УДК 655.5

Шпаковский Ю. Ф., кандидат филологических наук, доцент (БГТУ)

РАЗРАБОТКА ЭФФЕКТИВНЫХ МЕТОДОВ ПОИСКА ИНФОРМАЦИИ

В статье приведены результаты автоматической классификации документов. Для достижения цели выбрана функционально-стилевая концепция, в соответствии с которой выделено пять функциональных стилей речи (официально-деловой, научный, публицистический, художественный и разговорный) и сформирован опытный массив документов для анализа. На основании изучения литературы сформирован набор параметров классификации, который в дальнейшем использовался в дискриминантном анализе для классификации документов. Проведенный анализ позволил получить дискриминантную функцию. Ошибка классификации составляет менее 15%. Поиск оптимальных текстов с точки зрения их трудности для читателей позволит использовать информационные каналы более эффективно.

In article results of automatic classification of documents are resulted. For purpose achievement the is functional-style concept according to which it is allocated five functional styles of speech (officially-business, scientific, publicistic, art and colloquial) is chosen and the skilled document file for the analysis is generated. On the basis of literature studying the set of parameters of classification which was used further in the discriminant analysis for classification of documents is generated. The carried out analysis has allowed to receive discriminant function. The classification error makes less than 15%. Search of optimum texts from the point of view of their difficulty for readers will allow to use information channels more effectively.

Введение. На этапе создания и редактирования текста первостепенное значение имеют анализ и оценка трудности восприятия текста будущими читателями. Но для конечного потребителя информации важнейшую роль играет нахождение необходимой информации. Только после наличия у читателя нужного материала встает вопрос об адекватном его восприятии. В настоящее время следует учитывать тот факт, что наряду с печатными изданиями как средство коммуникации и неограниченного доступа к информационным ресурсам достаточно активно используется глобальная сеть интернет.

Всемирная сеть достаточно активно используется и в научной среде Последнее время интернет находит применение и как средство публикации научных трудов. Появились веб-сайты различных научных сообществ, высших учебных заведений, тематические страницы по разным научным дисциплинам, а также сайты научных журналов. Уже сегодня начинают появляться виртуальные лаборатории, где не только студенты, но и все желающие могли бы проводить различные эксперименты, ставить опыты, выбирая параметры и модели экспериментов по своему усмотрению. Более того, в последние годы появляются виртуальные кафедры и виртуальные университеты. Определенная часть образования принимает дистанционную форму.

Таким образом, потоки информации возрастают в геометрической прогрессии. И как точно отмечено одним из авторов, по мере развития интернета «вероятность существования нужной информации возрастает, а возможность ее нахождения уменьшается» [1].

С учетом данного парадокса цель исследования — разработка методов повышения эффективности поиска информации, в частности автоматической классификации документов. Для достижения поставленной цели необходимо решить следующие задачи:

- разработать процедуру автоматической классификации документов;
- создать схему машины поиска с использованием стилистического анализа;
- разработать макетные версии программы для классификации документов;
- протестировать разработанный программный продукт;
- выработать рекомендации по практическому использованию программы.

В данной статье приведены результаты автоматической классификации документов.

Основная часть. Существующие поисковые машины глобальной сети реализуют идеи, которые были сформулированы еще в 1970-х гг. для локальных информационно-поисковых систем (ИПС) [2].

Наиболее простой и распространенной является линейная модель поиска [2, 3]. Основными понятиями модели являются словарь, документ, база. Словарь — это упорядоченное множество терминов мощности D. Документ (поисковый образ документа) — двоичный вектор размерности D. Если термин входит в документ, то в соответствующем разряде вектора стоит 1, в противном случае — 0. База L — это матрица $N \times D$, строки которой соответствуют N документам. Тогда процедуру обработки запроса можно представить следующим образом: $L \times$

$\times q = r$, где q — вектор запроса, а r — отклик системы на запрос.

Усложнение модели происходит путем приписывания терминам документа и запроса весов, отражающих их значимость.

В качестве критериев оценки эффективности функционирования ИПС используются следующие [2]:

1. Полнота поиска — доля релевантных документов массива, присутствующая в выдаче.

2. Точность поиска — доля релевантных документов в выдаче.

3. Длина поиска — это среднее число нерелевантных документов, которые должен просмотреть пользователь, прежде чем будут просмотрены все релевантные.

4. Время реакции системы.

5. Форма представления результатов поиска.

6. Полнота массива, т. е. степень охвата всех релевантных документов, интересующих пользователей.

Современные информационно-поисковые системы (Google, Рамблер, Яндекс и др.) эффективно справляются с большинством запросов, предлагая поиск информации по следующим признакам: ключевые слова, язык, регион, формат файлов, дата, расположение слов, сайт или домен, права использования.

Однако для эффективного поиска необходимого материала этих параметров явно недостаточно.

Для повышения эффективности поиска информации предлагается на научной основе (используя статистические методы) дополнить список следующими признаками: функциональный стиль (научный, официально-деловой и др.), жанр (статья, репортаж, интервью и т. д.), вид издания по целевому назначению (научное, справочное, учебное издание и т. д.), читательский адрес (для взрослого читателя, для детей дошкольного, младшего, среднего или старшего школьного возраста).

Ниже приведены результаты классификации документов в соответствии с функциональным стилем.

Первые работы, демонстрирующие применение статистических методов в стилистике, появились на рубеже XIX и XX вв. Примером может служить стилистическое исследование некоторых грамматических форм на материале художественных произведений, проведенное А. Марковым [4].

Первое использование компьютера для исследования литературного стиля относится к началу 1950-х гг. [5]. Роберт Буза (Robert Busa) использовал вычислительную технику для создания конкорданса трудов Фомы Аквинского.

В 1960-е годы математические методы активно проникали в языкознание, в том числе в стилистику, одновременно вызывая горячие споры о правомерности и границах их применения.

В 1964 г. д-р филол. наук, проф. Б. Н. Головин, например, писал: «...едва ли можно получить убедительную картину развития русского литературного языка на протяжении XIX–XX вв. без обращения к вероятностно-статистическому многостороннему изучению языка в его дифференциации по стилям: ведь история литературного языка — это прежде всего история формирования и движения его стилей и их речевого воплощения в различных типах речи» [6].

В статье «К вопросу о применении статистики в языкознании» д-р филол. наук Л. Р. Зиндер также писал: «Употребительность тех или иных слов, грамматических конструкций, объем тех или иных синтаксических единиц оказывается в прямой зависимости от содержания текста, от жизненной ситуации, от цели высказывания и намерения автора, от адресата речи и от ряда других факторов, учесть которые практически невозможно. Только статистическая структура текста, как некая равнодействующая, дает возможность отнести тот или иной текст к соответствующему функциональному стилю (подъязыку), соотнести его хронологически, определить автора и т. д.» [7].

Примеры использования статистических методов в стилистике можно найти и в других более поздних работах [8–14].

Для решения нашей задачи следовало, прежде всего, выбрать функционально-стилевую концепцию. Опираясь на труды Б. Н. Головина, Т. В. Строевой и других исследователей, было выделено пять функциональных стилей речи: официально-деловой, научный, публицистический, художественный и разговорный.

В соответствии с концепцией был сформирован опытный массив документов для анализа.

Официально-деловой стиль представлен в опытном массиве текстами пяти законов Республики Беларусь последнего времени:

1. О поддержке малого и среднего предпринимательства. Закон Республики Беларусь от 01.07.2010 г. № 148-З.

2. О товарных биржах. Закон Республики Беларусь от 05.01.2009 г. № 10-З.

3. О кредитных историях. Закон Республики Беларусь от 10.11.2008 г. № 441-З.

4. Об ипотеке. Закон Республики Беларусь от 20.06.2008 г. № 345-З.

5. О государственном регулировании внешнеторговой деятельности. Закон Республики Беларусь от 25.11.2004 г. № 347-З.

Документы были отобраны из юридической базы данных портала pravo.by.

Коллекция документов научного стиля состояла из 10 статей по различным отраслям знаний:

1. Влияние катионной нестехиометрии на термохимические и транспортные характеристики фаз в системах SM(ND)—BA—CU—O / Н. И. Мацкевич [и др.] // Исследовано в России: в 15 т. — 2002. — Т. 5. — С. 1919–1929.

2. Сафонов, М. А. Ресурсный потенциал биоты ксилотрофных грибов / М. А. Сафонов // Вестник ОГУ. — 2005. — № 9. — С. 159–163.

3. Макаренко, П. В. Германский фактор в Октябрьской революции 1917 г. / П. В. Макаренко // Вопросы истории. — 2008. — № 5. — С. 30–45.

4. Малкова, Е. Е. Возрастная динамика проявлений тревожности у школьников / Е. Е. Малкова // Вопросы психологии. — 2009. — № 4. — С. 24–32.

5. Смирнова, А. Г. Угрозы и их изучение в социологии международных отношений / А. Г. Смирнова // Социологический журнал. — 2010. — № 2. — С. 35–50.

6. Кудрявцев, А. В. Повышение информативности измерений вибрации системой погружной телеметрии / А. В. Кудрявцев // Нефтегазовое дело. — 2011. — № 2. — С. 4–15.

7. Казаковская, В. В. Реактивные реплики взрослого и усвоение ребенком грамматики родного языка / В. В. Казаковская // Вопросы языкознания. — 2010. — № 3. — С. 3–29.

8. Власенко, А. А. Аномальный ход концентрационной зависимости электропроводности водного раствора NaCl в области малых содержаний электролита при бесконтактных измерениях проводимости / А. А. Власенко, В. И. Ермаков // Исследовано в России: в 15 т. — 2004. — Т. 7. — С. 667–672.

9. Микешина, Л. А. Эпистемологическое оправдание гипостазирования и реификации / Л. А. Микешина // Вопросы философии. — 2010. — № 12. — С. 44–54.

10. Моисеев, С. В. О реализации автоматов нейронными сетями / С. В. Моисеев // Интеллектуальные системы. — 2008. — № 1–4. — С. 283–316.

В опытный массив документов публицистического стиля вошли статьи на различные темы, опубликованные в последнее время на трех порталах: tut.by, naviny.by, belta.by.

Художественный стиль представлен произведениями с сайта lib.ru. Были отобраны следующие произведения: М. С. Бубеннов «Белая береза» (роман); А. Н. Толстой «Егор Абозов» (роман); Н. С. Лесков «На ножах» (роман); В. В. Ершов «Страх полета» (повесть); Б. Е. Штейман «Путешествие для одного» (повесть); Н. С. Лесков «Смех и горе» (повесть); Е. И. Замятин «Три дня» (рассказ); А. И. Куприн «Впотьмах» (повесть); И. А. Бунин «Легенда» (рассказ); А. И. Куприн «Просительница» (рассказ); Л. Н. Толстой «Метель» (рассказ).

Основной объем текстов разговорного стиля был отобран с форумов портала tut.by.

На основании изучения литературы и опытного массива документов был сформирован следующий набор параметров классификации:

1) средняя длина абзаца в словах;
2) средняя длина абзаца в буквах;
3) средняя длина абзаца в печатных знаках;
4) средняя длина предложения в словах;
5) средняя длина предложения в слогах;
6) средняя длина предложения в буквах;
7) средняя длина предложения в печатных знаках;
8) средняя длина слов в слогах;
9) средняя длина слов в буквах;
10) средняя длина слов в печатных знаках;
11) средняя длина слов по Деверу;
12) процент слов длиной в 5 букв и больше;
13) процент слов длиной в 6 букв и больше;
14) процент слов длиной в 7 букв и больше;
15) процент слов длиной в 8 букв и больше;
16) процент слов длиной в 9 букв и больше;
17) процент слов длиной в 10 букв и больше;
18) процент слов длиной в 11 букв и больше;
19) процент слов длиной в 12 букв и больше;
20) процент слов длиной в 13 букв и больше;
21) процент слов в 3 слога и больше;
22) процент слов в 4 слога и больше;
23) процент слов в 5 слогов и больше;
24) процент слов в 6 слогов и больше;
25) процент неповторяющихся слов;
26) средняя частота повторения слова;
27) процент существительных;
28) процент конкретных существительных;
29) процент абстрактных существительных;
30) процент прилагательных;
31) процент глаголов;
32) процент местоимений;
33) процент междометий;
34) процент предлогов;
35) процент союзов;
36) процент частиц;
37) процент наречий;
38) процент причастий;
39) процент деепричастий;
40) процент числительных;
41) процент возвратных глаголов;
42) процент сложных предложений;
43) процент простых предложений;
44) процент предложений с экспрессивной пунктуацией;
45) процент терминов;
46) количество иллюстраций;
47) количество таблиц;
48) процент условных обозначений.

Значения признаков исследуемых текстов были сведены в таблицу, фрагмент которой представлен ниже (табл. 1).

Таблица 1

Значения признаков исследуемых текстов

Номер признака	Стиль				
	научный	официальный	публицистический	разговорный	литературно-художественный
1	53,25	26,88	43,40	27,09	50,83
2	375,50	202,38	297,60	126,82	253,50
3	426,75	228,63	340,40	156,73	304,33
4	19,36	35,83	27,13	8,05	21,79
5	129,73	262,33	178,63	58,81	81,57
6	136,55	269,83	186,00	37,70	108,64
7	155,18	304,83	212,75	46,59	130,43
8	2,92	2,98	2,70	1,94	2,10
9	7,05	7,53	6,86	4,68	4,99
10	8,01	8,51	7,84	5,79	5,99
11	1,14	1,13	1,14	1,24	1,20
12	75,59	69,30	75,12	46,31	55,74
13	71,36	65,58	62,67	35,57	42,95
…	…	…	…	…	…
48	1,20	0,00	0,00	0,00	0,00

Так как характеристики текста измерялись в различных единицах, то все данные были стандартизированы. Для этого использовалась нормализация, приводящая все переменные к стандартной z-шкале. Стандартизированные переменные исследуемых текстов были сведены в таблицу, фрагмент которой представлен ниже (табл. 2).

Использование большого количества параметров является неэффективным по ряду причин: а) сильная взаимосвязанность признаков, что приводит к дублированию информации; б) неинформативность признаков, мало меняющихся при переходе от одного объекта к другому; в) возможность агрегирования, т. е. простого или «взвешенного» суммирования, по некоторым признакам. В связи с этим следует перейти к существенно меньшему числу наиболее информативных переменных.

Таблица 2

Стандартизированные переменные текстов

Номер признака	Стиль				
	научный	официальный	публицистический	разговорный	литературно-художественный
1	0,081	–0,083	0,019	–0,082	0,066
2	0,014	–0,006	0,005	–0,014	0,000
3	0,013	–0,006	0,005	–0,013	0,001
4	–0,029	0,128	0,045	–0,138	–0,006
5	–0,002	0,018	0,005	–0,013	–0,009
6	–0,001	0,016	0,005	–0,015	–0,005
7	–0,002	0,015	0,005	–0,013	–0,004
8	1,711	1,973	0,751	–2,566	–1,868
9	0,495	0,781	0,379	–0,919	–0,737
10	0,503	0,818	0,394	–0,922	–0,793
11	–15,082	–18,199	–11,710	30,617	14,374
12	0,067	0,029	0,064	–0,109	–0,052
13	0,066	0,042	0,029	–0,084	–0,053
…	…	…	…	…	…
48	3,333	–0,833	–0,833	–0,833	–0,833

С этой целью для выделенных параметров текстов была построена корреляционная матрица, фрагмент которой представлен в табл. 3. В дальнейшем из наиболее связанных признаков (коэффициенты корреляции данных признаков выделены цветом) в учет принимался только один параметр. Например, коэффициент корреляции между признаками 9 (средняя длина слов в буквах) и 10 (средняя длина слов в печатных знаках) равен единице, поэтому в анализ включался только признак 9.

Таблица 3

Корреляционная матрица между признаками текстов

Номер признака	1	2	3	4	5	6	7	8	9	10	11	…	48
1	1,000	0,854	0,887	–0,042	–0,290	–0,142	–0,135	0,115	0,038	0,014	–0,221	…	0,571
2	0,854	1,000	0,998	0,252	0,147	0,245	0,243	0,592	0,528	0,510	–0,657	…	0,738
3	0,887	0,998	1,000	0,213	0,089	0,195	0,194	0,538	0,471	0,452	–0,609	…	0,730
4	–0,042	0,252	0,213	1,000	0,918	0,974	0,979	0,735	0,773	0,766	–0,800	…	–0,168
5	–0,290	0,147	0,089	0,918	1,000	0,981	0,977	0,838	0,885	0,888	–0,837	…	–0,086
6	–0,142	0,245	0,195	0,974	0,981	1,000	1,000	0,835	0,874	0,871	–0,864	…	–0,072
7	–0,135	0,243	0,194	0,979	0,977	1,000	1,000	0,825	0,864	0,861	–0,858	…	–0,086
8	0,115	0,592	0,538	0,735	0,838	0,835	0,825	1,000	0,993	0,991	–0,982	…	0,458
9	0,038	0,528	0,471	0,773	0,885	0,874	0,864	0,993	1,000	1,000	–0,979	…	0,359
10	0,014	0,510	0,452	0,766	0,888	0,871	0,861	0,991	1,000	1,000	–0,973	…	0,352
11	–0,221	–0,657	–0,609	–0,800	–0,837	–0,864	–0,858	–0,982	–0,979	–0,973	1,000	…	–0,393
…	…	…	…	…	…	…	…	…	…	…	…	…	…
48	0,571	0,738	0,730	–0,168	–0,086	–0,072	–0,086	0,458	0,359	0,352	–0,393	…	1,000

Кроме того, на основании анализа значений признаков опытного массива все параметры были разделены на три группы:

1) параметры, средние значения которых слабо отличаются для разных стилей (процент союзов, предлогов, числительных, возвратных глаголов и др.);

2) параметры, средние значения которых относительно плавно меняются от стиля к стилю (процент слов и предложений разной длины, процент глаголов, существительных, прилагательных и др.);

3) параметры, значения которых резко выделяются в одном-двух стилях и примерно равны в других (средняя длина абзаца в словах, средняя длина предложения в словах, средняя длина слов в слогах, процент терминов, междометий и др.).

Например, использование большого количества терминов, нетекстовых элементов, абстрактных существительных характерно для научного стиля. Употребление коротких слов и предложений, конкретных существительных, междометий, предложений с экспрессивной пунктуацией характерно для разговорного стиля. В научном и официально-деловом стилях преобладают сложные предложения.

Предложения с экспрессивной пунктуацией, местоимения «я», «ты, «вы», частицы «ну», «вот», «ведь» наиболее ярко проявляются в разговорном стиле, слабее — в художественном стиле и практически отсутствуют в публицистическом, научном и официально-деловом стилях.

В дальнейшем параметры, средние значения которых слабо отличаются для разных стилей, из анализа исключались.

В качестве метода построения классификации был использован дискриминантный анализ, который на основании некоторых признаков (в нашем случае параметров текста) позволяет предсказать принадлежность объектов к двум или более непересекающимся группам.

Проведенный анализ позволил получить дискриминантную функцию, с использованием которой можно эффективно сортировать документы в соответствии с функциональным стилем (ошибки составляют менее 15%). Хуже всего классифицируются документы литературно-художественного и публицистического стилей, лучше всего — документы научного, официально-делового и разговорного стилей (около 5% ошибок).

Выводы. Автоматизация процесса поиска и классификации документов в соответствии с функциональным стилем обеспечит поисковым машинам нахождение оптимальных текстов с точки зрения не только их полезности, но и трудности для читателей. Это позволит использовать любые информационные каналы более эффективно.

Литература

1. Кешелава, В. Поисковые системы для Интернет / В. Кешелава // PC Week/RE. — 1997. — № 10. — С. 22–27.

2. Солтон, Дж. Динамические библиотечно-информационные системы / Дж. Солтон; пер. с англ. — М.: Мир, 1979. — 558 с.

3. Храмцов, П. Моделирование и анализ работы информационно-поисковых систем Internet / П. Храмцов // Открытие системы. — 1996. — № 6 (20). — С. 46–56.

4. Марков, А. А. Об одном применении статистического метода / А. А. Марков // Известия Императорской Академии наук. Сер. 6, т. X. — 1916. — № 4. — С. 239–242.

5. Allen, R. F. Computer-Aided Stylistic Analysis. A Case Study of French Texts / R. F. Allen // Computational Linguistics. An International Handbook on Computer Oriented Language Research and Applications. — Berlin: Walter de Gruyter, 1989. — P. 544–552.

6. Головин, Б. Н. О вероятностно-статистическом изучении стилевой дифференциации языка / Б. Н. Головин. — К.: Знание, 1964. — 21 с.

7. Зиндер, Л. Р. К вопросу о применении статистики в языкознании / Л. Р. Зиндер, Т. В. Строева // Вопросы языкознания. — 1968. — № 6. — С. 120–123.

8. Андреев, Н. Д. Статистико-комбинаторные методы в теоретическом и прикладном языковедении / Н. Д. Андреев. — Л.: Наука, 1967. — 403 с.

9. Головин, Б. Н. Язык и статистика / Б. Н. Головин. — М.: Просвещение, 1970. — 190 с.

10. Кауфман, С. И. Из курса лекций по статистической стилистике / С. И. Кауфман. — М.: МОПИ, 1970. — 319 с.

11. Кожина, М. Н. О речевой системности научного стиля сравнительно с некоторыми другими / М. Н. Кожина. — Пермь, 1972. — 395 с.

12. Осиновский, Д. Э. Компьютеризованный текстологический анализ исторических документов: возможности программы НТА / Д. Э. Осиновский, А. С. Ровный, Д. В. Новицкий // [Электронный ресурс]. Режим доступа: http://kleio.dcn-asu.ru/aik/bullet/22/23.shtml. — Дата доступа: 05.03.2011.

13. Применение статистических методов для интеллектуальной компьютерной обработки текстов / И. С. Ашманов [и др.] // Диалог'97: труды Междунар. семинара по компьютерной лингвистике и ее приложениям. — Ясная Поляна, 1997. — С. 33–37.

14. Хмелев, Д. Лингвоанализатор [Электронный ресурс]. Режим доступа: http://www.rusf.ru/books/analysis . — Дата доступа: 22.02.2011.

Поступила 06.04.2011