типичным операциям относится обоснование вывода, выдвижение гипотезы, введение термина и понятия, приведение фак-

тов и доказательств, подведение итогов и др. Как правило, эти операции более или менее явно помечаются общенаучными словами и выражениями, образующими общенаучный лексикон.

Наиболее явными маркерами мыслительных операций служат так называемые ментальные перформативные высказывания (например, *ниже рассмотрим, особо подчеркнем*), которые обычно квалифицируют применяемую операцию. В работе [5] описаны виды перформативных высказываний, опирающиеся на широкий круг ментальных перформативных глаголов (*опишем, предположим, заметим* и т. п.):

– канонические, с глаголом в первом лице множественного числа (*мы покажем*);

– «установочные», с модальным или оценочным словом (*необходимо/нетрудно заметить*);

– в форме деепричастия или деепричастного оборота (*резюмируя вышесказанное*);

– в безличной форме (*представляется, что*…).

В научных текстах встречаются также дескриптивные (косвенные) варианты ментальных перформативов, используемые либо для перифразирования (*эти данные приводятся в таблице 3*), либо для установления связей между высказываниями текста (*далее кратко изложен*).

Кроме перформативов используются также маркеры очередности (*во-первых, наконец и др.*); коннекторы – союзы и союзные слова (*однако, благодаря тому, что и т. п.*); слова-оценки (*возможно, по-видимому и т. п.*), часто встречающиеся и в текстах других стилей [6].

Все указанные виды дискурсивных маркеров имеют ярко выраженный метатекстовый характер [7], большинство из них функционируют в тексте как метатекстовые операторы, предполагающие в своем составе сентенциальный или атрибутивный аргумент: *подчеркивается, что S, рассмотрим N.*

К общенаучному лексикону относятся абстрактные существительные, называющие аппарат научно-познавательной деятельности (*вопрос, проблема, понятие, анализ, процедура, схема и др.*). Эти существительные называются общенаучными переменными [7], поскольку имеют обязательную атрибутивную валентность (*проблема N, понятие N*). Хотя они не используются в метатекстовой функции, они играют важную роль в структуризации научной информации. Общенаучные переменные обычно употребляются в научных текстах с перформативными глаголами (*ввести понятие, подвергнуть анализу*) [8].

Таким образом, общенаучный лексикон охватывает широкий круг семантически и грамматически разнородных слов и выражений общенаучной речи. Важно, что он не зависит от конкретной предметной области и сравнительно немногочисленен. Заметим, что лексикон состоит из общеупотребительных слов, поэтому их метатекстовая функция в конкретном предложении текста (т. е. выполняют ли они роль дискурсивного маркера) в общем случае может быть установлена только в результате исследования контекста их употребления.

Для анализа научно-технических документов предлагается использовать, кроме традиционных терминологического и морфологического словарей, следующие словарные средства, отображающие специфику научной прозы:

– словарь общенаучных слов и выражений;

– лексико-синтаксические шаблоны типичных фраз научной речи.

При построении словаря общенаучных слов и выражений была проведена классификация собранных ключевых слов. Выражения были разбиты на группы путем кластеризации гибридным алгоритмом в признаковом пространстве, характеризующем смысл и функции выражений в тексте, без учета их грамматической формы и синтаксических характеристик. В итоге получилось 53 группы, каждая из которых является классом слабой эквивалентности и включает в общем случае несколько семантически близких выражений разной грамматической природы. Каждой группе приписана соответствующая операция научного дискурса; эти операции частично приведены в табл. 1.

Для каждой единицы словаря общенаучной речи указывается ее классификационная группа (дискурсивная операция); для словосочетаний описываются их синтаксические характеристики (разрывность/неразрывность и др.).

Для распознавания в тексте словарного словосочетания необходима информация о семантико-синтаксических валентностях составляющих его слов. Такую информацию можно представить в виде нечеткого лексико-синтаксического шаблона, который фиксирует лексемы и их грамматическую форму, а также задает синтаксические условия заполнения своих пустых мест (валентностей). Кроме того, в виде шаблонов удобно представлять клишированные конструкции научной речи, составленные из нескольких словарных единиц и имеющие фиксированную синтаксическую структуру. К числу таких конструкций относятся определения новых терминов, состоящие из одного предложения, например, фраза «…значение, которое используется для расширения первоначального набора, мы будем называть существенным значением…». Указанная конструкция схематически может быть описана как

Таблица 1

Сравнение алгоритмов кластеризации

Алгоритм	Применимость к сильно сгруппированным данным	Необходимость указания количества кластеров	Чувствительность к входным параметрам	Применимость к неравномерно распределенным данным
Гибридный алгоритм	Да	Нет	Нет	Да
k-средних	Да	Да	Да	Да
Субстрактивный	Да	Нет	Да	Нет
Maxmin	Да	Нет	Да	Да
Fuzzy c-means	нет	да	Да	Да

NG_{ACC} [«мы»] «будем называть»T_{INS},

где «мы» и «будем называть» – совместно встречающиеся лексемы, причем слово «мы» может отсутствовать; T_{INS} – определяемый термин, выраженный согласованной именной группой, главное слово которой имеет форму творительного падежа; NG_{ACC} – определение или объяснение авторского термина, выраженное согласованной именной группой (возможно, расширенной придаточным предложением), главное слово которой имеет форму винительного падежа. Каждый шаблон реализуется отдельно обученной нечеткой нейросетью (табл. 2).

Представленная в словаре общенаучной речи и наборе нечетких шаблонов семантико-синтаксическая информация позволяет производить содержательный анализ научно-технических текстов – распознавание примененных дискурсивных маркеров и операций научного дискурса.

Разработка нечетких лексико-синтаксических шаблонов. Проблема, возникающая при разработке шаблонов конструкций, заключается в определении контекстов, однозначно сигнализирующих дискурсивный (метатекстовый) характер употребляемых слов и словосочетаний. Для ее решения необходимо проводить исследование контекстов употреблений конструкций, целесообразно использовать аппарат нечетких нейронных сетей.

Такое исследование было проведено для контекстов конструкций, определяющих новые термины. Нечеткими нейросетями было обработано около 50 научно-технических текстов, и из них были выделены те фразы, которые использовались при определении или пояснении нового термина. После их предварительного анализа было получено первоначальное множество лексем, входящих в конструкции определений, что позволило в дальнейшем автоматизировать процесс поиска новых конструкций и контекстов, посредством обученных нечетких нейросетей.

Так как количество разных контекстов было велико, контексты для каждой фиксированной лексемы (или для двух-трех совместно встречающихся лексем) были представлены отдельно обученной нечеткой нейросетью, что позволило выявить соответствующие синтаксические конструкции, которые затем были формализованы в виде нечетких лексико-синтаксических шаблонов, реализуемых настроенными нечеткими нейросетями.

В состав шаблонов входят следующие элементы:

– литералы, т. е. конкретные лексемы из словаря общенаучной речи (*определим*, *будем называть* и др.), а также сокращения (т. н.) и знаки препинания. Литеральные элементы заключаются в кавычки;

– символьные обозначения слов определенной части речи и грамматической формы, которые могут заполнять свободные места (слоты) шаблона; например, N – существительное, V – глагол, P – предлог, Pa – причастие;

– символьные обозначения определенных грамматических конструкций, например, Ng – именная группа, T – определяемый термин, выраженный именной группой (простой или расширенной);

– условия, уточняющие грамматические характеристики рассмотренных элементов и записываемые в угловых скобках, например: <Ng. number = V. number> означает, что число группы Ng и глагола V совпадают, а условие <person = 3> фиксирует употребление третьего лица.

При записи условий используются символьные обозначения грамматических характеристик: времени (tense), числа (number), лица (person), рода (gender), падежа (case) и конкретных падежей (например, nom – именительный, ins – творительный).

К примеру, шаблон Ng «,» Pa<«названный»> T<: case = ins> <Ng. case = Pa. case Ng. gender = Pa. gender Ng. number = Pa. number = T. number> описывает случаи вида «По результатам генерации форм, слова были разбиты на группы, названные профилями» (в этом примере подчеркнута фиксированная шаблоном лексема). В то же время, фраза «...устойчивого выражения, названного в заголовке, в левой (объясняемой) части словарной статьи» не вводит новый термин и не удовлетворяет шаблону, так как после причастия «названный» не стоит конструкция, имеющая требуемые в шаблоне характеристики.

Разработанные к настоящему моменту шаблоны, реализуемые настроенными нечеткими нейросетями, образуют нечеткую гибридную систему обработки научных текстов, покрывающую примерно 60–70 % определений терминов, встречающихся в научных текстах. Важно, что добавляя новые нечеткие шаблоны, учитывающие все более сложные конструкции и контексты, можно постепенно наращивать мощность процедуры распознавания в тексте операций научного дискурса.

Таким образом, с использованием математического аппарата, описанного выше, была реализована нечеткая

Таблица 2

Операции научного дискурса

Операции	Примеры слов и выражений
Описание и констатация	Укажем, что; характеризуя
Конкретизация и уточнение	В частности; в дополнение к
Причинно-следственные связи	По этой причине; следовательно
Актуализация темы	Перейдем к; рассмотрим
Выделение информации	Особо подчеркнем; необходимо отметить
Предположения и допущения	Предположим/допустим, что
Определения	Будем называть; по определению
Сравнение и противопоставление	С одной стороны; в отличие от; по сравнению с
Иллюстрация и приведение примеров	К примеру; например
Обобщение и резюмирование	Суммируя вышесказанное; в общем
Упорядочивание и перечисление	Во-первых; наконец
Помета общенаучной переменной	Идея, модель, результат
Выражение мнения и оценивание	Целесообразно считать; по-видимому

гибридная система обработки информации, позволяющая производить кластеризацию данных с лингвистическими атрибутами и выявлять зависимости в виде нечетких продукций. Представленный алгоритм позволяет проводить кластеризацию сильно сгруппированных и неравномерно распределенных данных: нечувствителен к входным параметрам и не требует указания количества кластеров.

Разработанная нечеткая гибридная система обработки информации была эффективно использована для анализа научных текстов. Охарактеризованы разрабатываемые для этого словарные средства – словарь общенаучной речи и нечеткие лексико-синтаксические шаблоны характерных фраз. Кратко описаны составные элементы шаблонов, язык их записи, а также методика их построения, базирующаяся на нечетком нейросетевом анализе. Все это дает возможность приступить к реализации процедуры распознавания дискурсивной структуры научно-технических текстов.

Библиографический список

1. Дюбуа, Д. Теория возможностей. Приложения к представлению знаний в информатике : пер. с фр. / Д. Дюбуа, А. Прад. М. : Радио и связь, 1990.

2. Горбоконенко, Е. А. Представление нечеткой информации в СУБД / Е. А. Горбоконенко, Н. Г. Ярушкина // Тр. 7-й нац. конф. по искусств. интеллекту. М. : Изд-во физ.-мат. лит., 2000.

3. Митрофанова, О. Д. Язык научно-технической литературы / О. Д. Митрофанова. М. : Изд-во МГУ, 1973.

4. Николаев, А. М. Описание семантики научного текста с позиций теории речевых актов (на материале рецензии на научно-техническую работу) / А. М. Николаев // НТИ. 1998. Сер. 2. № 7.

5. Рябцева, Н. К. Ментальные перформативы в научном дискурсе / Н. К. Рябцева // Вопросы языкознания. 1992. № 4.

6. Севбо, И. П. Сквозной анализ как шаг к структурированию текста / И. П. Севбо // НТИ. 1989. Сер. 2. № 2.

7. Вежбицка, А. Метатекст в тексте / А. Вежбицка // Новое в зарубежной лингвистике. М. : Прогресс, 1978. Вып. VIII.

8. Словарь глагольно-именных словосочетаний общенаучной речи. М. : Наука, 1973.

E. A. Engel

FUZZY HYBRID SYSTEM USAGE FOR DATA PROCESSING

The designed hybrid system for data processing is described. It coveres the features of the scientific texts detected on the base of cluster analysis and realized by the fuzzy neural networks for scientific texts analysis. The fuzzy hybrid system for scientific texts analysis.The results of tests are described.

Keywords: cluster analysis, fuzzy neural network, hybrid system.

УДК 519.21

Т. А. Ширяева, С. И. Сенашов

СТОХАСТИЧЕСКИЕ МОДЕЛИ АНАЛИЗА ФИНАНСОВЫХ СИСТЕМ

Предложен метод формирования оптимального портфеля ценных бумаг на основе формулы Кокса–Росса–Рубинштейна. Итогом работы является перечень конкретного вида мера риска и применение полученных мер к конкретным ценным бумагам и формирование предпочтения одних бумаг другим.

Ключевые слова: мера риска, стохастический анализ.

Эффективное управление движением капитала в рамках организаций (или физических лиц) предполагает временное вложение свободных средств в ценные бумаги для извлечения дополнительной прибыли.

Прежде чем вкладывать средства в ценные бумаги, необходимо изучить рынок ценных бумаг. Стоимость каждой ценной бумаги можно рассматривать как некоторую случайную величину. В любой ситуации желательно знать вид закона распределения.

Для решения данной задачи используют различные методы. В частности, можно рассматривать известные стандартные вероятностные распределения и проводить проверки гипотез, что случайная величина имеет свое распределение со своими параметрами.

В случае, если распределение установлено, то характеристикой ценной бумаги является функция риска, которая может быть определена различным образом. Поэтому, сравнивая функции риска для наборов ценных бумаг, можно выбрать бумаги с наименьшим риском, т. е. говорить о предпочтительность одних ценных бумаг по отношению к другим.

Две математические модели рынка ценных бумаг. Инвесторы условно могут быть разделены на два типа.