Математические структуры и моделирование
1999. Вып. 3, с.134-140.

УДК 025.4.036:681.3

РАСПОЗНАВАНИЕ СТИЛЕЙ РЕЧИ ПРИМЕНИТЕЛЬНО К ИНФОРМАЦИОННОМУ ПОИСКУ: ПОСТАНОВКА ЗАДАЧИ

П.И. Браславский

The problems of information overload have become more pressing with the increasing popularity of Internet. The main information retrieval mechanisms in the Net are based on keyword search. The variety of Internet documents often makes the technique ineffective and suggests an idea of using a functional genre attribute as an additional dividing parameter. A problem definition and a framework for implementation of such method are presented. The proposed method for recognizing text genres allows to improve Internet keyword searching.

1. Введение

Глобальная сеть Internet — один из самых интересных феноменов нашего времени. Стремительное развитие Сети кроме новых возможностей, достижений и успехов имеет множество проблем: научных, технических, социальных, психологических и др. Одна из важнейших — это поиск информации на безбрежных просторах Internet.

В настоящее время наиболее распространенным в Internet является поиск по ключевым словам. Мы предлагаем дополнить поиск по ключевым словам стилистическим анализом обнаруженных документов. По нашему мнению, отнесение текста документа к одному из стилей речи сделает поиск более эффективным. В данной работе содержится постановка задачи и схема реализации.

2. Интернет и поиск информации

Сегодня Internet, или «Сеть» с большой буквы, является синонимом Мировой Паутины (WorldWide Web, WWW). Временем рождения WWW принято считать 1990 год. Именно тогда в Европейской лаборатории физики элементарных частиц (CERN) в Женеве были созданы язык разметки (HyperText Markup

E-mail: pb@dpt.ustu.ru
Уральский государственный технический университет

Language, HTML) и протокол их передачи (HyperText Transfer Protocol, HTTP). Вторым «родителем» можно считать Национальный центр суперкомпьютерных приложений (National Centre for Supercomputing Applications, NCSA), где в 1993 году была создана первая программа-броузер «Mosaic». Новый способ представления информации обеспечивал не только удобный доступ, но и простой способ комбинирования информационных блоков друг с другом, а использование графики сделало внешний вид документов более привлекательным. С появлением Мировой Паутины достаточно закрытая до этого сеть Internet привлекла внимание миллионов потенциальных пользователей.

За свою короткую историю Сеть уже успела пройти несколько этапов. На первом, зародышевом этапе пространство Web было достаточно обозримым, и поиск информации осуществлялся с помощью простого «листания», т.е. перехода по ссылкам к релевантным документам в надежде найти нужную информацию. Та же процедура подходила для того, чтобы превратить Internet в развлечение, игру. Пользователей завораживал сам процесс перехода по ссылкам (что получило название *net surfing*) [9]. С развитием Мировой Паутины все более разнообразным становилось ее наполнение: от научных докладов и правительственных сообщений до кулинарных рецептов и анекдотов.

Сегодня Internet предоставляет универсальный способ доступа к огромному количеству документов, и больше, чем раньше, пользователей интересует осмысленный и точный поиск информации. Отражением этого процесса является развитие и совершенствование поисковых средств Internet. Примерами могут служить специальные серверы-«порталы» для входа в Сеть: Yahoo, Lycos и др. На этих серверах располагаются рубрикаторы и машины поиска, кроме того, пользователь может завести там личную страничку, электронный почтовый ящик и подписаться на извещения о сетевых новинках по определенной теме. Информационный поиск тесно связан с национальным языком, это дает импульс для развития локальных средств поиска. Не так давно в русскоязычном сегменте сети Internet появились такие проекты, как «Апорт», «Следопыт» [1], «Яndex». Основным методом поиска информации в Сети является поиск по ключевым словам.

Массив документов Web очень разнороден: по тематике, объему, структуре, стилю, языку. Не менее пестрой является армия пользователей Internet: по образовательному уровню, социальному положению, владению родным и иностранными языками; а также по запросам и ожиданиям, связанным с Internet. Это определяет недостатки поиска по ключевым словам, несмотря на совершенствование технологии поиска и развитие синтаксиса языка запросов [1, 9]. По тем же причинам попытки внедрить в качестве универсального достаточно мощный (а потому сложный) язык информационных запросов представляются неосуществимыми, а факультативные возможности существующих языков часто остаются невостребованными.

Одним из выходов в этой ситуации может быть тематическая рубрикация документов. В этом случае автор или эксперт должен отнести текст к одному или сразу нескольким разделам рубрикатора. В [9] предлагается система автоматического разделения информационного пространства документов на кате-

гории с помощью самоорганизующейся нейронной сети. После разбиения поиск осуществляется по ключевым словам внутри тематического раздела.

Наше предложение имеет другой характер и состоит в том, чтобы документы, найденные по ключевым словам, автоматически относить к одному из четырех-пяти стилей речи. Такой подход частично пересекается с тематической классификацией: тексты на некоторые темы могут появляться только в определенных стилях; однако большинство тем допускает раскрытие во множестве стилей. Нам представляется, что введением дополнительного дивидивного признака — стиля — поиск по ключевым словам может быть существенно усилен.

Сходное предложение было выдвинуто в работе [10], однако с выбором других методологических оснований (стилевая концепция, параметры стилей) и класса документов (телеконференции новостей USENET).

3. Стили речи и статистика

Тексты могут быть очень разными. Даже тексты на одну тему, посвященные одному предмету, могут быть написаны в разных стилях, жанрах, манерах.

Понятие «стиль» в лингвистике многозначно. В Лингвистическом энциклопедическом словаре [5] читаем: «Стиль (от лат. stilus, stylus — остроконечная палочка для письма) в языкознании — 1) разновидность языка, закрепленная в данном обществе традицией за одной из наиболее общих сфер социальной жизни и частично отличающаяся от других разновидностей того же языка по всем основным параметрам — лексикой, грамматикой, фонетикой; то же, что стиль языка. В современных развитых национальных языках существует три наиболее крупных стиля языка в этом значении: а) нейтральный, б) более «высокий», книжный, в) более «низкий», разговорный; 2) то же, что функциональный стиль; 3) общепринятая манера, обычный способ исполнения какого-либо конкретного типа речевых актов: ораторская речь, бытовой диалог, дружеское письмо и т.д.: стиль в этом смысле характеризуется не только набором (параметрами) языковых средств, но и композицией акта; 4) индивидуальная манера, способ, которым исполнены данный речевой акт или произведение, в т.ч. литературно-художественное; 5) то же, что языковая парадигма эпохи, состояние языка в стилевом отношении в данную эпоху».

Современные текстовые редакторы осуществляют некоторый стилистический анализ текста. Здесь «стиль» понимается в первом значении определения, как «стиль языка». В состав редактора входят словари, где слова имеют пометы «разговорный», «поэтический» или подобные. На основании этих помет редактор вырабатывает рекомендации по стилистическому «единству» документа по части лексики.

Функционально-стилевая концепция (соответствует пункту 2 определения) утверждает экстралингвистическую, социальную основу расслоения литературного языка. Исходным положением концепции является зависимость стиля речи от выполняемой им коммуникативно-общественной функции, от задач общения в соответствующей сфере. Эта концепция восходит к трудам

В.В.Виноградова, Ш.Балли, ученых Пражского лингвистического кружка. В 60-80-е годы в советской русистике было разработано развернутое обоснование этой концепции [3, 4].

Обычно различают пять функциональных стилей речи: научный, художественный, деловой, публицистический, разговорный. На сегодняшний день функционально-стилевая концепция в отечественном языкознании является наиболее разработанной и обоснованной. Использование функционально-стилевой системы, ее четырех-пяти категорий, представляется нам наиболее плодотворным и для задач информационного поиска.

(Однако в литературе можно найти и другие подходы. Так, в статье [10], которую можно рассматривать как начальный импульс для нашей работы, используется система из пятнадцати жанровых категорий, что больше соответствует пункту 3 в определении понятия «стиль».)

Статистика достаточно давно и широко используется для изучения стилистических особенностей речи [2, 4, 7, 8]. В теоретической стилистике применение статистических методов традиционно преследует следующие цели: вскрыть закономерности функционирования языка в различных сферах коммуникации, обогатить знания о стилистических нормах этих речевых разновидностей, глубже рассмотреть проблему мышления и речи, их взаимосвязь [4]. К этому направлению примыкают литературоведческие по духу исследования индивидуальных стилей отдельных авторов (см. пункт 4 определения понятия «стиль») [7, 8]. Более прикладной характер носят работы, относящиеся к нормативной стороне стилистики или к направлению «машинный перевод». Обычно применение статистических методов в стилистике означает подсчет количественных характеристик конкретных текстов, обработку результатов, последующее сравнение и качественные выводы. Наша задача является обратной: разработать автоматическую процедуру, на основании которой по количественным характеристикам произвольного текста он мог бы быть отнесен к одному из стилей речи.

4. Метод и аппарат

Разработка стилистического анализатора документов Internet для повышения эффективности поиска информации, на наш взгляд, должна включать в себя следующие этапы:

1. Отбор опытного массива документов.
2. Отбор параметров классификации.
3. Выбор и настройка метода классификации.
4. Оптимизация процедуры классификации.
5. Тестирование, проверка результатов.

Поясним содержание каждого из этапов.

На первом этапе производится отбор документов Сети, которые послужат опытным материалом, сырьем для разработки. Можно предложить три принципа для отбора: документы, относящиеся к предопределенным функциональным стилям речи (художественный, официально-деловой, публицистический,

научный и разговорный стили); группы документов, выдаваемые машинами поиска в ответ на достаточно общий запрос; случайный набор документов.

На втором этапе составляется набор параметров документов, по которому будет производиться классификация. Два фактора являются решающими для включения параметра в первичный набор: легкая вычислимость и потенциальная значимость для задач стилистической классификации. Первое требование вытекает из прикладного характера задачи. На практике это означает, что параметры берутся с «нижних» уровней языковой системы: графики, лексики, морфологии. Синтаксические параметры можно получить по косвенным легковычислимым признакам: например, по количеству знаков экспрессивной пунктуации или отдельных союзов. На этом этапе необходимо создать набор параметров «с запасом», так как последующие процедуры позволят выделить значимые и удалить лишние. Несмотря на то, что мы заранее не беремся учитывать многие параметры, несущие стилистическую информацию, можно надеяться на приемлемое для практических целей качество классификации. Такую уверенность дает системность стиля речи [4] (синхронное «проявление» стиля на всех уровнях языка), а также успешные опыты подобного рода с аналогичным подходом к отбору параметров [10].

На этапе 3 необходимо принять решение о выборе конкретного метода классификации. Для классификации текстов по их количественным характеристикам можно применять кластер-анализ (классификация без учителя) или дискриминантный анализ (классификация с учителем) [6]. Так, например, кластер-анализ использовался для выяснения степени близости индивидуальных стилей нескольких авторов [7], а дискриминантный анализ — для классификации содержания телеконференций USENET [10].

Нетрудно подобрать «образцовые» тексты для каждого функционального стиля, которые составили бы обучающую выборку в процедуре дискриминантного анализа. Однако нам представляется небезынтересным для начала использовать кластер-анализ для проверки того, насколько система пяти функциональных стилей подходит для такой специфичной сферы функционирования языка, как Internet. После этого, определившись с количеством стилей (классов), можно обратиться к методам дискриминантного анализа.

Кроме того, на третьем этапе предстоит выбор метрики для процедуры классификации. Метрика должна учитывать: различные шкалы измерения параметров; предварительные оценки важности отдельных параметров; стилевую системность (взаимосвязь параметров).

На этапе оптимизации процедуры классификации применяются методы снижения размерности пространства параметров и выявления наиболее информативных признаков [6].

На завершающем этапе производится проверка процедуры классификации. При неудовлетворительных результатах может понадобиться возврат к одному из предыдущих этапов или изменение формулировки задачи.

5. Заключение

При нашем подходе основной целью разработки автоматического стилистического анализатора является повышение эффективности поиска информации по ключевым словам. Наряду с этим в ходе выполнения работ могут быть получены результаты, ценные и с точки зрения теоретического языкознания. К ним относятся:

1. Новый взгляд на функционально-стилевую концепцию «с другого берега» – со стороны потребителя информации. Это позволяет по-новому взглянуть, например, на адекватность восприятия текста адресатом, успех коммуникативного акта.

2. Определение стилистических особенностей специфических «сетевых» текстов. Материальное средство фиксации текстов существенно влияет на их стиль (вспомним хотя бы латинские корни самого слова «стиль»), и Internet в этом аспекте - фактор не менее значительный, чем печатный станок пятьсот лет назад.

3. Изучение современной языковой парадигмы (см. пункт 5 определения понятия «стиль»). Речь является динамической системой, и значительные стилевые изменения могут происходить на относительно коротких временных промежутках. Тексты Internet являются удобным материалом для исследований по динамике стилей.

Литература

1. Ашманов И. и др. *Применение статистических методов для интеллектуальной компьютерной обработки текстов* // Труды международного семинара «Диалог 97» по компьютерной лингвистике и ее приложениям, 1997.

2. Головин Б.Н. *Язык и статистика.* – М.: Просвещение, 1970.

3. Кожина М.Н. *К основаниям функциональной стилистики.* – Пермь, 1968.

4. Кожина М.Н. *О речевой системности научного стиля сравнительно с некоторыми другими.* – Пермь, 1972.

5. *Лингвистический энцеклопедический словарь* / Гл. ред. В.Н.Ярцева. – М.: Сов. энциклопедия, 1990.

6. *Прикладная статистика: Классификация и снижение размерности: Справ. изд.* / С. А. Айвазян, В.М.Бухштабер, И.С.Енюков, Л.Д.Мешалкин; Под. ред. С. А.Айвазяна. – М.: Финансы и статистика, 1989.

7. Тулдава Ю. *Опыт классификации текстов с помощью кластер-анализа* // Труды по лингвостатистике. Кн. VII. – Тарту: Изд-во ТГУ, 1981.

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

9. Chen H., Schuffels Ch., Orwig R. *Internet Categorization and Search: A Self-Organizing Approach* // Journal of Visual Communication and Image Representation, Special Issue on Digital Libraries, Vol. 7, No. 1, 1996.

10. Karlgen J., Cutting D. *Recognizing Text Genres with Simple Metrics Using Discriminant Analysis* // Proceedings of 15th International Conference on Computational Linguistics (COLING), Kyoto, 1994.