Автоматическая классификация текстов

БЕЛОНОГОВ Г. Г.
ГИЛЯРЕВСКИЙ Р. С.
КОЗАЧУК М. В.
НОВОСЕЛОВ А. П.
ХОРОШИЛОВ А. А.

ВИНИТИ, г. Москва, Россия

Авторами статьи разработана концепция автоматической классификации текстов. Составлены частотные словари ключевых слов и словосочетаний по поисковым образам документов, извлеченным из БД ВИНИТИ На основе частотных словарей составлен машинный словарь для автоматической классификации. Построена программная модель системы автоматической классификации. Опытная эксплуатация этой модели подтвердила правильность принципов, положенных в ее основу.

Процесс классификации текстов состоит в их распределении по классам на основе признаков сходства и различия, отражающих наиболее существенные черты смыслового содержания этих текстов. Определение смыслового содержания текстов — труднейшая задача. Ее решение с помощью ЭВМ в настоящее время возможно лишь на путях формализации семантико-синтаксической структуры текстов. При этом могут использоваться формализованные модели различной степени сложности. Одной из них может быть простейшая понятийная модель, в которой смысловое содержание текстов описывается перечнями наименований содержащихся в них понятий. Наименования понятий выражаются в текстах как отдельными словами, так и (значительно чаще) словосочетаниями.

Статистические исследования показывают, что тексты, принадлежащие к различным тематическим классам, отличаются друг от друга не столько перечнями наименований понятий, сколько распределениями вероятностей их появления. Этим можно воспользоваться при построении систем автоматической классификации текстов.

В качестве простейших моделей описания текстов, принадлежащих к различным тематическим классам, можно принять распределения вероятностей появления в них различных слов (частотные словари). В этом случае принадлежность какого-либо текста к той или иной тематической области может быть определена путем сравнения его частотного словаря с частотными словарями, характеризующими различные классы текстов. Текст может быть отнесен к тому классу, где будет наблюдаться наибольшее сходство распределений.

Такая процедура автоматической классификации текстов была реализована авторами на ПЭВМ и позволила убедиться в правильности выбранного направления исследований. Но у нее было, как минимум, два недостатка: 1) здесь в качестве основной единицы смысла было выбрано не словосочетание (что было бы более естественным), а слово; 2) применялась слишком грубая модель, аппроксимирующая распределения вероятностей появления слов (весовые коэффициенты слов в тематических словарях могли принимать только два значения). Поэтому было принято решение перейти на более адекватную модель, основанную на использовании частотных словарей ключевых слов (КС) и словосочетаний (СС), составленных по поисковым образам документов (ПОД) реферативных баз данных ВИНИТИ.

Частотные словари ключевых слов и словосочетаний составлялись по массивам ПОД за 1981–1986 гг. и за 1999 г. по всем тематическим областям, представленным в этих массивах. Всего было составлено более 20 частотных словарей. Они были объединены в шесть частотных словарей, получивших условные названия "Автоматика и радиоэлектроника" (АИРЭ), "Биология", "Геология", "Машиностроение", "Физика", "Экономика". Словарь "Автоматика и радиоэлектроника" объединял ключевые слова и словосочетания по собственно АИРЭ, вычислительной технике, робототехнике, информатике, полиграфии и связи; словарь "Биология" — по биологии (включая физико-химическую биологию и биотехнологию), химии и экологии; словарь "Геология" — по собственно геологии, горному делу, географии и геофизике; словарь "Машиностроение" — по машиностроению, металлургии, транспорту, аэронавтике, космонавтике, электротехнике и энергетике; словарь "Физика" — по физике, астрономии, математике и механике; словарь — "Экономика" — по экономике промышленности. В табл. 1 представлены некоторые статистические данные об этих словарях.

Таблица 1

Статистические данные о частотных словарях ключевых слов и словосочетаний, составленных по ПОД БД ВИНИТИ

№ п/п	Тематическая область	Объем словаря	Объем исходного массива КС и СС
1	Автоматика и радиоэлектроника	65 199	205 419
2	Биология	861 446	5 184 130
3	Геология	53 065	159 815
4	Машиностроение	102 298	622 363
5	Физика	35 074	185 692
6	Экономика	9172	117 577
	Итого	1 126 254	6 474 996

Объединенные частотные словари были упорядочены по убыванию частот и каждый из них был разделен на десять участков, обеспечивающих одинаковое покрытие исходных массивов ключевых слов и словосочетаний, по которым эти словари составлялись. При этом было замечено, что во всех словарях, начиная с третьего от начала участка и до восьмого, отношение количества лексических единиц каждого последующего участка к количеству лексических единиц, содержащихся в предыдущем участке, было равно примерно двум. Поэтому было принято решение назначать весовые коэффициенты словам и словосочетаниям словаря с учетом этой закономерности: лексическим единицам третьего участка был

присвоен весовой коэффициент "32", четвертого участка — "16", пятого — "8", шестого — "4", седьмого — "2", восьмого — "1".

Таблица 2

Фрагменты политематического классификационного словаря наименований понятий
(объем словаря — 45 тыс. лексических единиц)

противомикробное средство * 2.08
противообрастающее вещество * 2.02
противоопухолевая активность * 2.08
противоопухолевое вещество * 2.32
противоопухолевое действие * 2.02
противоопухолевый иммунитет * 2.04
противоопухолевое средство * 2.32

протопланетарная туманность * 5.02
протопланетное облако * 5.04
протосолнечная туманность * 5.04
проточно-инжекционный анализ *2.16
проточно-инжекционная система * 2.02
проточная цитометрия * 2.08/3.02

профессиональная болезнь * 2.04
профессиональное воздействие * 2.08
профессиональная вредность * 2.02
профессиональная деятельность *2.01
профессиональное заболевание * 2.08
профессиональное облучение * 2.08/3.02
профессиональное образование * 6.04

прочностная характеристика * 3.02/4.01
прочность грунта * 5.04
динамическая прочность * 5.08
прочность корпуса * 4.02
прочность летательного аппарата * 4.01
прочность материала * 5.04
прочность породы * 3.01
прочность сцепления * 4.01
усталостная прочность * 4.08

проявочное устройство *1.08/4.01
проявочное устройство электрофотографического аппарата * 4.01

равновесная конфигурация * 5.04
равновесное состояние * 5.01
равновесная фаза * 5.04
равновесная форма * 5.04
равномерность обработки * 1.02
равномерное пространство * 5.04
радарное измерение * 3.02
радарное изображение * 3.02
радарная интерферометрия * 3.02
радарное наблюдение * 3.08
радарная съемка * 3.02

распределительный вал * 4.02
распределительная сеть * 1.04/4.16
распределительная система * 1.04
распределительное устройство * 1.08/4.02
распределительный центр * 4.08
распределительный щит * 4.02
распределительная электрическая сеть * 4.02
распространение акустических волн * 1.32/5.04
распространение в атмосфере * 5.04
распространение волн *1.02/5.08
распространение животных *3.08
распространение звука * 3.04/5.16
распространение издательской продукции * 1.16
распространение излучения * 1.04/5.04
распространение информации * 1.02

Лексическим единицам первого, второго, девятого и десятого участков весовые коэффициенты не назначались и при построении системы автоматической классификации они не использовались. Первый и второй участки содержали в основном однословные термины, девятый и десятый — относительно редкие термины. Исключение из системы классификации словарей двух первых участков было обусловлено малой "дифференцирующей силой" входящих в их состав лексических единиц, а двух последних — желанием сократить объемы словарей.

В отличие от первого варианта системы автоматической классификации текстов, при построении ее второго варианта было принято решение иметь не несколько классификационных словарей, а один сводный словарь, в котором для каждой лексической единицы указывался ее относительный вес в соответствующей тематической области. Фрагменты такого словаря приведены в табл. 2.

Здесь в левой части словаря стоят наименования понятий (ключевые слова и словосочетания), справа от них — индексы тематических областей (согласно табл. 1), а через точку — двузначные весовые коэффициенты. Если лексическая единица встречается в нескольких тематических областях, то сочетания индексов тематических классов и весовых коэффициентов отделяются друг от друга косой чертой.

При программной реализации системы классификации этот словарь был приведен к другой форме (табл. 3)

Таблица 3

Фрагмент машинной формы представления политематического классификационного словаря наименований понятий
(объем словаря 45 тыс. строк)

Хеш-коды Весовые коэффициенты

nDRqudlk 00 08 04 00 00 04

PDmчn Vg 00 08 04 00 00 00
MoЛpkgfE 00 02 00 00 00 00
2TrJbNMn 00 00 00 01 00 00
FsBAzzMt 00 00 00 08 00 00

ВбчНГZчn 00 00 04 00 00 00
ЬПMUkHfj 02 02 32 02 08 00
БраAjlAG 00 00 02 00 00 00
WyeHYdqp 00 00 00 00 04 00
ЖзБЬPoLt 00 00 02 00 04 00
УуZPAJJI 02 02 32 01 32 04
аюпAKlNy 00 00 00 00 02 00

Здесь в каждой строке таблицы первые слева восемь байтов содержат хеш-код (сжатый код) наименования понятия, а остальные байты — информацию о весах понятий в различных тематических областях. Веса понятий представляются сочетаниями из двух цифр. Они следуют друг за другом через пробел в порядке возрастания номеров тематических областей (см. табл. 1).

Построенный авторами второй вариант системы автоматической классификации текстов включал в свой состав машинный словарь наименований понятий, представленный в формате табл. 3, и комплекс программ, предназначенный для определения принадлежности текстов к тематическим областям. Классификация текстов осуществлялась с помощью следующих процедур:

1. Морфологический анализ текстов, проводимый в целях определения грамматических характеристик входящих в их состав слов.

2. Семантико-синтаксический анализ текстов, проводимый в целях распознавания в них словарных наименований понятий (словосочетаний и слов).

3. Определение для каждой тематической области сумм весов наименований понятий, опознанных в тексте. Текст считался принадлежащим той тематической области, для которой сумма весов наименований понятий оказывалась наибольшей.

Процедуры морфологического и семантико-синтаксического анализа текстов были заимствованы из системы русско-английского машинного перевода RETRANS [1,2] и модифицированы с учетом условий их функционирования в системе автоматической классификации.

Испытание второго варианта системы автоматической классификации текстов позволило убедиться в правильности принципов, положенных в его основу. В табл. 4 приведены результаты решений по трем текстам, принадлежащим к тематическим областям "Автоматика и радиоэлектроника", "Биология" и "Геология".

Таблица 4

Результаты экспериментов

1. Фрагмент текста по автоматике и радиоэлектронике

D 1. Применение экспертных регуляторов для систем управления динамическими объектами
D 2. Квазипериодическое управление линейной дискретной системой с выпуклыми симметричными ограничениями в задаче нуль-управляемости (случай сингулярной переходной матрицы)
D 3. Оценка запаса устойчивости минимально фазовых дискретных систем
D 3. Периодические режимы в частотно-импульсных системах
D 4. Синтез динамических регуляторов пониженного порядка по квадратичному критерию
D 5. Применение метода символических возмущений для обработки вырожденных ситуаций в геометрических алгоритмах
D 6. Модели сложных систем, построенные с использованием метода крупнозернистого параллелизма на транспьютерной сети
D 7. Клеточные автоматы в математическом моделировании и обработке информации
D 8. Общее решение задачи аналитического конструирования регуляторов, оптимальных по квадратичным функционалам
D 9. Определение временной задержки сигналов методом адаптивной цифровой фильтрации
D 10. Методы сжатия данных в вычислительных системах
D 11. Формирование стереоскопического изображения виртуальной сцены в системе "Гипервизор"
D 12. Синтез и визуализация трехмерных и стереоскопических медицинских диагностических изображений на персональном компьютере

Результат автоматической классификации

НАИМЕНОВАНИЕ ТЕМАТИКИ	ВЕСОВОЙ КОЭФФИЦИЕНТ
1. АВТОМАТИКА И РАДИОЭЛЕКТРОНИКА	6022
2. БИОЛОГИЯ	2898
3. ГЕОЛОГИЯ	1379
4. МАШИНОСТРОЕНИЕ	1733
5. ФИЗИКА	3224
6. ЭКОНОМИКА	880
ТЕМАТИЧЕСКИЙ КЛАСС ТЕКСТА:	
АВТОМАТИКА И РАДИОЭЛЕКТРОНИКА	6022

2. Фрагмент текста по биологии

Ref001. Информация, основанная на структурной биологии

Анатомия или структурная биология является фундаментальной по отношению к основным биомедицинским наукам. Важность анатомии заключается в том, что структура логически предшествует функции, а взятые вместе они обеспечивают фундамент клинической медицины. Трудности в изучении анатомии связаны первично с иерархической и гетерогенной природой биологических структур человека, которые простираются от молекул до целого организма, включая разнообразие величины и формы таких основных органов, как мозг, сердце, печень, почки. Типичная клетка может содержать 10 тыс. различных белков, каждый из которых может содержать сотни или тысячи атомов. Для того, чтобы распознать патологию, если она наблюдается, врач должен хорошо разбираться в сложности этой структурной организации. Познавательная способность человека помогает абсорбировать пространственную информацию визуально, однако визуальная система имеет важные ограничения. Другие трудности при изучении анатомии связаны с бурным развитием биомедицинских знаний, особенно на молекулярном уровне. В связи с этим необходимы компьютерные автоматические и полуавтоматические системы для сбора и хранения информации. В общих чертах описаны достижения в компьютеризации банка данных, составлен словарь анатомических терминов и их синонимов. Симпозиум по применению компьютеров в медицине состоялся в 1989 г. в США.

Результат автоматической классификации

НАИМЕНОВАНИЕ ТЕМАТИКИ	ВЕСОВОЙ КОЭФФИЦИЕНТ
1. АВТОМАТИКА И РАДИОЭЛЕКТРОНИКА	471
2. БИОЛОГИЯ	1548
3. ГЕОЛОГИЯ	464
4. МАШИНОСТРОЕНИЕ	278
5. ФИЗИКА	496
6. ЭКОНОМИКА	209
ТЕМАТИЧЕСКИЙ КЛАСС ТЕКСТА:	
БИОЛОГИЯ	1548

3. Фрагмент текста по геологии

D 1. О возрасте фауны грызунов (Rodentia, Mammalia) Буранской свиты Зайсанской впадины (Восточный Казахстан)
D 2. О возрасте большеземельской серии Тимано-Уральской области
D 3. Палеонтология и культура
D 4. Изменение систематического и этолого-трофического состава донных шельфовых сообществ на границе мела и палеогена
D 5. Флора сосудистых растений в водораздельных озерах востока Большеземельской тундры
D 6. Система млекопитающих и историческая зоогеография
D 7. Зоогеография раннеолигоценовых бассейнов Западной Евразии по двустворчатым моллюскам

D 8. Эндемичные и космополитные сообщества фораминифер и остракод в среднеюрских бассейнах Сибири

D 9. Шельфовые сообщества фораминифер Охотского моря

D 10. Радиолярии и возраст железомарганцевых конкреций

D 11. Морфология, структура и функциональное значение маргинальных образований беззамковых брахиопод семейства CranidaeMenke

D 12. Физико-химическое моделирование процессов гидротермального рудообразования

Результат автоматической классификации

НАИМЕНОВАНИЕ ТЕМАТИКИ	ВЕСОВОЙ КОЭФФИЦИЕНТ
1. АВТОМАТИКА И РАДИОЭЛЕКТРОНИКА	933
2. БИОЛОГИЯ	1279
3. ГЕОЛОГИЯ	5220
4. МАШИНОСТРОЕНИЕ	606
5. ФИЗИКА	808
6. ЭКОНОМИКА	656
ТЕМАТИЧЕСКИЙ КЛАСС ТЕКСТА:	
ГЕОЛОГИЯ	5220

ЗАКЛЮЧЕНИЕ

Проблема автоматической классификации текстов относится к числу трудных проблем моделирования интеллектуальной деятельности человека. Трудность ее решения обусловлена тем, что признаки, определяющие сходство и различие текстов, весьма многочисленны, а формы их представления — многообразны. Для решения этой проблемы необходимо создавать процедуры семантико-синтаксического анализа текстов и составлять словари наименований понятий большого объема.

В настоящей статье предпринята попытка решения проблемы автоматической классификации текстов путем их концептуального анализа и сопоставления распределений вероятностей появления концептов в анализируемых текстах с распределениями вероятностей их появления в текстах, заведомо принадлежащих к определенным классам. Эксперименты показали, что такой подход имеет хорошие перспективы.

В заключение мы хотели бы отметить, что предлагаемый нами метод классификации текстов предполагается использовать для автоматического выбора тематических настроечных словарей в системе русско-английского и англо-русского фразеологического машинного перевода RETRANS. Работы по этой системе поддерживаются Российским фондом фундаментальных исследований.

ЛИТЕРАТУРА

1. Белоногов Г. Г., Зеленков Ю. Г., Новоселов А. П., Хорошилов Ал-др А., Хорошилов Ал-сей А. Системы фразеологического машинного перевода. Состояние и перспективы развития // Научно-техническая информация. Сер. 2.— 1998.— № 12.

2. Белогонов Г. Г., Зеленков Ю. Г., Новоселов А. П., Хорошилов Ал-др А., Хорошилов Ал-сей А. Метод аналогии в компьютерной лингвистике // Научно-техническая информация. Сер. 2.— 2000.—№ 1.

3. Белоногов Г. Г., Зеленков Ю. Г. Еще раз о принципе аналогии в морфологии // Научно-техническая информация. Сер. 2.— 1995.— № 3.