УДК 004.912: 303.7

ВЕГА – КОМПЬЮТЕРНАЯ СИСТЕМА КЛАССИФИКАЦИИ И АНАЛИЗА ТЕКСТОВ

К.К. Боярский, Е.А. Каневский

ВЕГА представляет собой систему для классификации и обработки как анкетной, так и другого рода текстовой информации. Обсуждаются особенности анализа текстовой информации, основанной на контент-аналитическом сравнении фраз. Рассматриваются возможности системы и особенности ее функционирования. Большое внимание уделяется вопросам практического использования описываемой системы.
Ключевые слова: анализ текста, классификация, контент-анализ, словари, социологические анкеты, открытые вопросы, статистический анализ, шкалирование.

Введение

В текстовых массивах, циркулирующих в обществе, содержатся специфические познавательные возможности. Анализируются различные тексты: материалы средств массовой информации, политические материалы в виде программ партий и кандидатов в электоральных кампаниях, уставы партий и движений, биографии и дневники, научные публикации и др. Социология сама стимулирует появление в обществе специальных текстов, проводя конкурсы сочинений и автобиографий, организуя интервью.

Анализом содержания текстов занимаются многие исследователи при изучении влияния средств массовой информации на общественное мнение, документов истории и культуры, политического, экономического, юридического и даже экологического сознания общества. Одной из первых систем автоматизированного анализа текстов явилась General Inquirer (Гарвард, 1968), широко использующая различные словари [1]. Современные системы TACT и TextPack также основаны на использовании словарей [2].

Сегодня наибольшее распространение получили два метода анализа: кластерный анализ и контент-анализ. Математический аппарат кластерного анализа можно использовать для автоматического выделения естественных тематических групп из случайной однородной массы текстового материала (например, набора статей из различных журналов). Созданы методы классификации полнотекстовых баз данных (БД) на основе алгоритмов визуального эвристического кластерного анализа документов. Методы основаны на составлении частотных словарей и выделении тех слов, частота встречаемости которых во всех текстах БД превышает общеупотребительную частоту [3].

Другим методом качественно-количественного изучения содержания текстов является контент-анализ (КА). В процессе КА все многообразие текстов по интересующей исследователя тематике сводится к набору определенных элементов, которые затем подвергаются подсчету и анализу. На использовании контент-анализа построены программы TACT, ARRAS, TextPack, SYREX, SATO [4].

Обычно в качестве элемента содержания (единицы анализа) при «машинном» КА используют слово, которому ставят в соответствие определенную категорию. Это удобно, так как слово изначально выделено в тексте пробелами. Однако слово характеризуется лишь номинативной, назывной функцией. Единицей выражения мысли является предложение, которое используется в качестве единицы содержания при классическом («ручном») КА. Мы используем в качестве элемента содержания несколько другую единицу анализа – фразу, которая может состоять даже из одного слова. Каждая фраза является выражением одного суждения, одной мысли. При сравнении фразы считаются идентичными друг другу при совпадении двух–трех слов или одного–двух ключевых слов [5].

Любые попытки применения КА к текстовым массивам неизбежно связаны с проблемой классификации и, следовательно, с разработкой классификатора. Простейшая

структура классификатора обычно напоминает таблицу и содержит категории (группы) и их модальности (типы). Более сложный классификатор имеет древовидную структуру, состоящую из классов и групп.

В большинстве исследований заранее составляется формализованная, полностью закрытая схема классификации [6]. При повторяющихся исследованиях, например, в средствах массовой информации, стандартный классификатор даже помогает сравнивать результаты очередного обследования с предыдущими. Однако при анализе быстро меняющейся ситуации даже опытный исследователь, проводя обследования через 1–2 года, не может заранее создать полностью готовый классификатор. Причина ясна: сместилась тематика ответов респондентов, их волнует уже не то, что год или два тому назад. Поэтому на один и тот же вопрос (открытого типа) они отвечают совсем не так, как раньше. Очевидно, что в этом случае классификатор должен уточняться непосредственно в процессе КА, при осмыслении материалов данного опроса.

Система ВЕГА

На основе разработанных методик в 1991–1997 г.г. была создана система ДИСКАНТ [7]. Основная ее цель – обеспечить мощной компьютерной поддержкой работу различных исследователей и аналитиков, имеющих дело с анализом текстовой информации. ДИСКАНТ представляла собой систему для классификации и обработки как текстовой, так и другого рода анкетной информации, которая хранится в БД системы. Система была разработана под DOS и позволяла классифицировать содержание текстовой информации по множеству оснований, составлять указатели и частотные словари слов и фраз, осуществлять поиск слов в тексте и в словаре. Были разработаны разнообразные способы визуализации результатов в виде гистограмм, циклограмм, сопряженных двумерных диаграмм.

На основе этой системы сегодня разработана более совершенная система ВЕГА, предназначенная для работы с текстовой и цифровой информацией при проведении социальных и социологических исследований. Система в основном предназначена для обработки структурированной и, прежде всего, анкетной информации. Кроме того, система позволяет выполнять некоторые элементы анализа текста: составление словарей, подсчет встречаемости слов, поиск слов по словарю и по тексту и т.д. Система обеспечивает статистический анализ ответов на закрытые и полузакрытые вопросы.

В системе ВЕГА используется методика итерационной классификации текста. В качестве единицы анализа выбрана фраза. Слова, наиболее точно отражающие смысл фразы, выделяются прописными буквами – это ключевые слова. Выделение фраз и ключевых слов осуществляется вручную в процессе предварительной подготовки текста к анализу.

В системе вся исходная информация хранится в собственной БД. При этом в случае анкетной информации ответы на одну анкету составляют одну запись, ответы на один вопрос помещаются в одно поле. Каждое поле имеет свой номер, имя и определенный тип. Набор всех этих параметров для данной БД образует ее структуру. Выбранный набор типов полей позволяют наиболее адекватно хранить ответы респондентов, учитывая, что в анкетах имеются открытые, закрытые и полузакрытые вопросы. Для удобства пользователей и контроля вводимой информации предусмотрены следующие типы полей.

– ***Символьное*** содержит любую последовательность текстовых символов и может быть разделено на отдельные фразы. Размер текста в каждом поле – до 32767 символов. Разделенный на фразы текст имеет вид

  “НЕСТАБИЛЬНОСТЬ ¦ отсутствие РЕАЛЬНОЙ ВЛАСТИ ¦ несоблюдение ЗАКОННОСТИ”.

– ***Целое*** содержит одно или несколько целых положительных десятичных чисел, которые соответствуют номерам вариантов ответов, выбранных респондентом. Они разделяются друг от друга запятыми. Содержимое поля имеет вид
  "2" или "1,3,6" или "1, 3, 6".
– ***Вещественное*** содержит только одно вещественное десятичное число.
– ***Диапазон*** содержит одно или два положительных вещественных числа и вспомогательные знаки (">", "<", "–"). Содержимое поля имеет вид
  "<89.6732" или ">7845" или "56 – 238.349".

Кроме того, имеются составные поля, в которых числовая часть специальным знаком отделяется от символьной. Это позволяет производить независимую обработку разных частей такого поля. К полям составного типа относятся:

– ***Смешанное 1,*** первая часть которого представляет собой целый тип, а вторая – символьный. Содержимое поля имеет вид
  "7 ‡ Вооруженные силы".
– ***Смешанное 2,*** первая часть которого представляет собой символьный тип, а вторая – целый. Содержимое поля имеет вид
  "Вооруженные силы ‡ 1,7".

Классификация текста предполагает распределение всех текстовых фраз по классификатору. Последующий анализ результатов классификации, сравнение полей или массивов, проверка гипотез уже не представляют технических и содержательных трудностей. В *ручном варианте* классификация состоит из трех основных процедур [5, 8].

1) Разрабатывается классификатор, представляющий собой древовидную структуру классов и групп, приблизительно соответствующую содержанию текста. Отбираются фразы, наиболее часто встречающиеся в тексте.

2) Выполняется собственно классификация. При этом в каждый класс (и в каждую группу) помещается некоторое количество предварительно отобранных фраз. Желательно, чтобы в каждой такой фразе было выделено (прописными буквами) хотя бы одно ключевое слово. Эти фразы получают статус нормативных фраз.

3) Производится идентификация. Каждой фразе основного текстового массива ставится в соответствие одна из нормативных фраз, наиболее близкая ей по смыслу. В результате этой текстовой фразе приписывается порядковый номер выбранной нормативной фразы – своеобразный адрес в классификаторе, по которому в дальнейшем определяется класс и группа текстовой фразы.

В принципе полная реализация классификационного процесса осуществляется множественными итерациями всех трех перечисленных процедур. Такая итерационная схема позволяет выполнять анализ порциями. Например, сначала можно классифицировать массу более легких, типовых фраз, а затем уже переходить к нестандартным фразам, последовательно улучшая результаты.

Сущность предлагаемой методики сводится к тому, что вместо 100%-ной классификации всех фраз проводится классификация только части фраз, а все остальные фразы отождествляются с ними. Это дает двойное преимущество. Во-первых, сокращается объем работы по классификации фраз. Во-вторых, при любом изменении классификатора, – а это не исключение, а правило при анализе текстов – достаточно изменить класс и группу у ряда нормативных фраз, чтобы автоматически произошли соответствующие изменения и у всех остальных фраз, связанных с ними.

В режиме *автоматической классификации* система самостоятельно выполняет все три процедуры, описанные выше [5]. При этом очередная текстовая фраза сравнивается с уже имеющимися нормативными фразами. Если при сравнении обнаруживается похожая нормативная фраза, то ее порядковый номер приписывается этой текстовой фразе. В противном случае данная текстовая фраза объявляется нормативной, и для нее заводится новая группа в классификаторе. В результате автоматической классификации получается

классификатор, состоящий из одного класса, который содержит множество групп. Далее работа исследователя заключается в объединении этих групп в новые группы и в классы в соответствии с их тематикой с целью организации соответствующего классификатора. Практика показывает, что для достаточно точной классификации количество исходных групп в 5–6 раз превышает количество результирующих групп [9].

В результате выполнения этих процедур фиксируются все три составляющие процесса классификации текста: своего рода «каталожный ящик» – классификатор, приписанные к соответствующим ящичкам (классам и группам) нормативные фразы и система соответствия (идентификаторы) каждой исходной фразы с одной из нормативных фраз. Никаких других сведений по классификации система не хранит.

Процедура сравнения фраз, не совпадающих текстуально, заключается в том, что каждая текстовая фраза разлагается на отдельные слова, ее составляющие [5]. Для каждого такого слова в нормативном словаре (словаре, составленном по нормативным фразам) ищется слово, совпадающее с ним по правилам сравнения слов. Если такое слово в этом словаре имеется, то для этого слова фиксируется номер той нормативной фразы, в которой он встретился. Для каждой из встретившихся нормативных фраз подсчитывается ее вес. Та нормативная фраза, которая получит наибольший вес, и будет считаться наиболее близкой к анализируемой текстовой фразе.

Основные характеристики системы

В табл. 1 заголовки граф показывают основные режимы работы системы ВЕГА, реализованные в виде соответствующих экранных форм. Ячейки табл. 1 представляют различные возможности каждого режима.

<table>
<tr><th>ФАЙЛ</th><th>БАЗА ДАННЫХ</th><th>СЛОВАРИ</th><th>КЛАССИ-ФИКАЦИЯ</th><th>АНАЛИЗ ТЕКСТА</th><th>СТАТ. АНАЛИЗ</th><th>СЕРВИС</th></tr>
<tr><td>Создать</td><td>Просмотр записей</td><td>Вывод словаря</td><td>Классификатор</td><td>по классам</td><td>1-мерный</td><td>Сжатие базы</td></tr>
<tr><td>Открыть
Закрыть</td><td>Просмотр поля</td><td>Пермутационный вывод</td><td>Классификация</td><td>по признакам</td><td>2-мерный</td><td rowspan="5">Параметры:
– общие
– шрифты
– идентичность</td></tr>
<tr><td>Импорт
Экспорт
Печать</td><td rowspan="4">Структура</td><td rowspan="4">Новый словарь</td><td>Нормативные (фразы)</td><td>по сочетаемости классов</td><td>3-мерный</td></tr>
<tr><td rowspan="3">Идентификация</td><td rowspan="3">по количеству фраз</td><td>Таблица</td></tr>
<tr><td>Объединение БД</td><td>МСА</td></tr>
<tr><td></td><td>Шкала</td></tr>
</table>

Таблица 1. Структура системы ВЕГА

Основной формой является БАЗА ДАННЫХ. При переходе на эту форму всегда выводится закладка Просмотр записей. В этом режиме осуществляется ввод и редактирование информации, вывод ее в файл или на печать и переход во все другие режимы. В текстовые окна выводится содержимое ряда полей выбранной записи. Обеспечивается возможность разбиения текстовых ответов респондента на фразы и выделения ключевых слов. Кроме того, обеспечивается выделение фразы для последующей классификации (занесение в Сундук – вспомогательную БД для хранения отобранных фраз с целью их последующей классификации) или непосредственный переход к ее классификации. Закладка Структура обеспечивает задание полей. При отсутствии записей допускается любое редактирование вновь введенной структуры, при наличии хотя бы одной записи – только просмотр и редактирование имен полей.

Закладка Объединение баз служит для формирования новой БД на основе уже имеющихся. Режим обеспечивает перепись содержимого выбранных полей из одной БД в другую с одновременной конвертацией типов полей.

Форма СЛОВАРИ обеспечивает анализ частоты встречаемости отдельных элементов и различного вида поиски. Как известно, широкое использование словарей характерно для различных методов анализа текста. В системе ВЕГА словари (вместе с пермутационным выводом) служат, прежде всего, для изучения текста и разработки классификатора [10], хотя могут использоваться и по своему прямому назначению. При выводе словаря обеспечивается просмотр его содержимого, вывод его в файл или на печать, его удаление, а также различного вида поиски. В окне вывода словаря для каждого айтема (словарной единицы) указывается количество слов или фраз, объединенных в ней. Для выбранного айтема приводятся все адреса его вхождений в исходный текст – номер записи и номер поля в БД, что обеспечивает простой переход к контексту. При пермутационном выводе обеспечивается такое представление текста, при котором заданные тем или иным способом слова словаря расположены в центре экрана, а контекст их окружения укорочен до размера стандартной строки экрана. В начале каждой строки указан адрес выводимого текста. Пермутационный вывод является одним из основных инструментов исследователя при первичном анализе текста. Могут быть созданы словари четырех типов: по словам, по ключевым словам, по фразам и частотный словарь. При этом в трех первых типах словарей айтемы расположены по алфавиту, а в последнем – по частоте встречаемости.

Форма КЛАССИФИКАЦИЯ обеспечивает создание классификатора и распределение всех текстовых фраз по нему. Вывод классификатора осуществляется в виде таблицы (служит для ввода и редактирования) или в виде дерева (используется для объединения групп и классов и изменения конфигурации классификатора). Закладка Классификация обеспечивает классификацию фраз, хранящихся в Сундуке. В процессе классификации исследователь присваивает классифицируемой фразе выбранные им класс и группу, после чего эта фраза становится нормативной. Обеспечен просмотр и редактирование нормативных фраз. Возможно непосредственное изменение класса и группы у отдельной фразы. Можно удалить нормативную фразу. Если по этой нормативной фразе идентифицирована хотя бы одна текстовая фраза, то такие идентификаторы гасятся.

Закладка Идентификация позволяет идентифицировать текстовые фразы в нескольких режимах. В *ручном режиме* каждая фраза основного текстового массива уже не классифицируется, а ставится в соответствие с одной из нормативных фраз. Результаты идентификации можно проверить в режиме «Просмотр». Возможно использование одного из *автоматических режимов* идентификации. При выборе режима «Автоидентификация» автоматически выполняется только идентификация текстовых фраз – к этому моменту должны существовать классификатор и набор нормативных фраз. При выборе режима «Создание классификатора» не только автоматически выполняется идентификация, но одновременно создаются классификатор и нормативные фразы, т.е. весь процесс классификации выполняется автоматически. Полученный таким образом классификатор обычно содержит множество групп, расположенных в одном единственном классе, и дополнительно требует объединения этих групп [9].

Форма АНАЛИЗ ТЕКСТА обеспечивает анализ распределения фраз из заданных полей по классам и группам. Обеспечивается стыковка с Microsoft Exel и SPCC на уровне передачи в них первичных результатов анализа. Для сопоставления текстовых и числовых ответов следует сформировать вторичные признаки. Для каждого такого признака запоминается его имя, номера используемых им полей и условия его вычисления. После этого система позволяет анализировать распределение фраз по классификатору и вторичным признакам с представлением результатов в виде таблицы и двумерных гистограмм. Результаты анализа сочетаний классов в каждой анкете для выбранного набора полей представляются в виде гистограмм (рис. 1). Этот анализ позволяет понять, сочетания каких тем встречается в ответах чаще всего.

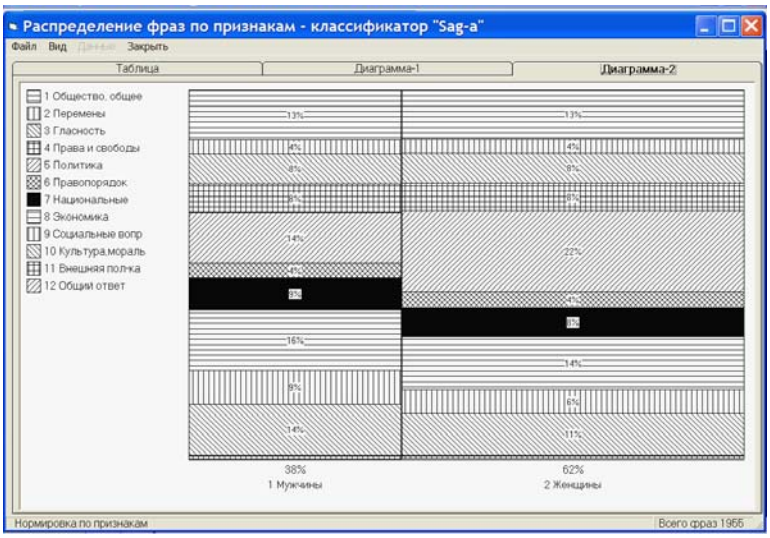

Рис. 1. Двумерная гистограмма распределения ответов по 12 классам для мужчин и женщин

Форма СТАТ.АНАЛИЗ обеспечивает возможность статистического анализа ответов респондентов на закрытые или полузакрытые вопросы анкеты. Статистическая обработка данных в рассматриваемой системе производится только для полей целого или смешанного типа. Вывод результатов обеспечивается в виде таблиц и диаграмм. Предварительно необходимо определить максимальное количество рангов (градаций) для каждого из этих полей. В ряде случаев данные, например возраст респондента, могут иметь значительный разброс. В таких случаях можно воспользоваться шкалированием, разделив весь диапазон изменения данных на ограниченное число градаций (рис. 2). Затем можно приступать к статистическому анализу ответов на одноальтернативные вопросы. При одномерном и двухмерном статистическом анализе результаты выводятся в виде таблиц и диаграмм, кроме того, подсчитывается ряд дополнительных параметров. При трехмерном статистическом анализе результаты выводятся в виде многомерной таблицы, в которой для каждого имеющегося набора рангов выводятся количество анкет и их процентное содержание. При многоальтернативном статистическом анализе результаты выводятся в виде двух многомерных таблиц.

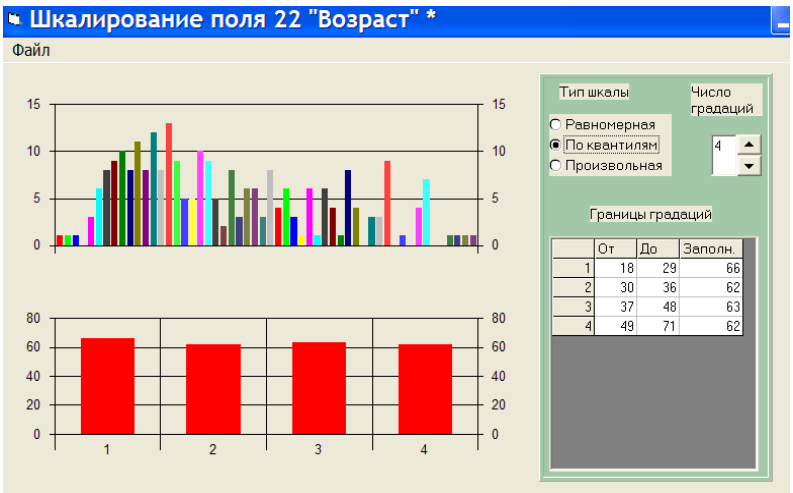

Рис. 2. Шкалирование поля «Возраст»

Форма СЕРВИС обеспечивает задание параметров системы, которые сохраняются и после выхода из нее. Можно задавать кодировку для импортируемой информации, тип пермутационного вывода, шрифты для текста и отдельно для пермутационного вывода, а также веса простых и ключевых слов и уровень отсечки фраз.

Дополнительные возможности системы

Система позволяет анализировать и отдельные тексты, не имеющие жесткой структуры – эссе, биография и т.п. [10]. Их анализ может быть осуществлен двумя путями.

Для свободных текстов можно создать БД, состоящую из одного поля символьного типа. При этом в одной записи можно разместить до 32 Кбайт текста. В этом случае каждый абзац текста будет соответствовать одной записи.

Для текстов, не имеющих жесткой структуры (типа автобиографий), можно создать БД, состоящую из нескольких полей символьного типа, сопоставив каждому из них узкую тематическую направленность (отношения в семье, отношение к искусству и т.п.). Импортируемый текст в этом случае может состоять из ряда разделов (параграфов), каждый из которых соответствует одной записи. Параграф состоит из нескольких абзацев, каждый из которых соответствует какому-либо полю записи в БД. Каждый такой абзац начинается с указателя номера поля. Номера полей могут быть расположены в любом порядке и могут даже повторяться в пределах одной записи. В последнем случае все абзацы, имеющие одинаковый номер поля, будут помещены в одно поле.

Даже для текстов такого рода система ВЕГА позволяет строить разнообразные словари и классифицировать информацию с последующей статистической обработкой.

Литература

1. Coxon A.R.M., Trappes-Lomax H.R.N. INQUIERER III (Edinburg's version). – Edinburg Univer., Jan.1977, Rep. 92. – 43 p.
2. Popping Roel. Computer-Assisted Text Analysis. – London: Sage Publications, 2000. – 240 p.
3. Сбойчаков К.О. Автоматизированная система классификации текстов на основе визуального эвристического кластерного анализа [Электронный ресурс]. – Режим доступа: http://www.gpntb.ru/win/elbib/trud3/f1_06.htm, свободный.
4. Компьютеризированный анализ текста [Электронный ресурс]. – Режим доступа: http://edu.tsu.ru/historynet/informatika/posobia/istgf_kleo/analis.htm, свободный.
5. Каневский Е.А. Методы классификации текста // Труды Международного семинара Диалог’98 по компьютерной лингвистике и ее приложениям. – Казань: ООО «Хэтер», 1998. – С. 488–497.
6. Коробейников В.С. Методы качественно-количественного анализа содержания документов // Методы анализа документов в социологических исследованиях. – М.: ИСИ АН СССР, 1985. – С. 10–66.
7. Каневский Е.А., Саганенко Г.И., Гайдукова Л.М., Клименко Е.Н. Система анализа текстов // Социология: 4М. – 1997. – № 9. – С. 198–216.
8. Каневский Е.А., Лезин Г.В. Анализ текстов // Экономико-математические исследования: математические модели и информационные технологии. Вып. 2. – СПб: ЗАО «Центр стратегических анализов общественных процессов», 2001. – С. 260–285.
9. Боярский К.К., Каневский Е.А. Вега – система для работы с текстами // Экономико-математические исследования: математические модели и информационные технологии. Вып. 6. – СПб: Нестор История, 2008. – С. 184–200.

10. Саганенко Г.И., Каневский Е.А. Боярский К.К. Контексты эмпирического познания в социологии и возможности программы ВЕГА // Телескоп. – 2008. – № 6. – С. 43–55.

Боярский Кирилл Кириллович – Санкт-Петербургский государственный университет информационных технологий, механики и оптики, кандидат физ.-мат. наук, доцент, boyarin9@yandex.ru

Каневский Евгений Александрович – Санкт-Петербургский экономико-математический институт РАН, кандидат технических наук, вед. научный сотрудник, kanev@emi.nw.ru