Уральский государственный технический университет

На правах рукописи

Браславский Павел Исаакович

Методы повышения эффективности поиска научной информации (на материале Internet)

Специальность 05.13.16 – Применение вычислительной техники, математического моделирования и математических методов в научных исследованиях

Диссертация на соискание ученой степени кандидата технических наук

Научный руководитель –
доктор технических наук
профессор С.Л.Гольдштейн

Екатеринбург – 2000

Оглавление

Введение 4

1. Internet. Стили речи. Тезаурус 8

1.1. Internet и средства поиска информации 8
1.1.1. Краткая история 8
1.1.2. Основы построения машин поиска 11
1.1.3. Российский Internet 14
1.1.4. Современные тенденции развития средств поиска 20

1.2. Стили речи и классификация текстов 23
1.2.1. Различные подходы к определению понятия «стиль» 23
1.2.2. Статистические методы в стилистике и классификация текстов 25

1.3. Тезаурус 30
1.3.1. Краткая история 30
1.3.2. Тезаурусы в информационном поиске 31

1.4. Результаты и выводы 33

2. Классификация документов по стилям 36

2.1. Подходы к решению задачи стилистической классификации 36
2.1.1. Система стилей 36
2.1.2. Методика классификации 38

2.2. Опытный массив документов 40

2.3. Параметры классификации 42
2.3.1. Первичный набор параметров 42
2.3.2. Общая методика вычисления параметров 47
2.3.3. Формальные параметры 47
2.3.4. Формально-семантические параметры 49

2.4. Параметры опытного массива и их первичная статистическая обработка 50

2.5. Сокращение набора параметров 56

2.6. Классификация опытного массива документов 60
2.6.1. Условия применения процедур дискриминантного анализа 60
2.6.2. Первая классификация документов опытного массива 61
2.6.3. Главные компоненты – новые параметры классификации 65

2.7. Результаты и выводы 71

3. Расширение запроса с помощью тезауруса 72

3.1. Описание метода 72

3.2. Модель тезауруса и процедура формирования запросов 75

3.3. Программная реализация 79

3.4. Результаты и выводы 81

4. Проверка результатов. Практические рекомендации 83

4.1. Стилистическая классификация документов 83
4.1.1. Классификация тестового массива документов 83
4.1.2. Возможные реализации 91

4.2. Расширение запроса с помощью тезауруса 92
4.2.1. Запросы к ИПС «Яндекс» 92
4.2.2. Замечание о реализации 103

4.3. Результаты и выводы 103

Заключение 105

Библиографический список 108

Приложения 116

П1. Опытный массив документов 116
П1.1. Официально-деловой стиль 116
П1.2. Научный стиль 119
П1.3. Публицистический стиль 123
П1.4. Художественный стиль 125
П1.5. Разговорный стиль 128

П2. Списки для вычисления формально-семантических параметров 131
П2.1. Общенаучная лексика (37) 131
П2.2. Названия официальных документов (43) 132
П2.3. Научные корнеаффиксы приставочного типа (70) 132

П3. Диаграммы «ящики с усами» для параметров первичного набора 134

П4. Тестовый массив документов 144

П5. Тезаурус по компьютерной лингвистике 150
П5.1. Словник 150
П5.2. Набор семантических отношений 151
П5.3. Примеры установленных связей 152

П6. Акты внедрений и испытаний результатов работы 154

Введение

Эффективность научно-исследовательских работ напрямую зависит от качества их информационного обеспечения, а поиск информации является ключевым этапом любого научного исследования. На сегодняшний день глобальная сеть Internet – важнейший источник информации для всех областей знаний, однако поиск специализированной научно-технической информации при помощи Internet зачастую оказывается малоэффективным.

Сегодня с трудом верится в то, что на начальной стадии своего развития Internet был по преимуществу сетью научных и образовательных учреждений, а наиболее популярная служба Internet – WWW, – была разработана в европейском центре ядерных исследований CERN как способ объединения разнородных *научных* ресурсов и облегчения переходов между ними.

Один из авторов заметил, что “по мере развития Интернет обостряется парадокс: вероятность существования нужной информации возрастает, а возможность ее нахождения уменьшается” [33]. Это происходит потому, что наполнение сети очень разнородно, громадно по объему, быстро и нерегулярно обновляется, плохо поддается структуризации и управлению.

Internet унаследовал хаотичность структуры от своего предшественника – сети ARPANET, которая разрабатывалась по заказу военного ведомства США в конце 60-х – начале 70-х гг. Одним из главных требований заказчика была высокая живучесть сети, поэтому в основе ARPANET лежали полная независимость отдельных сегментов и отсутствие централизованного управления. Однако вряд ли разработчики могли тогда предположить, насколько стремительно будет развиваться этот проект. По данным Internet Software Consortium (www.isc.org), в январе 2000 года к Internet было подключено 72 398 092 машины (против 37 машин сети ARPANET в 1972 году).

Появление “всемирной паутины” (WorldWide Web), новых программ и протоколов; развитие средств связи и доступность персональных компьютеров сделали достаточно закрытую до этого сеть привлекательной для мил-

лионов новых пользователей. 1995 год был назван годом триумфа Internet. Впрочем, триумфальными для Internet можно назвать и все последующие годы.

В связи с бурным развитием Internet проблема поиска информации приобрела особую остроту и стала общезначимой. Поиск информации в Internet перестал быть темой исключительно специализированных изданий[1], а службы поиска превратились в коммерчески успешные компании. По данным Nielsen Media Research, к поисковым службам Internet сегодня обращается около 71% многомиллионной армии пользователей (цит. по [23]). По результатам опросов, наиболее требовательными к качеству информационного поиска оказываются научные работники (см., например, [53]).

По сравнению с локальными информационно-поисковыми системами (ИПС) задача поиска информации в Internet значительно усложняется[2].

В настоящее время в Internet представлены два основных вида служб поиска информации: 1) тематические каталоги ресурсов и 2) машины поиска (МП) по ключевым словам. Эти универсальные средства обладают целым рядом недостатков с точки зрения поиска научной информации.

Процесс отнесения документа к одному из разделов тематического каталога не поддается полностью автоматизации, поэтому каталоги охватывают ограниченное количество ресурсов и «не успевают» за ростом сети.

Машины поиска по ключевым словам охватывают больше ресурсов и чаще обновляются. Однако нередко они оказываются малоэффективными с точки зрения поиска научной информации из-за большого уровня шума (ссылок на нерелевантные документы), ограниченных возможностей языков запросов и формы представления результатов поиска.

---

[1] См., например: Быковский Е. Взгляд сверху вниз // Итоги. – 1999. – 24 августа. – С.56-57.

[2] Internet можно сравнить с очень большой библиотекой, где отсутствует какая-либо специализация и хранятся самые разные книги: от научных фолиантов и поэтических сборников до телефонных справочников и каталогов «Товары – почтой». Причем книги появляются в “библиотеке” совершенно самопроизвольно и в труднопредсказуемых местах. Интересы посетителей “библиотеки” не менее разнообразны, чем ее содержание. Со всем этим должен справляться “библиограф” – поисковая служба Internet.

Поэтому сегодня особую актуальность приобретают исследования, направленные на повышение эффективности поиска научной информации в Internet. Решение проблемы лежит в области разработки теоретических основ, методов и средств использования слабо структурированных информационных баз в научных исследованиях.

Целью нашей работы является разработка методов повышения эффективности поиска научной информации:

1) автоматической стилистической классификации документов и
2) расширения запроса с помощью тезауруса.

Оба метода ориентированы на усиление поиска русскоязычных научных документов с помощью универсальных МП.

Для достижения указанной цели в работе поставлены и решены следующие задачи:

- разработка процедуры автоматической классификации документов по стилям;
- разработка структурной схемы машины поиска с использованием стилистического анализа;
- разработка модели расширения запроса на основе тезауруса с сильно дифференцированным набором семантических отношений;
- разработка структурной схемы поиска информации с использованием процедуры расширения запроса на основе тезауруса;
- разработка макетных версий программ, реализующих предложенные методы;
- тестирование разработанных методов;
- выработка рекомендаций по практическому использованию методов.

Научная новизна работы состоит в следующем:

- разработана процедура автоматической стилистической классификации текстовых документов;
- введен показатель стилистической информативности документа;

- разработана модель расширения запроса на основе тезауруса с сильно дифференцированным набором семантических отношений.

Разработанные методы повышения эффективности поиска научной информации реализованы в виде макетных версий программ стилистического анализа и ассистента формирования запросов на основе тезауруса. В работе предложены эффективные с точки зрения реализации структурные схемы поиска научной информации с использованием разработанных методов.

Разработанные программы прошли испытания и внедрены в компании “Конвекс” (Екатеринбург), Свердловской областной универсальной научной библиотеке им. В.Г.Белинского, НИИ ЦветМет (Екатеринбург). Результаты работы используются в научных исследованиях и учебном процессе на кафедре риторики и стилистики русского языка Уральского государственного университета и на кафедре вычислительной техники Уральского государственного технического университета.

Основные результаты и положения работы докладывались и обсуждались на XXXV международной научной студенческой конференции «Студент и научно-технический прогресс» (Новосибирск, 1997), всероссийской конференции “Информационные технологии, системы управления и электроника” (Екатеринбург, 1997), семинаре “Методы прикладной математики и информационные технологии в многодисциплинарных исследованиях и проектах” (Омск, 1998), 30-ой региональной молодежной конференции “Проблемы теоретической и прикладной математики” (Екатеринбург, 1999), четвертом и пятом рабочих совещаниях по электронным публикациям EL-PUB-99 и EL-PUB-2000 (Новосибирск, 1999 и 2000), рабочем совещании “Новые Интернет-технологии” (Петрозаводск, 2000).

Основное содержание работы опубликовано в 5 печатных и 3 электронных работах.

1. Internet. Стили речи. Тезаурус

1.1. Internet и средства поиска информации

1.1.1. Краткая история

Когда говорят о развитии Internet за последние 5-6 лет, эпитетов обычно не жалеют. Действительно, темпы и размах поражают. Развитие Сети происходит по всем направлениям: увеличивается число хостов, объемы и разнообразие доступной через Internet информации, растет скорость передачи данных и численность пользователей, совершенствуются стандарты и программное обеспечение. Internet называют одним из главных факторов глобальных социальных и экономических изменений, происходящих сегодня во всем мире, и говорят о повышении этой роли в будущем.

Историю Internet обычно ведут от сети ARPANET, презентация которой состоялась в 1968 г. В 1969 г. в сети было четыре компьютера, в 1971 – 14, а в 1972 – уже 37. В 70-е гг. проходил процесс отладки технологии и роста Internet. В 1982 г. были стандартизированы основные протоколы: Transmission Control Protocol (TCP) и Internet Protocol (IP) [75].

В марте 1989 г. Тим Бернерс-Ли (Tim Berners-Lee) из CERN предложил руководству этого международного научного центра концепцию новой распределенной информационной системы, которую назвал WorldWide Web (WWW). В основе концепции лежала гипертекстовая технология, которая должна была объединять разнородные информационные ресурсы и облегчать переходы между ними. В 1990 году эти предложения были приняты и проект стартовал. Так началось развитие самой популярной технологии Internet. Первое сообщение об WWW было послано в телеконференции alt.hypertext, com.sys.next, comp.text.sgml и comp.mail.multi-media в августе 1991 года [75].

Само по себе появление WWW не стало революционным событием. Прошло еще целых полтора года до того момента, когда программа-браузер Mosaic, разработанная Марком Андресеном (Mark Andressen) из Националь-

ного Центра Суперкомпьютерных Приложений (NCSA), обеспечила бурный рост популярности "паутины". 1995 год был назван годом триумфа WorldWide Web. Сегодня Internet и World Wide Web многими воспринимаются как синонимы.

Павел Храмцов указывал [66], что «в октябре 1993 года количество зарегистрированных серверов WWW равнялось 500, а к июню 1994 года оно достигло 1500 и продолжает стремительно расти». По данным агентства Netcraft (www.netcraft.com), на 17.07.2000 во всем Internet насчитывалось 17 119 262 www-серверов. В мае 1999 г. их было в три раза меньше – 5 414 325.

По результатам исследования агентства Cyveillance, Inc. (www.cyveillance.com), опубликованным 10.07.2000, размер Internet оценен в 2,1 млрд. уникальных документов (URL). Ежедневно эта цифра увеличивается в среднем на 7,3 млн. При сохранении таких темпов роста в феврале 2001 г. в Internet будет насчитываться 4 млрд. документов.

По мере развития сети все более острой становилась проблема организации и поиска информации.

В 1989 г. широкой публике была представлена информационная система Gopher, разработанная в университете шт. Миннесота. Gopher представляет собой иерархический каталог распределенных ресурсов Internet. Система развивалась достаточно динамично, так что скоро встал вопрос об обозримости самого каталога. Эти проблемы была призвана решить система индексирования каталогов Gopher – Veronica (Very Easy Rodent-Oriented Net-wide Index of Computerized Archives), разработанная в 1992 г. в университете шт. Невада. До 1995 г. Gopher-серверы по темпам роста опережали www-серверы, однако постепенно Gopher утратил свое значение.

К 1991 г. относится анонс информационно-поисковой системы WAIS (Wide Area Information Servers). WAIS должна была перенести накопленный опыт разработки традиционных ИПС [58, 78, 79] на распределенные ресурсы Internet. Система WAIS предоставляет пользователю возможность поиска в полнотекстовых базах данных по запросу из ключевых слов. Современные

поисковые машины WWW во многом сходны с WAIS если не внешним видом, то механизмами функционирования [64].

На начальном этапе развития WWW стали возникать списки (коллекции) ссылок по определенной тематике, которые поддерживались энтузиастами [30]. Сам процесс последовательного перехода по гиперссылкам был приемлемой стратегией поиска при ограниченности пространства WWW, кроме того, воспринимался отчасти как развлечение [77]. Именно тогда появился английский фразеологизм *net surfing*[1].

В 1994 г. создание специальных средств поиска информации в пространстве WWW было уделом одиночек, *web-heads*, как называет их один из авторов [83]. Разработки, дожившие до Internet-бума 1995 года, получили новый импульс для развития, в том числе и в виде более или менее щедрых инвестиций[2]. В 1996 г. машины поиска стали широко обсуждаемой темой, а компании-лидеры успешно акционировались.

Именно в это время появились поисковые серверы «Yahoo!» (www.yahoo.com), AltaVista (www.altavista.com), HotBot (www.hotbot.com), WebCrawler (www.webcrawler.com), Lycos (www.lycos.com).

История российского сегмента Internet ведется с конца 80-х – начала 90-х годов, когда первые локальные сети ряда организаций науки и образования, построенные по IP-технологии, получили небольшие каналы связи с зарубежными сетями Internet. Первые сети СССР и России – это Relcom, Sovam Teleport, FreeNet, Radio-MSU, RunNet, Glasnet [56, 66].

В конце 1996 года в российском Internet появилась поисковая машина «Рэмблер» (www.rambler.ru) компании «Стек» (www.stack.net) [24].

---

[1] Роман Лейбов писал "об утраченных радостях серфинга": «Мы ходили по этому лесу без уверенности в том, что вернемся назад, с одной ненадежной (связь-то была та еще) ниточкой в руках. Мозаика, а за ней Нетоскоп позволили ходить по нескольким дорожкам сразу. ...Революцией было, конечно, появление ... Альтависты; хотя до нее разные пауки бегали, но толку от них было немного. Мы ходили по этому лесу без карты, наугад. Это была эпоха Великих Географических Открытий. Теперь наступило время межконтинентальных перелетов.» (Бессрочная ссылка от 20.04.98, http://www.russ.ru/ssylka/98-04-20.htm)

[2] Об истории успеха проекта «Yahoo!» см., например: Леви С. Претенденты, шаг вперед! // Итоги. – 1998. – 2 июня. – С.44-46.

Несколько позже появилась поисковая машина «Апорт» (www.aport.ru), разработанная компанией «Агама» (www.agama.ru). Об истории разработки ИПС «Артефакт» (компания «Интегрум», www.integrum.com) и «Апорт» см. [49].

23 сентября 1997 года считается датой официального открытия российской поисковой машины “Яндекс”, разработанной компанией COMPTEK (www.comptek.ru). Об истории “Яндекс” см. http://www.yandex.ru/skazki/story.html.

1.1.2. Основы построения машин поиска

Существующие сегодня поисковые машины Internet во многом реализуют идеи, которые были сформулированы еще в 70-х гг. для традиционных (локальных) ИПС [58[1], 78, 79].

Наиболее простой и распространенной является линейная модель поиска [58, 65]. 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 же случае – 0. *База* L – это матрица N x D, строки которой соответствуют N документам. Тогда процедура обработки запроса представима следующим образом: $L \cdot q = r$, где q – вектор запроса, а r – отклик системы на запрос.

Усложнение модели происходит путем приписывания терминам документа и запроса весов, отражающих их значимость. Использование весов позволяет определить меру близости «документ-запрос» s как косинус угла между вектором запроса и документа:

$$s(q,d) = \frac{\sum_i q_i d_i}{\sqrt{\sum_i q_i^2 \sum_i d_i^2}}, \qquad (1)$$

где $0 \le s(q, d) \le 1$,

$q=(q_1,..,q_n)$ – запрос, q_i – вес i-го термина запроса,

$d=(d_1,..,d_n)$ – документ, d_i – вес i-го термина документа.

Основные компоненты машины поиска (рис. 1.1): интерфейс, программа построения индекса (робот-индексатор) и сам индекс (индексная база) [1, 64, 66, 79].

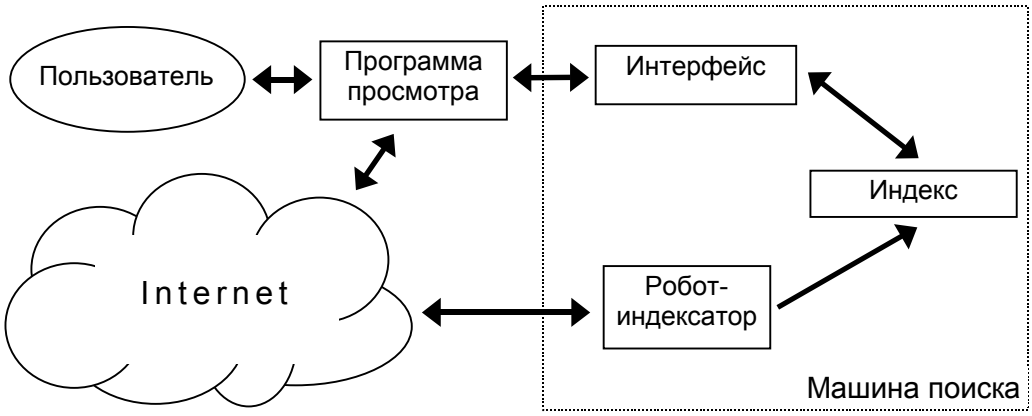

Рис. 1.1. Обобщенная архитектура информационно-поисковой системы Internet

Интерфейс выполняет несколько задач. Во-первых, он осуществляет перевод запросов, сформулированных пользователем в соответствии с языком запросов системы, в запросы к базе данных индекса. Во-вторых, интерфейс отображает отклик системы: группирует и ранжирует выданные документы, отображает места вхождения слов запроса в найденных документах. Наконец, после выдачи результатов интерфейс предлагает дополнительные стратегии поиска: “Искать в найденном”, “Найти похожие документы”.

Способ организации индекса – *know how* разработчиков машин поиска, поэтому подробные описания отсутствуют в доступных источниках. Однако можно сказать, что обязательными компонентами являются: таблица идентификаторов страниц, словарь (таблица ключевых слов) и инвертированный список [1, 64]. Таблица идентификаторов страниц устанавливает соответствие между числовым идентификатором и адресом ресурса. Словарь содержит числовые идентификаторы ключевых слов (т.е. слов, используемых для ин-

[1] Эту книгу Дж.Солтона Павел Храмцов назвал «библией разработчиков ИПС» [64].

дексирования). Инвертированный список – это массив документов, “точками входа” в который служат ключевые слова. Каждому такому слову соответствуют список идентификаторов документов, в которых оно встречается, с указанием точной позиции.

Программа построения индекса[1] обходит серверы и, углубляясь по ссылкам, индексирует документы. Именно этот компонент отличает ИПС Internet от локальных аналогов. От того, с какой частотой происходит обновление (актуализация) индексной базы, во многом зависит эффективность поиска. Так, по данным [23], в среднем 5% ссылок в выдаче указывают на несуществующие или устаревшие документы.

Кроме того, система может поддерживать архив запросов, включать тезаурус, средства проверки орфографии и синтаксического разбора естественно-языкового запроса.

Слова русского языка обладают ярко выраженной флективностью, т.е. они изменяются в зависимости от грамматической формы. Для среднего русского существительного таких форм двенадцать — одна основная и 11 косвенных (6 падежей единственного и 6 — множественного числа). У русского глагола косвенных форм — порядка 100 (с учетом причастия – до 180). Поэтому для эффективной работы с русскоязычными документами необходимо проводить поиск с учетом всех грамматических форм слов запроса, а для этого в свою очередь нужны средства морфологического анализа [6, 7, 49].

Языки запросов могут существенно различаться. Наиболее распространенными возможностями являются:

- формирование запроса с использованием логических связок AND, OR и AND NOT;
- ограничение на расстояние между словами запроса в документе;
- поиск по определенным полям документа;
- приписывание весов словам запроса;

[1] В литературе нет единого термина для обозначения этого компонента машины поиска. В различных англоязычных работах встречаются: *indexing robot, search agent, worm, infobot, crawler*.

- ограничения на дату/размер/расположение документа и другие.

Ранжирование найденных документов происходит на основе меры близости “запрос-документ” (см. формулу (1)). В ИПС WAIS ранжирование производится на основе отношения частоты встречаемости термина внутри документа к частоте встречаемости во всем массиве [65, 66]. Дополнительно учитывается близость слов запроса в документе.

При использовании коррекции запроса по релевантности («Найти похожие документы») запрос дополняется терминами указанного документа, а термины запроса, совпавшие с терминами документа, получают больший вес.

Существует множество критериев оценки эффективности функционирования ИПС. Основными принято считать следующие [7, 58]:

1. *Полнота поиска* – доля релевантных документов массива, присутствующая в выдаче.
2. *Точность поиска* – доля релевантных документов в выдаче.
3. *Усилия*, интеллектуальные или физические, затрачиваемые пользователем на формулирование запросов, помощь в процессе поиска и просмотр выдаваемой информации. Один вариантов этого критерия – *длина поиска*. 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.

1.1.3. Российский Internet

Российский сегмент Internet (Рунет) повторяет в миниатюре все мировые тенденции развития Сети: растет количество машин и пользователей, увеличивается объем информации, совершенствуются средства поиска.

Информационное наполнение. По данным Алексея Тутубалина (http://www.lexa.ru/ru-survey/), на конец 1999 года в российском сегменте Internet (домены ru и su) насчитывалось 362 463 хоста.

По информации Российского научно-исследовательского института развития общественных сетей (РосНИИРОС), управляющего доменом ru, на 04.07.2000 было зарегистрировано 46 266 серверов Internet (http://www.ripn.net/nic/dns/list/).

По результатам исследования Алексея Тутубалина (http://www.lexa.ru/ru-survey/), на конец 1999 г. в российском сегменте Internet насчитывалось 22 475 www-серверов[1] (рис. 1.2).

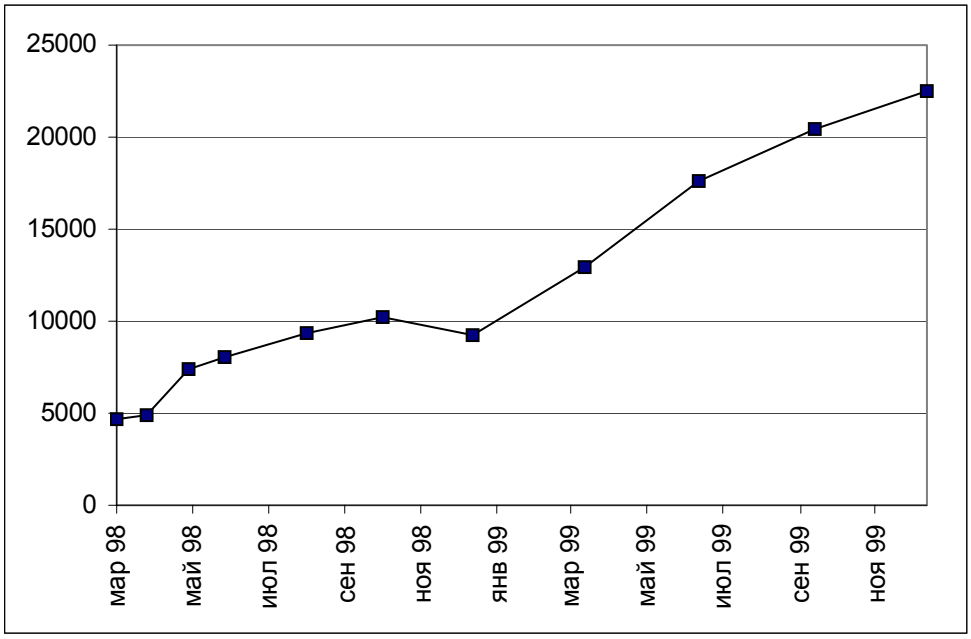

Рис. 1.2. Динамика распространенности www-сервиса в доменах ru и su. По данным Алексея Тутубалина (http://www.lexa.ru/ru-survey/)

Количественную оценку Рунета можно так же получить, обратившись к статистике российских поисковых машин.

Данные поисковой системы «Апорт» на 1 июля 2000 г. (http://www.aport.ru/WINP/info.htm):

Количество уникальных серверов: .....211 018.
Количество уникальных URL:.....10 424 348.
Объем проиндексированной информации: .....123,91Гб.

[1] В исследовании учитывались только «официальные» www-серверы, т.е. серверы с именем www.domainname или с именем, совпадающим с названием домена.

Данные ИПС «Яндекс»[1] на июль 2000 года (http://www.yandex.ru/chisla.html):

Количество уникальных серверов: .....142 074.
Количество уникальных URL:.....20 238 632.
Объем проиндексированной информации: .....214,06 Гб.

Динамика роста индекса «Яндекс» представлена на рис. 1.3.

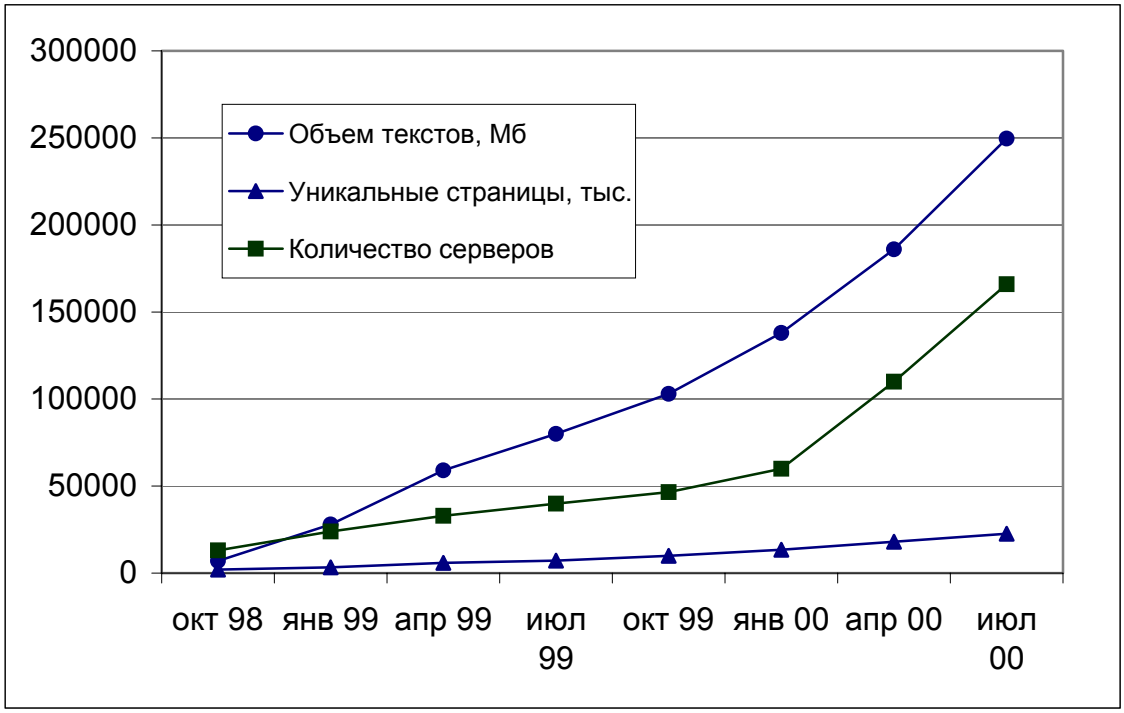

Рис. 1.3. Рост российского сегмента Internet.
По данным «Яндекс» (www.yandex.ru)

Из приведенных данных видно, что российский Internet быстро растет: увеличивается число серверов и объем информационных ресурсов. Причем наряду с развитием "в ширину" (увеличение числа серверов и страниц), можно отметить и развитие "в глубину" – увеличивается объем текстов на каждом сервере.

Пользователи. Максимальная взрослая (от 18 лет и старше) российская аудитория Internet в феврале 2000 по данным агентства monitoring.ru (www.monitoring.ru) [54] составила 6,6 миллиона человек, т.е. 6% взрослого населения страны (110,5 миллиона человек). В эту аудиторию входят те, кто

[1] «Яндекс» автоматически относит к Рунету серверы в доменах su, ru, am, az, by, ge, kg, kz, md, ua, uz. Кроме того, серверы вносятся в базу, если на них найден текст на русском языке или если владельцы ресурсов убедят администрацию поисковой машины в том, что их сервер интересен пользователям русскоязычного Internet.

имеет хотя бы единичный опыт посещения Internet. Ядро аудитории российского Internet образуют пользователи, которые проводят в Internet не менее 3 часов в неделю. Их 800 тысяч человек.

Аудитория Internet в России быстро растет. Если принять за 100% максимальную аудиторию по состоянию на февраль 2000 года, то всего за два с небольшим года до этого, в конце 1997 года, количество пользователей Internet составляло только 18% от аудитории февраля 2000 г.

В феврале 2000 г. в России количество тех, кто намерен начать пользоваться Internet, составило 12,5 миллиона человек. Из них 2,8 миллиона предполагают сделать это в ближайшие полгода, а 1,9 миллиона – в ближайшие три месяца. Эти данные подтверждаются анализом объективной динамики роста российской аудитории Internet.

Если наблюдающийся в последнее время закон роста сохранится, то в течение 2000 года российская максимальная аудитория Internet вырастет в 1,4 раза и к концу года составит около 9 миллионов человек, а недельная – около 4 миллионов. К концу 2002 года эти показатели могут достичь уровня 27 и 10 миллионов человек соответственно.

Тематические предпочтения максимальной российской аудитории Internet, по данным агентства moniring.ru [53], распределились следующим образом (пять наиболее популярных тем):

1. Новости ..... 32%[1]
2. Наука, образование 32%
3. Развлечения (анекдоты, игры) 25%
4. Музыка, литература 23%
5. Общение, чаты 22%

Интересно, что для ядра аудитории Internet – наиболее активной части пользователей, – наука и образование бесспорно выходят на первое место среди популярных тем (39%).

[1] Респонденты могли отмечать несколько категорий, поэтому общая сумма больше 100%.

Пользователи, принявшие участие в опросе «Яндекс» (www.yandex.ru/polling/9.html), на вопрос «Для Вас Internet –это…» ответили следующим образом (в порядке уменьшения популярности ответов):

1. Справочник..... 23,76%
2. Инструмент исследования 15,45%
3. Развлечение 14,15%
4. Источник новостей 12,32%
5. Средство общения 9,49%
6. Способ заработка 8,55%
7. Библиотека 7,02%
8. Другое 9,26%

По результатам того же опроса «Яндекс», больше всего пользователям Internet не хватает информации, в основном научной (библиотеки, документации, публикации), а также просто нужной (12,3%).

Некоторое представление об информационных потребностях пользователей и методах их работы с машинами поиска можно получить на основе статистики слов запросов.

В ноябре 1998 первая десятка списка самых популярных слов запросов на сервере «Апорт» выглядела так:

1. РЕФЕРАТ..... 2,51%
2. МОСКВА 1,30%
3. РАБОТА..... 1,15%
4. MP3..... 1,00%
5. СЕКС 0,92%
6. ПОРНО..... 0,87%
7. ПРОГРАММА 0,74%
8. SEX..... 0,69%
9. ИГРА 0,68%
10. РУССКИЙ..... 0,68%

Первые 100 самых популярных слов составляли треть (33,4%) всех слов запросов.

Однако можно отметить положительную тенденцию: если в 1997 г. средняя длина запроса к «Яндекс» составляла 1,2 слова, то к апрелю 2000 г. она увеличилась до 2,7 (http://www.yandex.ru/skazki/story.html).

Приведенные данные демонстрируют, что российская аудитория Internet быстро растет, а ее информационные запросы, интересы и ожидания весьма разнообразны.

Машины поиска. На сегодняшний день в Рунете работают три “большие” машины поиска[1]. Это уже упомянутые «Рэмблер» (www.rambler.ru), «Апорт» (www.aport.ru) и «Яндекс» (www.yandex.ru).

Данные об объемах информации, проиндексированной «Апорт» и «Яндекс», приведены выше. Точные данные об индексной базе «Рэмблер» отсутствуют. В табл. 1.1 сведены основные выразительные возможности языков запроса.

Таблица 1.1

Возможности языков запросов российских машин поиска

Операция	«Апорт»	«Рэмблер»	«Яндекс»
И	AND, &, +	AND, &	&&, +
ИЛИ	OR, I	OR, I	I
НЕТ	NOT, -	NOT, !	~~, ~
Группировка	( )	( )	( )
Словосочетания	" "	Не применяется	" "
Расстояние в словах	wN(...), [N,...]	$NEAR	.../(N M)*)...
Расстояние в предложениях	sN(...), {N,...}	Не применяется	...&&/(N M)...
Весовые коэффициенты	Не применяется	+, –	:w**)
Маска	*	*, ?	Не применяется
Поиск в поле документа	title=, keywords=, alt=, link=, comment=, text=	$All, $URL, $Title, $Header, $Essence, $Address	$title(), $anchor(), #keywords=, #link=, #hint=
Указание словоформы	!	По умолчанию	!

Примечания

*)N, M – число слов и предложений соответственно

**)w – натуральное число

Ранжирование. На результат ранжирования «Рэмблера» влияет встречаемость слова запроса в следующих полях документа (в порядке значимости):

- <TITLE>,

[1] Эти машины стремятся максимально полно представлять русскоязычное наполнение Internet. Существуют намного более скромные проекты, например, машины поиска Екатеринбурга: www.sverdlovsk.ru или grep.convex.ru.

- <H1 … H6>,
- <ADDRESS>,
- <B>, <STRONG>.

Чем чаще слово встречается в комбинации этих полей, тем более вероятно, что поисковая система «Рэмблер» поместит ссылку на документ ближе к началу списка результатов поиска. Кроме того, чем ближе располагается поисковое слово к началу документа, тем больше весовой коэффициент данного слова.

Метод ранжирования найденных документов «Яндекс» был описан Михаилом Масловым на форуме «Яндекс» (http://www.yandex.ru/forum.html) 13 октября 1999 г.

В большинстве случаев (более 90% запросов) для «Яндекс» решающее значение имеют не веса слов, а механизм приоритетов. Это значит, что документы, в которых все слова запроса находятся рядом и в том же порядке, что и в запросе, ранжируются выше всех остальных. Далее следуют документы, в которых есть все слова на указанном ограничении контекста (т.е. в пределах предложения – по умолчанию, – или всего документа). Можно считать, что остальные документы не полностью соответствуют запросу пользователя. Именно к ним применяется процедура нечеткого ранжирования – начинают играть заметную роль и количество различных слов запроса, и их веса. Количество вхождений каждого слова в документ при указанном ограничении контекста тоже имеет значение, но небольшое.

Судя по всему, более сложных методов ранжирования в российских машинах поиска не применяется.

1.1.4. Современные тенденции развития средств поиска

Можно выделить несколько направлений дальнейшего развития средств поиска информации в Internet (с использованием [23, 83]).

1. Оптимизация архитектуры поисковой машины, касающаяся алгоритмов индексации, организации массивов и баз данных, обработки запросов и т.п. Именно в этом направлении движется, например,

компания Fast Search&Transfer (www.fast.no) – разработчик поисковой службы www.alltheweb.com.

2. Развитие стандартов описания информационных ресурсов Internet, которые бы облегчали процесс индексирования. К этому направлению относится, например, развитие языка XML[1].
3. Персонализация (профилирование). Запоминая опции поиска, а также поддерживая личный архив запросов и откликов, признанных релевантными, машина поиска может настраиваться на конкретного пользователя. На основе этих данных система может делать заключения о тематических интересах и предпочтительных стратегиях поиска конкретного пользователя.
4. Специализация. Уже сегодня в связи с возрастающей мультимедиальностью Сети наблюдается специализация поисковых служб по типам информации. В первую очередь это касается цифровых форматов представления изображения, звука и видео. Соответствующие службы индексируют не только текстовые описания файлов (имя, расширение, текст поля ALT, текстовая информация файла MP3), но и само содержание файла, так что поиск можно проводить с учетом близости или по образцу. Примерами такого рода может служить технологии индексирования и поиска изображений, разработанные компаниями ImageLock (www.imagelock.com) и Excalibur [29, 31]. Кроме того, можно наблюдать специализацию поисковых служб по темам, региону и т.д.
5. Бóльшая вовлеченность пользователя в процесс поиска. Если взглянуть на распределение «нагрузок» в процессе поиска (см. рис. 1.1), то «центр тяжести» смещается в клиентскую, пользовательскую часть. Поиск информации становится все более интерактивным, настраиваемым и как следствие – персонализированным.

6. Развитие факультативных возможностей поиска. Разработчики универсальных машин поиска стараются удовлетворить все более разнообразные интересы растущего числа пользователей с помощью все более разнообразных возможностей языка запроса. Такие дополнения оформляются как расширение набора базовых возможностей.
7. Использование более сложных моделей поиска, которые учитывают человеческую психологию. На смену модели, основанной на символьном совпадении запроса и документа, приходят модели, учитывающие неявные предположения и когнитивные установки пользователя в процессе поиска (рис. 1.4).
8. Использование агентов[2] и ассистентов[3] поиска. Настроив своего личного агента, пользователь поручает ему поиск необходимой информации и ее первичную обработку: сортировку, ранжирование, закачку.

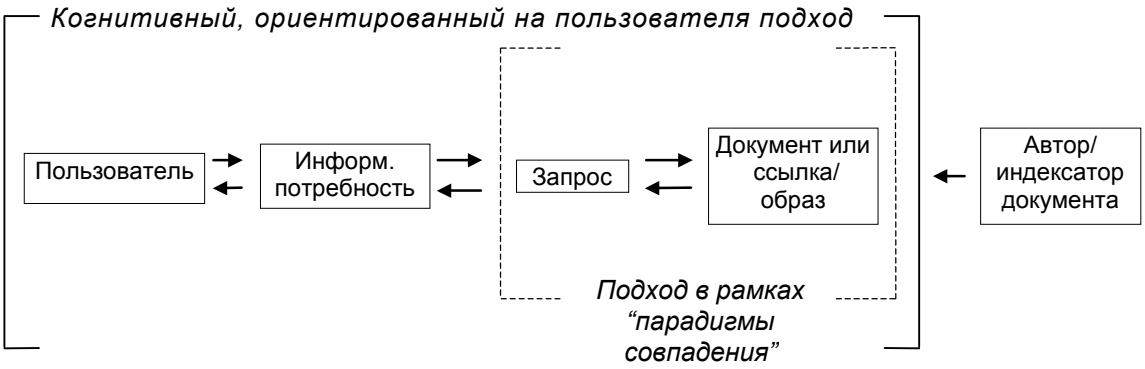

Рис. 1.4. Взаимодействие пользователя с информационно-поисковой системой [79]

[1] См., например, Храмцов П. XML через призму программирования // Открытые системы, – 1999. – №9-10. – С. 21-27. или Печерский А. Интернет учит новый язык //Компьютерра, 1999. – №27-28 (305-306). – С. 36-38, а также www.xml.com, www.xmlinfo.com, www.w3.org.

[2] Агенты – это программы, обладающие автономностью, самостоятельностью, персонализацией и даже – антропоморфностью. Агенты как новая метафора развития программного обеспечения стали главной темой журнала «Компьютерра» №7(285) за 1999 г.

[3] Примером ассистента поиска может служить программа «ДИСКо Искатель» российской компании «Арсеналъ» (http://www.ars.ru/products/disco/find.htm).

1.2. Стили речи и классификация текстов

1.2.1. Различные подходы к определению понятия «стиль»

На интуитивно понятном уровне можно сказать, что *тема* – это то, *о чем* текст, а *стиль* – то, *как* раскрыта тема. Т.е. *стиль* можно противопоставить *теме* (*план содержания* противопоставляется *плану выражения*).

Таким образом, наша задача – дополнить тематический поиск (запрос, состоящий из ключевых слов, – это приближенное описание содержания документа) стилистическим признаком. Ясно, что тексты на некоторые темы могут появляться только в определенных стилях, однако большинство тем допускает раскрытие во множестве стилей.

Общим местом для работ, посвященных стилистике, стало цитирование академика Виноградова [20]: “В области искусствознания, литературоведения и лингвистики трудно найти термин более многозначный и разноречивый – и соответствующее ему понятие – более зыбкое и субъективно-неопределенное, чем термин *стиль* и понятие *стиля*”.

Действительно, в лингвистическом энциклопедическом словаре [59] приводится несколько значений:

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:
   - нейтральный,
   - более “высокий”, книжный,
   - 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.

Функционально-стилевая концепция (соответствует пункту 2 определения) восходит к трудам В.В. Виноградова, Ш. Балли, ученых Пражского лингвистического кружка. Концепция утверждает экстралингвистическую, социальную основу расслоения литературного языка. Исходным положением является зависимость стиля речи от выполняемой им коммуникативно-общественной функции, от задач общения в соответствующей сфере. Обычно различают пять функциональных стилей речи: *официально-деловой, научный, публицистический, художественный*[1] *и разговорный*.

Функционально-стилевая концепция является наиболее разработанной и обоснованной в отечественном языкознании. В 60-80-е в СССР велось активное изучение функциональных стилей, результатом чего стало издание многочисленных монографий, учебников и научных сборников [8, 19, 39, 40, 42, 55].

Такой подход не является единственно возможным. Так, в статье [81], которую можно рассматривать как начальный импульс для нашей работы, используется система из пятнадцати жанровых категорий, что больше соответствует пункту 3 в определении понятия “стиль” (см. выше).

[1] Исходя из прикладного характера задачи, мы рассматриваем *художественный стиль* наряду с другими, не учитывая его «особый статус» в системе функциональных стилей.

В качестве примера также можно привести современные текстовые редакторы, которые осуществляют определенный стилистический анализ текста.

Так, в текстовом редакторе MS Word есть возможность выбора из трех наборов стилистических правил:

- «Строгая проверка» (все правила) – рекомендуется для проверки художественных текстов, публицистики и т.д.
- «Для деловой переписки» – рекомендуется для проверки деловой прозы: корреспонденции, текстов контрактов, научных статей и пр.
- «Для обычной переписки» – рекомендуется для проверки неформальных текстов – писем, внутрифирменных докладных записок, журнальных и газетных статей на общие темы и т.п.

Здесь *стиль* понимается в скорее в первом значении определения (см. выше), как “стиль языка” (традиция, идущая еще от античности). При таком подходе *стиль* становится нормативной (или рекомендательной) харакетристикой.

1.2.2. Статистические методы в стилистике и классификация текстов

По данным [41], первые работы, демонстрирующие применение статистических методов в стилистике, появились на рубеже XIX и XX веков. Примером может служить стилистическое исследование некоторых грамматических форм на материале художественных произведений, проведенное Марковым[1].

Первое использование компьютера для исследования литературного стиля относится к началу 50-х гг. [76]. Тогда Роберт Буза (Robert Busa) использовал вычислительную технику для создания конкорданса трудов Фомы Аквинского.

---

[1] Марков А.А. Об одном применении статистического метода // Известия Имп. Академии наук. Сер. 6. Т. X. – 1916. – №4.

В 60-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 году Б.Н. Головин, например, писал [21]: “...едва ли можно получить убедительную картину развития русского литературного языка на протяжении XIX-XX веков 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”.

В другой работе тот же автор так дал такое определение [22]: “Функциональные стили языка – это типы его функционирования, соответствующие различиям социальной практики коллектива и отличающиеся друг от друга *существенными различиями вероятностей языковых единиц и их категорий* (курсив наш. – *П.Б.*), достаточными для их совокупного качественного опознавания людьми на интуитивном уровне восприятия речи”.

В статье “К вопросу о применении статистики в языкознании” 1968 года [26] читаем: “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.”

В теоретической стилистике применение статистических методов традиционно имеет целью вскрыть закономерности функционирования языка в различных сферах коммуникации, обогатить знания о стилистических нормах этих речевых разновидностей, глубже рассмотреть проблему взаимоотношения мышления и речи [40]. К этому направлению примыкают литературоведческие по духу исследования индивидуальных стилей отдельных авторов (см.

пункт 4 определения понятия “стиль”) [62, 76, 63]. Более прикладной характер носят работы, относящиеся к нормативной стороне стилистики или к направлению “машинный перевод”.

Работы [3, 21, 22, 32, 40, 68, 76] содержат примеры продуктивного использования статистических методов в стилистике и создают основу для решения поставленной нами задачи. Новейшие примеры применения аппарата прикладной статистики в лингвистике демонстрируют работы [52, 48, 63].

Близкими нашей задаче являются примеры *тематической классификации* текстов в задачах информационного поиска.

В книге [58] описываются методы разбиения большого массива на кластеры близких документов для повышения точности и скорости поиска. Классификация происходит на основе поисковых образов документов (ПОД) – векторов пространства словаря (индексационных терминов) системы.

Метод самоорганизующихся карт Кохонена (Kohonen self-organizing maps, SOM) является дальнейшим развитием идей тематической классификации документов. Данная технология была разработана группой ученых, возглавляемой профессором Хельсинкского технологического университета Кохоненом (Kohonen). Первая публикация на эту тему появилась в январе 1996 года по адресу http://websom.hut.fi/websom/. Метод предназначен для организации доступа к массивам полнотекстовых документов в виде двумерной карты, раскраска которой отражает плотность распределения образов документов.

В статье [77] описывается автоматическая многоуровневая категоризация документов Internet с помощью SOM. После разбиения массива документов на категории («наука», «техника», «бизнес», «политика», «развлечения» и т.д.) пользователь может передвигаться по уровням иерархии и инициировать поиск по ключевым словам в соответствующей категории.

В работе [50] описывается опыт применения метода SOM к русскоязычным документам информационной системы ПАРК (см. http://is.park.ru/park/websom/demo.asp).

В статье [82] рассматривается проблема группировки близких / похожих / одинаковых документов при ответе на информационный запрос. Тематическое перекрытие может происходить при наполнении информационной базы из различных источников (например, разные газетные статьи, посвященные одному событию). Для такой группировки используются четыре иерархических агломеративных метода кластер-анализа:

- метод одиночной связи,
- метод полной связи,
- взвешенный метод «средней связи»,
- метод Уорда

и пять типов метрик.

Примерами классификации по *индивидуальным стилям* являются работы [62, 63].

В более ранней из них [62] с помощью методов кластер-анализа классифицируются тексты пяти авторов. В качестве параметров классификации используются:

- покрываемость текста словоформами,
- лексический спектр (доля словоформ с заданной частотой),
- динамика роста словаря.

На основании результатов классификации можно делать заключения о близости авторских стилей.

Программа «Лингвоанализатор» [63] создана для определения авторства литературных произведений. Для атрибуции программа использует следующие параметры:

а) число служебных слов (предлогов, союзов и частиц);

б) используемые в словах морфемы (приставки, корни, суффиксы, окончания) и их последовательности;

в) сложность используемых грамматических конструкций;

г) словарь, используемый автором.

Работы Юсси Карлгрена (Jussi Karlgren) [80, 81] наиболее близки нашей. Первая из них [81] послужила начальным импульсом для нашего исследования. В этой работе авторы используют корпус образцов английских текстов Брауна (Brown corpus of English text samples) – коллекцию текстов одинаковой длины, соответствующую 15 жанровым категориям:

- репортаж,
- редакционный материал,
- обзор,
- текст на религиозную тему,
- текст о хобби/увлечениях,
- популярные сведения,
- беллетристика,
- правительственный (официальный) документ,
- научный текст,
- художественная литература,
- детектив,
- научная фантастика,
- приключения,
- героический роман,
- юмор.

Для классификации используется модуль линейного дискриминантного анализа статистического пакета SPSS. Для процедуры классификации использовались следующие параметры:

- количество наречий,
- количество букв,
- количество длинных слов (> 6 букв),
- количество предлогов,
- количество местоимений 2-го лица,
- количество слов «Therefore»,
- средняя длина предложения в словах,
- средняя длина предложения в буквах,
- количество местоимений 1-го лица,
- количество слов «Me»,
- количество причастий настоящего времени,
- количество предложений,
- количество слов «I»,
- средняя длина слова в буквах,
- количество слов «It»,
- количество существительных,
- количество глаголов настоящего времени,
- количество слов «That»,
- количество слов «Which».

Авторы указывают на возможность применения процедур автоматической классификации по стилям в задачах информационного поиска, в частности – для категоризации сообщений телеконференций USENET.

1.3. Тезаурус

1.3.1. Краткая история

Слово *тезаурус* происходит от греческого *θησαυροξ* – сокровищница, запас, клад; в широком смысле используется для обозначения словаря, содержащего синонимы, антонимы, родственные слова, а также иерархическое дерево понятий с отношениями от общего к частному (род – вид), от целого к части.

Термин «тезаурус» достаточно древнего происхождения. Впервые его применил в значении, близком сегодняшнему, еще в XIII в. Брунетто Латини, автор систематизированной энциклопедии. Свой труд он назвал “Книга о сокровище”. Особенно большую известность получил тезаурус, составленный в 1852 году англичанином Роджетом “для облегчения выражения мыслей и помощи при написании сочинений”, который в 1956 году был использован Кембриджской группой по исследованию языка в работах по автоматическому переводу [71].

Трудоемкостью составления тезауруса объясняется тот факт, что на сегодняшний день существует всего два полноценных универсальных тезауруса: уже упомянутый тезаурус Роджета для английского и Робера для французского языка [6].

Вероятно, первой, кто употребил термин тезаурус в связи с вычислительными машинами, была Анетта Мастерман (см. [73]). В 1954 году она обратила внимание на то, что тезаурус – достаточно общее средство для описания семантических структур естественного языка и предложила использовать это средство в области машинного перевода.

В семидесятых годах можно было наблюдать целый тезаурусный бум, связанный с разработкой информационно-поисковых систем. Именно тогда

термин *тезаурус* прочно вошел в обиход как лингвистов, так и специалистов по информационным технологиям.

В работе [71] содержится такое определение термина *тезаурус*: “Тезаурус представляет собой словарь понятий и словарь классификационных связей между ними. Основным назначением тезауруса является перевод на информационно-поисковый язык смыслового содержания документов и запросов, относящихся к различным областям науки и техники, охватываемым классификационной схемой тезауруса. ...Тезаурус призван обеспечить единое унифицированное и формализованное представление информации в автоматизированных системах управления.”

Большая коллекция определений термина “тезаурус” собрана в книге [44]. Технология разработки описана в работах [17, 47, 58, 60, 69, 71, 72, 75].

Интересно, что исследователи, занятые прикладными аспектами тезаурусов, часто приходили к мысли об универсальном характере тезаурусного подхода. Так, Ю.И. Шемакин приводит следующее определение [71]: “Все понятия естественного языка, служащие для описания окружающего мира, представляют всеобщий тезаурус мира, отражающий весь универсум наших знаний”. В начале 70-х Ю.А. Шрейдер призывает отвлечься от узко прикладных аспектов применения тезаурусов в ИПС и взглянуть на тезаурус как на систему знаний [73]. С.Е. Никитина [46, 47] рассматривает тезаурус не только как инструмент информационного поиска, но и как средство анализа логико-понятийного аппарата науки[1]. В этом же русле – подхода к тезаурусу как системе знаний, – лежат наши работы [17, 18].

1.3.2. Тезаурусы в информационном поиске

Как уже было сказано выше, тезаурусы широко применялись в информационном поиске начиная с 70-х гг. В то время оперативная память, вычислительная мощность и дисковое пространство были ограниченными и очень

[1] При таком взгляде тезаурусы вызывают непосредственные ассоциации с идеями Людвига Витгенштейна о *семьях слов*, об исследовании *родственных связей слов* как одном из способов анализа языка. См. об этом

дорогими ресурсами. Для экономии этих ресурсов и повышения скорости работы поисковый образ документа (ПОД) можно формировать из ограниченного контролируемого набора терминов – словаря дескрипторов. С помощью тезауруса можно задавать классы эквивалентности терминов, при этом каждому классу соответствует один дескриптор. Такая организация позволяет производить замену значимых для поиска слов на дескрипторы как при индексировании, так и при обработке информационного запроса.

Сегодня индексирование в большинстве случаев является полнотекстовым – т.е. в индексной базе хранятся все слова документа[1]. Поэтому на этапе индексирования необходимость в тезаурусе отпадает, зато он может пригодиться при формировании запроса.

Тезаурус может использоваться для расширения запроса (*query expansion*): к словам запроса добавляются близкие им слова из тезауруса. Это позволяет находить релевантные текстовые фрагменты, вообще не содержащие слов запроса. Расширение запроса является одним из способов повышения полноты поиска (т.е. доли релевантных документов в отклике от общего количества релевантных документов).

Применение этого метода в поисковых системах Internet наталкивается на очевидную трудность: разнообразие наполнения Сети требует применения наиболее универсальных и полных тезаурусов, которых нет и которые вряд ли можно построить. С другой стороны, простое объединение по "ИЛИ" синонимов и прочих ассоциированных терминов тезауруса в силу тех же особенностей Internet может существенно снизить точность поиска (т.е. долю релевантных документов в общем числе найденных). Кроме того, такие возможности скорее всего останутся невостребованными большинством пользователей "поисковиков".

Витгенштейн Л. Философские исследования // Витгенштейн Л. Философские работы. Часть 1. Пер. с нем. – М.: Гносис, 1994. – Ч. I, §65.

[1] Даже если не все, то большинство. Храниться могут в том числе знаки препинания и слова из стоп-словаря (не нагруженные смыслом служебные слова, местоимения и т.п.). Эти слова не используются при поиске по индексу, но они нужны для восстановления связного текста по индексной базе.

На сегодняшний день тезаурусное расширение запроса не используются ни в одной из российских машин поиска. В 1997 г. компания «МедиаЛингва» (www.medialingua.ru) анонсировала выход тезауруса с 150 000 словарными входами, однако этот продукт до сих пор так и не вышел на уровень “промышленный эксплуатации”.

Пример использования тезауруса демонстрирует машина поиска AltaVista (www.altavista.com). Результаты поиска можно уточнить через контекст – набор слов, синхронно встречающихся со словами запроса[1] [83].

Если создание "тезауруса всего на свете" представляется нереальным, то построение тезауруса терминов узкой предметной области вполне осуществимо. Свойства терминов – системность, устойчивость и регулярность взаимосвязей, отсутствие субъективности и экспрессии, – делают возможным описание терминосистем с помощью тезаурусов. Технологиям построения тезаурусов посвящена специальная литература (см. выше).

Наиболее адекватного представления тезаурусом терминосистемы можно достичь с помощью набора сильно дифференцированных семантических связей. В таком тезаурусе в дополнение к универсальным связям "род-вид", "часть-целое", "синоним", "ассоциация" используются связи, специфические для конкретной предметной области. Общее количество типов однонаправленных связей может достигать нескольких десятков. Такой тезаурус не делит термины на непересекающиеся группы, а наоборот – связно представляет терминологию предметной области. Именно таким тезаурусам посвящены работы С.Е. Никитиной [47, 46], а также наша работа [17].

1.4. Результаты и выводы

На основании изучения печатных и электронных источников, конкретных продуктов и технологий, можно сделать следующие выводы.

---

[1] Когда речь идет о тезаурусах, необходимо учитывать разницу между *синтагматическими* и *парадигматическими* отношениями между словами. Первые говорят о семантической близости слов на основе частоты их совместной встречаемости в конкретных текстах. Процесс построения таких тезаурусов хорошо поддается автоматизации [58]. Именно такой тезаурус реализован на AltaVista. Парадигматические же связи говорят о взаимоотношениях *понятий*, выраженных терминами (например, *род-вид*) и не всегда выводимы из отдельных текстов.

1. В последнее время наблюдается бурное развитие Internet. В сети представлено все больше информации об объектах, предметах и областях научных исследований. Однако эта информация слабо структурирована, поэтому проблемы поиска научной информации приобретают все бóльшую актуальность.

2. Российский сегмент Internet (Рунет) повторяет в миниатюре все мировые тенденции развития Сети: растет количество машин и пользователей, увеличивается объем информации, совершенствуются средства поиска. Растет роль Internet в информационном обеспечении научно-исследовательских и опытно-конструкторских работ.

3. Российские машины поиска Internet несколько отличаются по объемам проиндексированной информации, возможностям языков запросов, методам ранжирования результатов поиска. Наличие нескольких поисковых служб, выполненных на высоком технологическим уровне, является, безусловно, большим достижением российского Internet. Однако все российские МП обладают на сегодняшний день целым рядом недостатков с точки зрения направленного поиска научной информации.

4. Научные работники по сравнению с другими пользователями имеют специфические информационные потребности и более требовательны к качеству информационного поиска. Специализация, полиферация, бóльшая ориентация на пользователя, расширение возможностей языков запросов и различных настроек поиска являются современными тенденциями развития средств поиска информации в Internet.

5. В русле этих тенденций лежат предлагаемые нами методы повышения эффективности поиска русскоязычных научных документов: автоматическое определение стиля документа и тезаурусное расширение запроса. На сегодняшний день эти методы не реализованы ни на одной из общедоступных универсальных машин поиска Рунета или в качестве автономных ассистентов поиска.

6. Термин стиль в языкознании многозначен, существует несколько подходов к определению этого понятия. Наиболее разработанной и обоснованной в отечественном языкознании является функционально-стилевая концепция.

7. В литературе описан ряд опытов по автоматической классификации текстов по стилям. В качестве прототипа для решения нашей задачи мы выбрали метод, описанный в работе [81].

8. Несмотря на богатую традицию использования в информационном поиске, сегодня тезаурусы находят лишь ограниченное применение в современных ИПС Internet. В качестве прототипа для разработки процедуры расширения запроса мы избрали тезаурус с сильно дифференцированными семантическими отношениями, описанный в работах С.Е.Никитиной [46, 47].

2. Классификация документов по стилям

2.1. Подходы к решению задачи стилистической классификации

2.1.1. Система стилей

Для решения нашей задачи – автоматической классификации документов по стилям, – в первую очередь необходимо выбрать *систему стилей*, т.е. на *какие* именно стили делятся документы.

В качестве основы для разработки процедуры автоматической стилистической классификации мы выбрали функционально-стилевую концепцию и соответствующую ей систему пяти функциональных стилей. Основными аргументами такого выбора стали разработанность концепции в отечественном языкознании и экстралингвистическая основа концепции.

Однако с точки зрения поиска информации не все функциональные стили равнозначны. Так, маловероятно, что кто-то станет направленно искать листинг чата. Тексты художественного стиля ищут обычно по автору и названию, реже – по более мелким (и характерным) жанрам, например: “научная фантастика”, “фэнтази”, “женский роман” и т.д. Скорее, объектом поиска может стать текст официально-делового, научного или публицистического стиля, а также стилей, занимающих промежуточное положение: научно-популярного, научно-учебного и т.п.

Стилистическая проверка, которую проводят современные текстовые редакторы «по вертикали» (см. п. 1.2.1), носит в основном нормативно-рекомендательный характер. Применение средств стилистического анализа такого рода оправдано на этапе создания и редактирования текста.

На основании анализа различных взглядов на *стиль* можно ввести новую стилистическую характеристику – *информативность* (рис. 2.1). «Стилистическая информативность» («горизонталь») больше соответствует задачам информационного поиска.

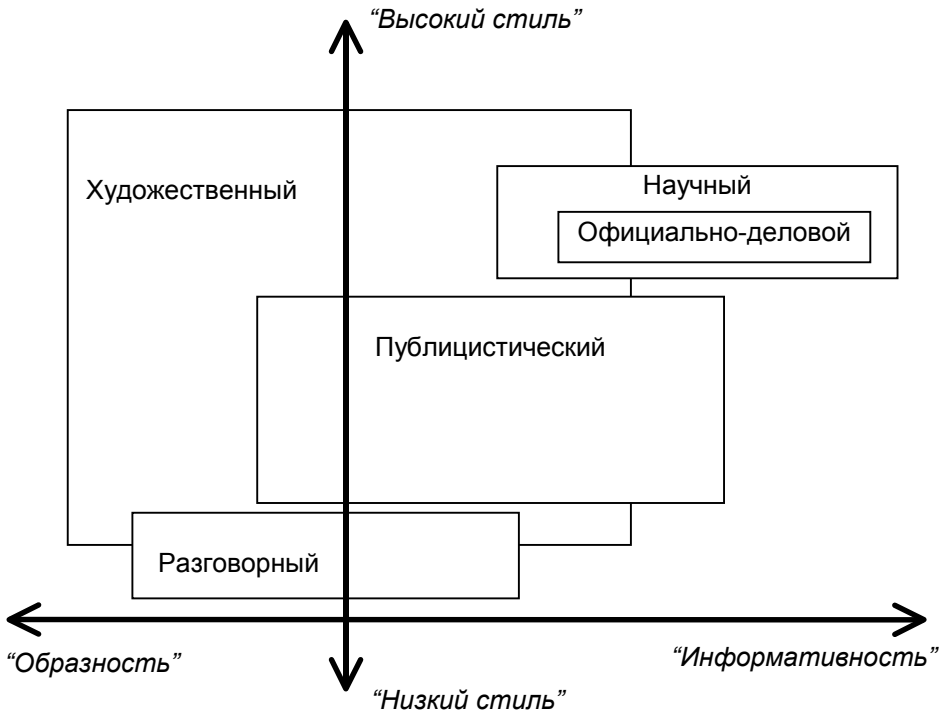

Рис. 2.1. Функциональные стили, стили языка и информативность

Для достижения прагматических целей коммуникации автор использует соответствующие ресурсы языка, т.е. цели автора проявляются в *стиле* текста. Именно это обстоятельство должна отражать «стилистическая информативность». На одном полюсе характеристики находится установка автора на объективную трансляцию информации без искажений (научный, официально-деловой документ), на другом – установка эмоционально воздействовать на читателя (художественная литература). Публицистический стиль занимает промежуточное положение на этой шкале.

Подчеркнем, что характеристика не отражает *семантику* текста[1] (например, насыщенность текста нужной, полезной информацией), а лишь его *стилистический* строй. Характеристика также не является *оценочной*[1].

[1] Прекрасный пример несовпадения стиля и содержания текста – «Квантовая теория танца» Я.И.Френкеля: «Сущность квантовой теории танца, представляющей собой своего рода компромисс между классической механикой условно периодических движений и классической эмоциодинамикой, заключается в следующем. Танцующие могут описывать определенные квантовые орбиты, не испуская и не поглощая при этом никаких эмоций. Последние испускаются и поглощаются прерывным образом при переходах с одной квантованной орбиты на другую. При этом в противоположность тому, что имеет место в случае электронных плясок в боровском атоме, эмоциональное излучение, как и поглощение, сопровождается переходом не на более низкий, а, наоборот, на более высокий уровень, т.е., другими словами, возбуждением. Таким образом, во время танца (особенно парного) возбуждение танцующих неизменно возрастает, пока не наступит релаксация, вызываемая истощением.» (Физики шутят. – М.: Мир, 1993.).

Переход к новой стилистической характеристике мы предлагаем совершить, отталкиваясь от функционально-стилевой концепции. Во-первых, как уже было сказано выше, концепция является хорошо разработанной и обоснованной. Во-вторых, такой подход не только плодотворен для решения нашей прикладной задачи, но может также обогатить наши знания о стилях речи.

Подход к функциональным стилям “с другой стороны” – со стороны потребителя информации, – позволяет по-новому взглянуть, например, на адекватность восприятия текста адресатом, успех коммуникативного акта. Кроме того, речь является динамической системой, и значительные стилевые изменения могут происходить на относительно коротких временных промежутках. Поэтому тексты Internet являются удобным материалом для исследования динамики стилей, изучения современной языковой парадигмы (см. пункт 5 определения понятия “стиль” в п. 1.2.1). Сегодня это актуально еще и потому, что меняется материальное средство фиксации текстов: бумагу и ручку (печатную машинку) заменяет компьютер; а Internet в этом аспекте – фактор не менее значительный, чем печатный станок пятьсот лет назад. Такая ситуация безусловно оказывает влияние и на стиль текстов (вспомним происхождение самого слова “стиль”).

2.1.2. Методика классификации

Из прикладного характера поставленной проблемы следует, что метод классификации должен быть достаточно простым в вычислительном плане. Из задачи *автоматической* классификации текстов по стилям вытекает задача *автоматического* вычисления классификационных параметров текста. Параметры также должны быть легковычислимыми, а их набор – в каком-то

По мнению Налимова [45], природа юмора как раз и состоит в нетипичном использовании языковых средств.

[1] Т.е. на основании характеристики нельзя сделать, например, заключение, что роман Толстого «Война и мир» хуже или лучше, более или менее значителен, чем «Курс теоретической физики» Ландау и Лифшица. Можно только сказать, что эти произведения *отличаются по стилю*.

смысле оптимальным. Кроме того, параметры должны быть содержательны с точки зрения стилистики.

В качестве метода построения классификации мы выбрали дискриминантный анализ (ДА) – один из разделов прикладной статистики. ДА является устойчивым и надежным методом изучения различий между несколькими классами и построения классификаций на основе обучающих выборок [9, 36, 51].

Составление набора параметров становится наиболее сложной и ответственной задачей. Обширная литература по стилистике не может дать нам готовое решение, т.к. задача классификации текстов по стилям в таком виде в стилистике просто не ставится. Кроме того, в этой литературе стили характеризуются в первую очередь качественно, например: *эмоциональностью, установкой на отсутствие экспрессии или объективную передачу информации, императивностью, отвлеченностью/конкретностью* и т.д. Количественные параметры имеют иллюстративный характер, а рассмотрение редко бывает комплексным (обычно речь идет об одном стиле или сравниваются два-три стиля).

В такой ситуации процесс получения набора параметров классификации целесообразно разбить на два этапа.

На первом мы создаем первичный набор параметров “с запасом”. Два фактора являются решающими для включения параметра в первичный набор: легкая вычислимость и потенциальная значимость для задач стилистической классификации. Второй этап – оптимизация набора параметров.

Из требования “простоты” вытекает, что параметры берутся в основном с “нижних” уровней языковой системы (графики, лексики, морфологии) и имеют формальный характер. Хотя на уровне синтаксиса сосредоточены многие стилистические ресурсы, а значит – и характерные признаки, проводить полноценный синтаксический разбор предложений для нашей задачи представляется чрезмерно затратным. Некоторые синтаксические параметры

можно получить по косвенным легковычислимым признакам (например, по количеству знаков экспрессивной пунктуации или отдельных союзов).

Основная единица для вычисления параметров – отдельное слово. Во-первых, именно слово является единицей индексирования машин поиска[1]. Во-вторых, как сказано выше, мы отказались от определения синтаксических параметров.

Чтобы сохранить универсальность подхода, мы выделяем только *лингвистические* параметры (относятся непосредственно к тексту), и не касаемся внешних (относятся к оформлению текста в различных форматах – HTML, TXT, TeX[2], PostScript, PDF).

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 нам, с одной стороны, *системность стилей речи*: стиль “проявляется” синхронно на всех уровнях языковой системы[3] [39, 40]. С другой стороны, нашу уверенность подкрепляет успех аналогичных опытов [63, 81].

2.2. Опытный массив документов

Взятая за основу функционально-стилевая концепция определила наш подход к формированию опытного массива документов для анализа.

---

[1] Поэтому, например, если в тексте встретился фразеологизм *белка в колесе*, то слова *белка* и *колесо* будут проиндексированы отдельно, а предлог *в* будет проигнорирован как служебное слово.

[2] В формате TeX обычно оформляются математические тексты. Однако желание авторов перевести их в более универсальный формат и сделать, таким образом, более доступными, приводит к созданию процедур автоматической генерации HTML из TeX. См., например: *Моргунова Н.Н., Финогенов А.А., Устюжанин А.М.* Опыт использования конверторов LaTeX2HTML и TTH при создании HTML-страниц: Доклад IV рабочего совещания по электронным публикациям – EL-PUB-99, Новосибирск, Академгородок, ИВТ СО РАН, 21-23 апреля 1999 г. – http://www.ict.nsc.ru/ws/elpub99/, а также сайт *WebEQ* компании Design Science (www.webeq.com).

[3] По этому поводу один из авторов пишет [19]: “В функционирующем языке действует общий закон балансированного (или компенсирующего) распределения нагрузки (речевых усилий) между различными секторами и элементами его структурно-функциональной системы. ...В каждом речевом стиле вырабатывается и утверждается самой практикой некая средняя константа речевых усилий, необходимых для производства и восприятия речи. Овладение речевым стилем связано с выработкой у человека определенной установки на соответствующую константу речевых усилий, имеющую как количественные, так и качественные характеристики.”

Официально-деловой стиль представлен в опытном массиве текстами 50 федеральных законов последнего времени. Эти документы были отобраны из юридической базы данных “Консультант Плюс” (www.consultant.ru). Дополнительным критерием отбора была длина текста. Например, были отсеяны законы о ратификации договоров, которые обычно содержат 2-3 строчки.

Коллекция документов **научного стиля** более разнообразна. Сюда вошли 54 статьи по физике, математике, химии, биологии и инженерным наукам. Практически все электронные версии научных статей, размещенные в Internet, имеют печатные аналоги.

Публицистический стиль – напротив, – представлен только Internet-журналистикой. В опытный массив документов вошли статьи на общественно-политические темы, опубликованные в период с декабря 1999 по февраль 2000 на трех новостных веб-сайтах: Gazeta.ru (27), Vesti.ru (28) и Polit.ru (6) – всего 61 статья. Первые два сайта представляют собой Интернет-аналоги ежедневных бумажных газет (что отражено в их названиях и подчеркнуто “газетной версткой”). Polit.ru – скорее новостная лента политической направленности, где публикуются обзорные статьи.

Художественный стиль в нашем исследовании представлен рассказами участников конкурса сетевой литературы “Тенета-98” (http://www.teneta.rinet.ru/1998/rasskaz/). В опытный массив мы включили 79[1] рассказов, поданных на конкурс. Сетевая литература – одно из интересных явлений Internet. Нам представляется закономерным использовать для анализа произведения, которые увидели свет в Internet и в принципе не нуждаются в бумажном носителе, а не литературную классику[1].

Основной объем текстов **разговорного стиля** принадлежит екатеринбургскому чату “На Плотинке” (www1.ekaterinburg.com/leisure/chat/), всего 42 фрагмента. Благодаря удобной функции “Показать последние 100 сообще-

[1] Из 80 рассказов, поданных на конкурс, мы исключили рассказ Данилы Давыдова “Подземные жители”. Повествование в рассказе ведется от лица главного героя, который сильно заикается, что отражено на письме. В такой ситуации автоматическое вычисление параметров не срабатывает.

ний”, реализованной на этом сайте, каждый фрагмент содержит ровно 100 сообщений участников чата. Кроме того, два фрагмента взято с чата “Сайберия” (http://www.son.ru/chat/) и четыре – с чата “В пещере у монстра” (cave.extrim.ru). В данном случае объем каждого фрагмента – примерно 2-3 экрана. Дополнили коллекцию образцов разговорного стиля 13 листингов диалогов (14 разных участников), которые велись с помощью программы ICQ (“аська”). Таким образом, всего был использован 61 фрагмент.

В дальнейшем все элементы опытного массива – листинги чатов, фрагменты диалогов, рассказы, научные и публицистические статьи, законы, – мы называем *документами*. Всего опытный массив содержит 305 документов.

Прил. 1 содержит описание опытного массива: списки документов официально-делового, научного, публицистического и художественного стилей с указанием источников, а также пример документа разговорного стиля.

2.3. Параметры классификации

2.3.1. Первичный набор параметров

На основании изучения литературы по стилистике русского языка [8, 19, 21, 39-42, 55], опытного массива документов и в соответствии с критериями, изложенными в п. 2.1.2, мы сформировали первичный набор параметров классификации (табл. 2.1). Все параметры мы разделили на *формальные* и *формально-семантические*.

Параметры первой группы – это либо подсчет определенных знаков (разделителей между словами или предложениями; формул), либо определенных грамматических характеристик слов. В случае формальных параметров конкретный смысл анализируемого слова или предложения нас “не интересует”. Вычисление формально-семантических параметров – это сравнение

---

[1] Хотя сетевая литература все еще испытывает что-то вроде комплекса перед “большой” бумажной литературой. По условиям конкурса “Тенета” произведения победителей печатаются отдельной книжкой.

каждого анализируемого слова с элементами заданных списков (использование “словарей”)[1].

Таблица 2.1

Первичный набор параметров классификации

Уровень	Параметры	
	формальные	формально-семантические
графика	формулы	smiles (“улыбки”)
словообразование	нет	приставки, характерные для научного стиля
лексика	средняя длина слова	• общенаучная лексика • названия официальных документов • слова организации логики повествования
морфология	• распределение по частям речи • существительные среднего рода • возвратные глаголы • аббревиатуры	• личные местоимения 1-го и 2-го лица: *я, ты, мы, вы* • частицы *бы* • частицы *ну, вот, ведь*
синтаксис	• цепочки имен существительных в родительном падеже • средняя длина предложения в словах • доля предложений с экспрессивной пунктуацией	доля предложений с подчинительными союзами

Дадим краткую характеристику параметров первичного набора.

Формулы. Формулы характерны для естественнонаучных и математических текстов.

Smiles. «Улыбки» («мордочки», «смайлики») – непременный атрибут неформального общения в сети Internet.

Словообразование. Для словообразовательной системы научной речи характерна количественная ограниченность общенаучных словообразовательных аффиксов и их функциональных значений. Особенно продуктивны в

[1] Лучше всего различие между двумя группами параметров можно продемонстрировать на примере известного стихотворения «Бармаглот» Л.Кэрролла (цит. по: Падни Дж. Льюис Кэррол и его миф. – М.: Радуга, 1982. – С.57.):

Варкалось. Хливкие шорьки
Пырялись на наве,
И хрюкотали зелюки,
Как мюмзики в мове.

Хотя смысл слов в этом стихотворении туманен, мы вполне можем определить их грамматические характеристики и синтаксические роли.

словообразовании научной лексики корнеаффиксы приставочного типа, в основном греческого и латинского происхождения [19].

Лексика. Одна из особенностей разговорной речи – стремление к *экономии речевых усилий*, что проявляется как в редуцировании (упрощении) отдельных слов[1], так и в использовании более «простых» из возможного синонимического ряда[2]. Средняя длина слова отражает эту особенность.

В лексике научной речи можно выделить ядро, общее для всех научных направлений. Это т.н. общенаучная лексика – общие категориальные понятия, одинаково актуальные для всех или для значительного ряда наук, общий понятийный фонд науки в целом [8, 19, 67].

В официально-деловой речи можно также выделить класс часто употребляемой специфической лексики – например, названия официальных документов [8].

Одна из задач научного текста – описание хода мысли, изложение многоэтапного рассуждения. Поэтому в научных текстах часто употребляются слова организации логики повествования (*во-первых, во-вторых, итак, следовательно, затем* и т.п.) [8, 19].

Морфология. Морфологические параметры образуют самую большую группу формальных параметров.

Количественные характеристики морфологии стилей речи разной степени детализации можно найти в работах [3, 19, 22, 27, 28, 32, 37, 40, 55, 68][3].

На основе анализа специальной литературы можно сделать предварительные заключения о морфологических особенностях различных стилей речи.

---

[1] Ср. компьютерный жаргон: «клава» – клавиатура, «комп» – компьютер, «прога» – программа, «генерить» – генерировать.

[2] В разговорной речи скорее будет употреблено слово *бегемот*, а не *гиппопотам*; *фокусник*, а не *престидижитатор*; *машина*, а не *автомобиль* и т.д.

[3] Недостаток большинства этих источников – использование для анализа выборок небольшого объема (часто трех-пяти текстов). Кроме того, не всегда понятно, какие именно тексты послужили материалом для исследования, и какая использовалась методика. Практически нигде не удается найти интегральную картину распределения классов слов по стилям: обычно одновременно рассматривается не более трех стилей. Данные “Частотного словаря” [68] лишены этих недочетов (общий объем обработанного материала – 1 056 382 слова), однако деление на стили (жанры) представляется не очень логичным: *художественная проза, драматургия, газетно-журнальные и научно-публицистические тексты*.

Наиболее важная морфологическая характеристика – соотношение частоты употребления глаголов и имен существительных (и шире – глагольных и именных частей речи). По данным Кожиной [40], в научном стиле на каждую тысячу словоупотреблений приходится 337 имен существительных и 90 глаголов, а в официально-деловом 421 и 60 соответственно. В художественных текстах глаголы (как и наречия) более активны: в художественной речи глаголов употребляется в среднем в 2,5 раза больше, чем в деловой, и в 1,7 раза больше, чем в научной [40]. В разговорном стиле наблюдается бóльшая активность незнаменательных слов (местоимения, частицы), соответственно меньше существительных и прилагательных. В официально-деловой речи частота наречий минимальна.

В текстах официально-делового стиля часто употребляется страдательный залог, который в русском языке выражается возвратной формой глагола [8]. Для научной речи характерно значительное количество возвратных глаголов, могущих сочетать потенциально и страдательное значение, и активно-безобъективное [19].

В научных текстах больше существительных среднего рода (как более отвлеченных по значению) [8, 41].

В книжных стилях речи чаще встречаются причастные и деепричастные обороты (доля причастий в разговорной речи – 0,1%, в книжной – 6% [41]).

В публицистическом стиле более активно используются аббревиатуры и словосложения терминологического характера.

По данным [57], около 25% словоупотреблений частиц в разговорной речи занимают две частицы: *вот* и *ну*. В [41] указывается на повышенную частоту частиц *ну, вот, ведь* в разговорной речи. Поэтому мы выделили эти частицы в отдельный список.

Местоимения 1-го и 2-го лица почти полностью отсутствуют в официально-деловой речи и, наоборот, преобладают в разговорной. В научных текстах нет личных местоимений *ты*, *вы* и соответствующих форм глаголов;

редко употребляется форма *я*, чаще – авторское *мы*. Поэтому отдельный параметр (наряду с общим числом местоимений в тексте) – доля местоимений личных местоимений 1-го и 2-го лица: *я, ты+вы, мы*.

Научная и официально-деловая речь имеют установку на однозначность и точность. Поэтому здесь, в отличие от художественного и разговорного стилей, практически не используется сослагательное наклонение. Об употреблении сослагательного наклонения мы можем судить по частице *бы*.

Синтаксис. В книжных стилях речи (научный, официально-деловой) часто встречаются цепочки существительных в форме родительного падежа, особенно богат ими канцелярит.

Книжные стили речи отличаются также более активным использованием сложных предложений. Наряду с более частым употреблением причастных и деепричастных оборотов, это ведет к увеличению средней длины предложения [8]. Длина предложения косвенно отражает сложность его структуры.

Экспрессивная пунктуация (?, !, …) практически отсутствует в научном и официально-деловом стилях, чаще употребляется в разговорном и художественном.

Сложноподчиненные предложения отличаются более книжным характером, чем сложносочиненные, и более высокой частотой употребления в сфере книжной речи. Сложноподчиненные предложения в разных стилях книжно-письменной речи составляют от 27 до 31%, а в разговорной – лишь 10% [41]. Способность сложноподчиненного предложения выражать логические (причинно-следственные и условные) отношения объясняет особенно высокий процент употребления в научной речи с ее специально подчеркнутой строгой логичностью. Доля сложноподчиненных предложений в научных текстах составляет 60%, соответственно насыщенность научной речи подчинительными союзами выше, чем других сфер общения [8].

2.3.2. Общая методика вычисления параметров

Описываемая ниже методика вычисления реализована в виде консольного приложения, написанного на языке Visual C ++.

Анализу подвергаются текстовые документы (*plain text*) и документы HTML в Windows-кодировке. Документы других форматов (например, Word и Adobe Acrobat – PDF) необходимо предварительно конвертировать в текстовые файлы.

В каждом документе для вычисления параметров используются первые 1000 слов плюс слова до конца текущего предложения, или весь текст, если его длина меньше тысячи слов. В дальнейшем все выводы о числовых параметрах *документа* относятся именно к этому начальному, обработанному фрагменту текста.

Словом считается последовательность букв (которая может содержать внутри себя дефис) между двумя разделителями. Слова, содержащие цифры, пропускаются.

Все параметры (кроме средней длины предложения и средней длины слова) являются относительными: абсолютное значение делится либо на общее количество слов, либо на общее количество слов определенной части речи, либо на количество предложений.

2.3.3. Формальные параметры

Формулы. В документах HTML формулы описываются по-разному. Сложные формулы обычно вставляются в виде картинок, а для оформления более простых используются стандартные средства языка гипертекстовой разметки. Поэтому мы заменили подсчет собственно формул на подсчет знаков “=”, что дало хорошие результаты для документов научного стиля (на которые и ориентирован этот параметр).

Морфология. Все морфологические характеристики рассчитывали с помощью модуля морфологического анализа LINGUIST компании “Агама” (www.agama.com). По информации разработчиков, основной словарь модуля

морфологического анализа и синтеза позволяет распознавать более четырех миллионов словоформ без использования алгоритмов сложного дефисного словообразования. Модуль выполнен в виде динамической библиотеки Windows. По словоформе модуль строит нормальную (словарную) форму и указывает ее грамматические характеристики. Словоформы, для которых модуль не возвращает ни одной нормальной формы, при вычислении морфологических параметров не учитываются.

По аналогии с предыдущими исследованиями морфологии функциональных стилей (см. п. 2.3.1) и в соответствии с возможностями модуля LINGUIST в качестве самостоятельных морфологических классов были выделены:

1. Существительные[1]
2. Прилагательные[2]
3. Местоимения
4. Числительные
5. Наречия
6. Глаголы
7. Причастия
8. Деепричастия
9. Предлоги
10. Союзы
11. Частицы
12. Междометия
13. Прочие[3]

Очевидно, что вычисленные таким образом формальные параметры не являются абсолютно точными.

Так как процесс должен быть полностью автоматическим, при вычислении морфологических параметров не учитывается грамматическая омонимия. Поэтому, например, слова *стекло, падали* всегда относятся к существительным, как и *рабочий, учащийся*; а *печь, течь* – к глаголам.

Ошибки определения границ слова могут возникать из-за переносов или вставки невидимых символов и тэгов HTML в слово.

Знак «=» не всегда указывает на формулу, а может, например, использоваться в оформлении документа в формате *plain text*.

Ошибки определения конца предложения могут возникнуть из-за сокращенных слов с точкой. В текстах разговорного стиля часто наблюдается

---

[1] К существительным мы также относим следующие категории модуля морфологического анализа: *имена собственные, отчества, фамилии, географические названия, аббревиатуры*.

[2] Несмотря на то, что краткие прилагательные имеют стилистическую окраску, мы не выделяем их в самостоятельный класс из-за особенностей функционирования модуля морфологического анализа.

[3] В этот разряд попадают те слова, которые модуль морфологического анализа отнес к *предикативам* или *вводным словам*.

полное игнорирование пунктуации, к тому же распространена парцелляция[1]. В научных текстах знаки “...”, “!” могут использоваться не только для обозначения конца предложения (например, $n!$ или $x_1, \ldots, x_n$).

Однако мы считаем, что в целом неточность вычисления параметров не фатальна и компенсируется полной автоматизацией процесса.

2.3.4. Формально-семантические параметры

Как мы уже указывали выше, процедура вычисления формально-семантических параметров состоит в сравнении каждого слова (или начального участка слова) с элементами конкретных списков.

Smiles. Мы учитываем наиболее распространенный тип смайлов: [;|:]{-}[(|)][2], т.е. последовательность “глаза” (одного из двух типов), “нос” (может отсутствовать) и “рот” (“грустный” или “веселый”).

Научные приставки. С помощью словарей[3] мы выделили 70 корнеаффиксов приставочного типа латинского и греческого происхождения, характерных для научного стиля. Так как подсчет этого параметра проходит формально (сравнением начального участка слова с элементом списка, без полного анализа состава слова), в список не были включены префиксы, которые потенциально значимы, однако совпадают с началом слов другого строения (ср.: *топология – топорище; метафизика – металлический, метательный*).

Подчинительные союзы. С помощью “Русской грамматики”[4] мы выделили 20 подчинительных союзов и союзных слов (*если, поскольку, поэтому, хотя, чтобы* и др.), которые должны указывать на сложноподчиненные предложения. Существенный неполнота этого списка (как и списка *слов организации логики повествования*) – использование только простых (однословных) элементов (в соответствии с принципом “слово – базовая единица анализа”).

---

[1] Парцелляция – способ речевого представления предложения несколькими коммуникативно самостоятельными единицами – фразами, является средством речевой экспрессии.
[2] В нотации Бэкуса-Наура.
[3] Словарь иностранных слов. – 10-е изд., стереотип. – М.: Русс. яз., 1983. – 608 с.; Словарь русского языка: В 4-х т. / АН СССР, Ин-т русс. яз.; Под ред. А.П.Евгньевой. – 2-е изд., испр. и доп. – М.: Русс. яз., 1981.
[4] Русская грамматика. В 2-х т. Т II. Синтаксис / АН СССР, Ин-т русс. яз.– М.: Наука, 1981.

Названия официальных документов. Список названий официальных документов был взят нами из юридической базы данных "Консультант Плюс": *акт, декларация, директива, договор, закон, заявление, извещение* и т.д., всего 43 названия.

Общенаучная лексика. Список из 37 слов общенаучной лексики был сформирован нами на основе анализа двух частотных словарей [67, 68]. В список попали слова, у которых отношение частоты встречаемости в научных текстах (по данным [67]) к частоте "во всех текстах" (по данным [68]) больше 10. Список возглавляют слова (максимальное отношение): *уравнение, функция, плоскость, образовывать, формула, коэффициент*.

Полные списки корнеаффиксов, характерных для научного стиля; названий официальных документов и слов общенаучной лексики приведены в прил. 2.

2.4. Параметры опытного массива и их первичная статистическая обработка

В соответствии с методикой подсчета параметров всего было обработано 305 фрагментов текста. Общий объем обработанного материала – 239 696 слов, по 227 257 из них модулем морфологического анализа были построены нормальные формы (установлены грамматические характеристики). 7971 слово из общего объема состояло из латинских букв. Доля слов из русских букв, для которых модулем морфологического анализа не было построено ни одной нормальной формы, составляет по отдельным стилям:

- разговорный – 15 %;
- художественный – 2,7%;
- публицистический – 4,6%;
- научный – 6,9%;
- официально-деловой– 1,9%;

для всего опытного массива – 6,1%.

В дальнейшем мы используем следующие обозначения параметров (переменных в терминологии прикладной статистики) и стилей (соответственно – классов):

Переменные:

1. VERBS Глаголы
2. NOUNS Существительные
3. ADJECT Прилагательные
4. PRONOUNS Местоимения
5. INTERJEC Междометия
6. PREPOS Предлоги
7. CONJUNCT Союзы
8. PARTICLE Частицы
9. ADVERB Наречия
10. PARTICIP Причастия
11. GERUND Деепричастия
12. NUMERAL Числительные
13. NEUTRUM Существительные среднего рода
14. REFLEX Возвратные глаголы
15. ACRONYM Аббревиатуры
16. GENITIVE Пары существительных в родительном падеже
17. WORD_LEN Средняя длина слова в буквах
18. SENT_LEN Средняя длина предложения в словах
19. EXPRESS Предложения с экспрессивной пунктуацией
20. FORMUL Формулы (по знаку «=»)
21. SMILES «Улыбки» (Smiles)
22. I Местоимение я
23. YOU Местоимения ты + вы
24. WE Местоимение мы
25. NU_VOT Частицы ну, вот, ведь
26. BY Частицы бы
27. SUBORD Подчинительные союзы
28. SC_TERM Слова общенаучной лексики
29. ORDER Слова организации логики повествования
30. DOCS Названия официальных документов
31. SC_PREF Слова с научными корнеаффиксами приставочного типа

Стили:

CHAT Разговорный
LITER Художественный
PUBL Публицистический
SCIENCE Научный

LAW	Официально-деловой
TOTAL	Весь опытный массив

Данные, полученные с помощью нашей программы, подвергались предварительной обработке в программе Excel компании Microsoft (www.microsoft.com). Основная статистическая обработка данных производилась при помощи пакета программ STATISTICA компании StatSoft (www.statsoft.com).

В табл. 2.2 приведены результаты рассчета параметров опытного массива документов по каждому стилю: их минимальные (*min*), максимальные (*max*) и средние значения (x_{cp}), а также стандартные отклонения (*S*).

В прил. 3 содержатся категоризированные диаграммы типа «ящики с усами» для всех параметров по стилям (за исключением SMILES и FORMUL[1]). Точка соответствует среднему значению, «ящик» – стандартной ошибке, а «усы» – стандартному отклонению.

Кроме того, рассчитана выборочная матрица корреляции (табл. 2.3 содержит часть матрицы корреляции, соответствующую группе взаимно коррелированных параметров[2]) и проведен тест Шапиро-Уилкса на нормальность распределения [2] (табл. 2.4).

[1] Каждый из этих параметров «проявляется» только в одной из групп опытного массива (CHAT и SCIENCE соответственно), а в остальных группах равен нулю.
[2] В таблицу не вошли параметры SENT_LEN и GENITIVE, которые сильно связаны между собой (коэффициент корреляции равен 0,81) и слабо – с остальными параметрами.

Таблица 2.2

Параметры опытного массива документов

		VERBS	NOUNS	ADJECT	PRONOUN	NUMERAL	INTERJEC	PREPOS	CONJUNC	PARTICLE	ADVERB	PARTICIP	GERUND	NEUTRUM	REFLEX	ACRONYM	GENITIVE
CHAT	*min*	0,109	0,074	0,007	0,105	0,000	0,000	0,024	0,021	0,132	0,039	0,005	0,000	0,037	0,019	0,000	0,000
	max	0,219	0,298	0,078	0,222	0,009	0,045	0,084	0,085	0,276	0,102	0,050	0,042	0,278	0,286	0,027	0,136
	x_{cp}	0,167	0,194	0,033	0,161	0,002	0,016	0,051	0,050	0,210	0,068	0,028	0,006	0,151	0,156	0,001	0,044
	S	0,024	0,040	0,013	0,027	0,002	0,009	0,013	0,013	0,031	0,017	0,011	0,007	0,060	0,062	0,004	0,032
LITER	*min*	0,091	0,140	0,022	0,059	0,000	0,000	0,031	0,014	0,068	0,023	0,016	0,001	0,074	0,080	0,000	0,000
	max	0,239	0,351	0,106	0,227	0,019	0,018	0,085	0,063	0,242	0,118	0,074	0,028	0,313	0,303	0,023	0,391
	x_{cp}	0,161	0,244	0,064	0,126	0,006	0,003	0,055	0,037	0,158	0,065	0,045	0,013	0,179	0,171	0,001	0,105
	S	0,027	0,050	0,020	0,039	0,004	0,003	0,010	0,011	0,030	0,017	0,013	0,007	0,049	0,046	0,004	0,079
PUBL	*min*	0,079	0,265	0,061	0,036	0,000	0,000	0,023	0,023	0,068	0,022	0,030	0,000	0,109	0,055	0,000	0,098
	max	0,165	0,410	0,175	0,119	0,027	0,007	0,084	0,061	0,221	0,077	0,102	0,020	0,286	0,245	0,095	1,133
	x_{cp}	0,120	0,335	0,107	0,075	0,007	0,000	0,058	0,038	0,130	0,049	0,066	0,009	0,203	0,149	0,010	0,467
	S	0,019	0,034	0,024	0,019	0,005	0,001	0,013	0,008	0,029	0,012	0,017	0,005	0,045	0,039	0,017	0,219
SCIENCE	*min*	0,047	0,245	0,074	0,019	0,000	0,000	0,023	0,010	0,055	0,004	0,044	0,002	0,189	0,052	0,000	0,033
	max	0,129	0,499	0,199	0,081	0,019	0,024	0,101	0,117	0,156	0,087	0,141	0,041	0,754	0,302	0,049	1,796
	x_{cp}	0,090	0,396	0,130	0,047	0,005	0,001	0,061	0,033	0,089	0,030	0,092	0,016	0,336	0,152	0,004	0,576
	S	0,020	0,054	0,027	0,013	0,004	0,004	0,014	0,022	0,022	0,016	0,022	0,010	0,109	0,057	0,008	0,378
LAW	*min*	0,015	0,427	0,075	0,010	0,000	0,000	0,010	0,000	0,024	0,000	0,046	0,000	0,141	0,000	0,000	0,038
	max	0,100	0,608	0,278	0,052	0,061	0,000	0,087	0,032	0,136	0,037	0,140	0,026	0,450	0,305	0,064	4,625
	x_{cp}	0,048	0,497	0,184	0,029	0,009	0,000	0,046	0,009	0,071	0,008	0,091	0,005	0,291	0,108	0,006	1,169
	S	0,018	0,037	0,048	0,011	0,012	0,000	0,020	0,008	0,019	0,007	0,023	0,005	0,064	0,067	0,014	0,969
TOTAL	*min*	0,015	0,074	0,007	0,010	0,000	0,000	0,010	0,000	0,024	0,000	0,005	0,000	0,037	0,000	0,000	0,000
	max	0,239	0,608	0,278	0,227	0,061	0,045	0,101	0,117	0,276	0,118	0,141	0,042	0,754	0,305	0,095	4,625
	x_{cp}	0,123	0,321	0,097	0,093	0,006	0,004	0,054	0,034	0,136	0,047	0,062	0,010	0,224	0,150	0,004	0,423
	S	0,049	0,113	0,058	0,055	0,006	0,008	0,015	0,018	0,055	0,026	0,030	0,008	0,095	0,057	0,011	0,582

Продолжение табл. 2.2

		WORD_ LEN	SENT_ LEN	EXPRESS	FORMUL	SMILES	I	YOU	WE	NU_VOT	BY	SUBORD	SC_TERM	ORDER	DOCS	SC_PREF
CHAT	*min*	4,025	2,917	0,303	0,000	0,007	0,139	0,056	0,000	0,016	0,000	0,052	0,000	0,000	0,000	0,000
	max	5,531	7,293	0,787	0,058	0,343	0,489	0,446	0,071	0,177	0,050	0,388	0,006	0,005	0,008	0,010
	x_{cp}	4,754	4,245	0,607	0,001	0,161	0,314	0,259	0,014	0,084	0,010	0,148	0,001	0,000	0,001	0,001
	S	0,301	0,734	0,095	0,008	0,083	0,064	0,082	0,018	0,038	0,011	0,050	0,002	0,001	0,002	0,001
LITER	*min*	4,540	3,839	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,045	0,000	0,000	0,000	0,000
	max	6,295	24,227	0,942	0,000	0,000	0,406	0,415	0,087	0,134	0,052	1,276	0,005	0,006	0,016	0,005
	x_{cp}	5,306	11,466	0,231	0,000	0,000	0,201	0,086	0,012	0,030	0,014	0,393	0,001	0,001	0,001	0,001
	S	0,413	3,997	0,159	0,000	0,000	0,128	0,081	0,020	0,027	0,014	0,199	0,001	0,001	0,002	0,001
PUBL	*min*	5,514	11,040	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,304	0,000	0,000	0,000	0,000
	max	7,039	27,867	0,257	0,000	0,000	0,176	0,061	0,081	0,034	0,077	1,050	0,010	0,010	0,023	0,004
	x_{cp}	6,165	16,810	0,036	0,000	0,000	0,024	0,005	0,013	0,009	0,016	0,587	0,003	0,001	0,006	0,001
	S	0,290	3,308	0,061	0,000	0,000	0,040	0,013	0,023	0,011	0,019	0,172	0,002	0,002	0,006	0,001
SCIENCE	*min*	6,359	4,930	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,056	0,002	0,000	0,000	0,000
	max	8,150	21,182	0,197	0,923	0,000	0,111	0,000	0,150	0,000	0,057	1,061	0,118	0,015	0,034	0,047
	x_{cp}	7,102	12,588	0,013	0,087	0,000	0,024	0,000	0,019	0,000	0,003	0,350	0,045	0,002	0,006	0,009
	S	0,409	3,523	0,035	0,164	0,000	0,031	0,000	0,035	0,000	0,010	0,206	0,028	0,003	0,008	0,010
LAW	*min*	6,033	10,243	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,009	0,000
	max	8,354	73,933	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,222	0,048	0,000	0,108	0,009
	x_{cp}	7,217	25,076	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,213	0,012	0,000	0,036	0,001
	S	0,580	13,436	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,278	0,009	0,000	0,019	0,002
TOTAL	*min*	4,025	2,917	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	max	8,354	73,933	0,942	0,923	0,343	0,489	0,446	0,150	0,177	0,077	1,276	0,118	0,015	0,108	0,047
	x_{cp}	5,999	13,520	0,191	0,015	0,032	0,123	0,075	0,012	0,026	0,010	0,346	0,011	0,001	0,009	0,002
	S	1,031	8,960	0,246	0,076	0,074	0,145	0,113	0,023	0,038	0,014	0,244	0,020	0,002	0,015	0,005

Таблица 2.3

Коэффициенты корреляции

	VERBS	NOUNS	ADJECT	PRONOUN	PARTICLE	ADVERB	PARTICIP	WORD_LEN	EXPRESS	INTERJEC	SMILES	I	YOU	NU_VOT
VERBS	1,00	-0,88	-0,85	0,79	0,75	0,80	-0,75	-0,83	0,64	0,46	0,38	0,68	0,58	0,58
NOUNS	-0,88	1,00	0,85	-0,87	-0,86	-0,85	0,77	0,88	-0,74	-0,54	-0,50	-0,74	-0,66	-0,64
ADJECT	-0,85	0,85	1,00	-0,81	-0,79	-0,75	0,67	0,89	-0,71	-0,54	-0,48	-0,71	-0,66	-0,64
PRONOUN	0,79	-0,87	-0,81	1,00	0,77	0,70	-0,78	-0,87	0,76	0,53	0,53	0,78	0,74	0,66
PARTICLE	0,75	-0,86	-0,79	0,77	1,00	0,76	-0,77	-0,84	0,79	0,61	0,61	0,68	0,68	0,65
ADVERB	0,80	-0,85	-0,75	0,70	0,76	1,00	-0,69	-0,76	0,58	0,38	0,36	0,60	0,51	0,48
PARTICIP	-0,75	0,77	0,67	-0,78	-0,77	-0,69	1,00	0,76	-0,71	-0,50	-0,49	-0,69	-0,65	-0,63
WORD_LEN	-0,83	0,88	0,89	-0,87	-0,84	-0,76	0,76	1,00	-0,76	-0,58	-0,53	-0,77	-0,71	-0,68
EXPRESS	0,64	-0,74	-0,71	0,76	0,79	0,58	-0,71	-0,76	1,00	0,75	0,75	0,77	0,86	0,80
INTERJEC	0,46	-0,54	-0,54	0,53	0,61	0,38	-0,50	-0,58	0,75	1,00	0,70	0,61	0,71	0,86
SMILES	0,38	-0,50	-0,48	0,53	0,61	0,36	-0,49	-0,53	0,76	0,70	1,00	0,61	0,71	0,70
I	0,68	-0,74	-0,71	0,78	0,68	0,60	-0,69	-0,77	0,77	0,61	0,61	1,00	0,71	0,71
YOU	0,58	-0,66	-0,66	0,74	0,68	0,51	-0,65	-0,71	0,86	0,71	0,71	0,71	1,00	0,78
NU_VOT	0,58	-0,64	-0,64	0,66	0,65	0,48	-0,63	-0,68	0,80	0,86	0,70	0,71	0,78	1,00

Примечание. Выделены коэффициенты корреляции, по модулю большие 0,70.

Таблица 2.4

Результаты теста Шапиро-Уилкса на нормальность распределения

(W – значение статистики, α – соответствующий уровень значимости)

	CHAT(61)		LITER(79)		PUBL(61)		SCIENCE(54)		LAW(50)		TOTAL(305)	
	W	α	W	α	W	α	W	α	W	α	W	α
VERBS	0,9847	0,85	0,9887	0,94	0,9805	0,69	0,9696	0,33	0,9740	0,51	0,9602	0,00
NOUNS	0,9887	0,95	0,9741	0,34	0,9720	0,36	0,9466	0,03	0,9503	0,06	0,9522	0,00
ADJECT	0,9495	0,03	0,9671	0,14	0,9564	0,07	0,9606	0,14	0,9732	0,48	0,9292	0,00
PRONOUNS	0,9810	0,71	0,9580	0,04	0,9781	0,59	0,9821	0,78	0,9550	0,10	0,9208	0,00
NUMERAL	0,8384	0,00	0,9486	0,01	0,9267	0,00	0,8814	0,00	0,7366	0,00	0,7683	0,00
INTERJEC	0,9606	0,11	0,7237	0,00	0,4182	0,00	0,2535	0,00	--	--	0,6212	0,00
PREPOS	0,9661	0,20	0,9866	0,88	0,9826	0,78	0,9842	0,86	0,9653	0,25	0,9846	0,57
CONJUNC	0,9817	0,74	0,9744	0,35	0,9742	0,44	0,8269	0,00	0,9249	0,00	0,9741	0,02
PARTICLE	0,9822	0,76	0,9931	0,99	0,9680	0,24	0,9236	0,00	0,9613	0,18	0,9516	0,00
ADVERB	0,9477	0,02	0,9879	0,92	0,9801	0,68	0,9297	0,00	0,8672	0,00	0,9532	0,00
PARTICIP	0,9728	0,39	0,9771	0,46	0,9823	0,76	0,9679	0,29	0,9730	0,47	0,9517	0,00
GERUND	0,7619	0,00	0,9578	0,04	0,9609	0,11	0,9264	0,00	0,8552	0,00	0,8971	0,00
NEUTRUM	0,9692	0,28	0,9751	0,38	0,9627	0,14	0,8845	0,00	0,9754	0,56	0,9356	0,00
REFLEX	0,9788	0,62	0,9793	0,56	0,9769	0,54	0,9709	0,37	0,9424	0,03	0,9803	0,22
ACRONYM	0,2140	0,00	0,3150	0,00	0,6455	0,00	0,5373	0,00	0,5009	0,00	0,4471	0,00
GENITIVE	0,9131	0,00	0,8737	0,00	0,9361	0,00	0,9229	0,00	0,8012	0,00	0,6693	0,00
WORD_LEN	0,9864	0,90	0,9664	0,13	0,9808	0,71	0,9723	0,42	0,9665	0,28	0,9450	0,00
SENT_LEN	0,8910	0,00	0,9496	0,01	0,9334	0,00	0,9858	0,90	0,8063	0,00	0,8117	0,00
EXPRESS	0,9381	0,01	0,9061	0,00	0,6546	0,00	0,4520	0,00	--	--	0,7500	0,00
FORMUL	0,2116	0,00	--	--	--	--	0,5702	0,00	--	--	0,2412	0,00
SMILES	0,9597	0,10	--	--	--	--	--	--	--	--	0,4980	0,00
I	0,9832	0,80	0,8965	0,00	0,6641	0,00	0,7755	0,00	--	--	0,7758	0,00
YOU	0,9846	0,85	0,8722	0,00	0,4637	0,00	--	--	--	--	0,6994	0,00
WE	0,7898	0,00	0,6398	0,00	0,6287	0,00	0,6214	0,00	--	--	0,6067	0,00
NU_VOT	0,9662	0,20	0,8824	0,00	0,7549	0,00	--	--	--	--	0,7244	0,00
BY	0,8251	0,00	0,8853	0,00	0,8110	0,00	0,3353	0,00	--	--	0,4504	0,00
SUBORD	0,8726	0,00	0,9358	0,00	0,9650	0,18	0,8811	0,00	0,6997	0,00	0,9127	0,00
SC_TERM	0,7822	0,00	0,7622	0,00	0,8965	0,00	0,9438	0,02	0,8642	0,00	0,5876	0,00
ORDER	0,4872	0,00	0,7698	0,00	0,6269	0,00	0,6061	0,00	--	--	0,5475	0,00
DOCS	0,6878	0,00	0,6409	0,00	0,8382	0,00	0,7343	0,00	0,8993	0,00	0,6409	0,00
SC_PREF	0,4346	0,00	0,4580	0,00	0,5842	0,00	0,8288	0,00	0,4130	0,00	0,4580	0,00

2.5. Сокращение набора параметров

Интерпретация результатов вычисления параметров опытного массива с позиций стилистики не входит в задачи данной работы. В дальнейшем мы рассматриваем параметры почти исключительно формально, с точки зрения их значимости для процедуры классификации.

На основании анализа значений параметров документов опытного массива все параметры можно разделить на три группы:

1) параметры, средние значения которых слабо отличаются для разных стилей;
2) параметры, средние значения которых относительно плавно меняются от стиля к стилю;
3) параметры, значения которых резко выделяются в одном-двух стилях и примерно равны в других.

Параметры **первой группы** – это CONJUNCT, PREPOS (рис. 2.4), ORDER (рис. 2.3)[1]. Эти параметры как наименее информативные с точки зрения классификации можно исключить из первичного набора.

Параметры **второй группы** в основном отражают морфологический строй текста. К этой группе относятся: VERBS, NOUNS, ADJECT, PRONOUNS, PARTICLE, PARTICIP, ADVERB, NEUTRUM, GENITIVE, WORD_LEN, BY, ACRONYM.

В первую очередь хочется отметить результат, который связан с одновременным исследованием пяти функциональных стилей. Табл. 2.2 и диаграммы прил. 3 наглядно демонстрируют монотонный рост средней доли существительных (NOUNS) и прилагательных (ADJECT) и монотонное же уменьшение средней доли местоимений (PRONOUNS), наречий (ADVERB), глаголов (VERBS) и частиц (PARTICLE) от разговорного к официально-деловому стилю (см. рис. 2.2).

В целом значения этих морфологических параметров неплохо согласуются с результатами аналогичных исследований (см. п. 2.3.3)[2].

Третью группу составляют параметры: SMILES, INTERJECT, EXPRESS, I, YOU, NU_VOT, FORMUL, SC_TERM, SC_PREF, DOCS. Первые два характерны для разговорного стиля (CHAT). Параметры EXPRESS, I,

---

[1] Малые изменения значений параметров, соответствующих служебным частям речи (предлогам и союзам) вполне закономерны; а малые различия доли слов организации логики повествования (ORDER) можно объяснить неполнотой соответствующего списка. В официальных документах в качестве средства организации изложения используется рубрикация, а также представление предложений в виде списков, отсюда – полное отсутствие слов организации логики повествования.

[2] Если говорить о расхождениях, то полученное нами соотношение долей существительных и наречий в текстах художественного, публицистического и научного стилей качественно отличается от приведенного в

YOU, NU_VOT наиболее ярко проявляются в разговорном стиле, слабее – в художественном, и практически «молчат» в публицистическом, научном и официально-деловом. Следующие три параметра (FORMUL, SC_TERM, SC_PREF) – наоборот, – резко выделяются по значениям в научных текстах, а последний из этого списка (DOCS) – в официальных документах.

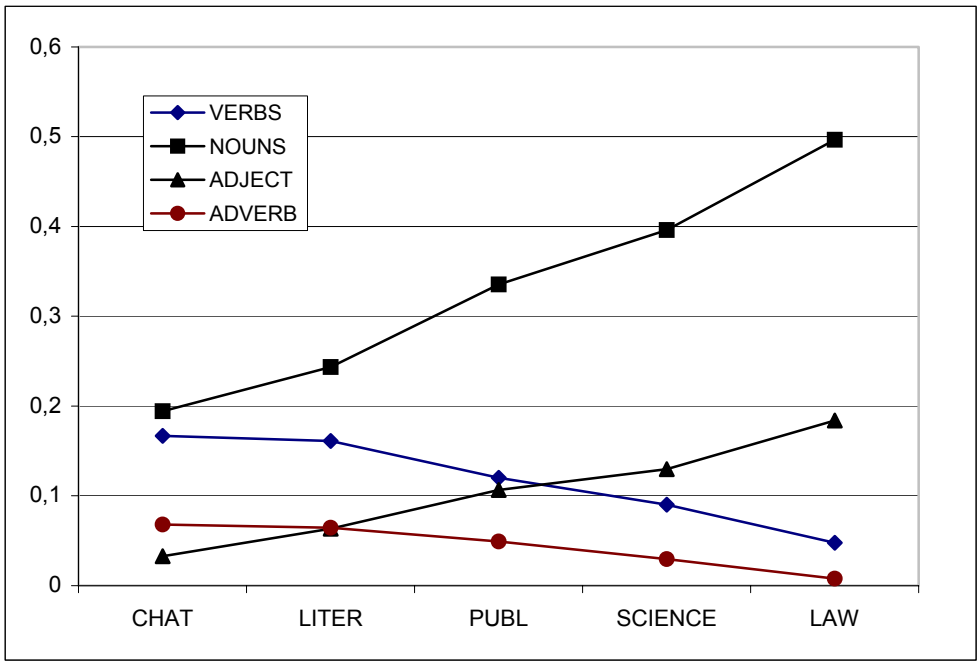

Рис. 2.2. Средние значения параметров VERBS, NOUNS, ADJECT и ADVERB по классам

В отдельную группу мы выделили параметры, поведение которых расходится с нашими предположениями и теорией. К ним относятся NUMERAL, GERUND, SUBORD, REFLEX. Несколько неожиданны «провалы» параметров NUMERAL в научных текстах и GERUND – в публицистических (рис. 2.3). Ожидалось, что доля возвратных глаголов (REFLEX) будет выше в текстах научного и официально-делового стилей (рис. 2.4). Расхождение между полученным и предполагаемым поведением параметра SUBORD можно отнести на счет неполноты списка подчинительных предлогов. Эти параметры мы также исключили из первичного набора.

[22, с.123]. Или, например, наши результаты не позволяют распространить на разговорный стиль Internet утверждение, что “характерная черта разговорной речи – малая употребительность наречий” [55, с. 139].

Кроме того, из набора оставшихся параметров мы удалили параметры с большой дисперсией (т.е. с большим уровнем шума с точки зрения классификации): ACRONYM, WE, BY (см. табл. 2.2 и диаграммы прил. 3).

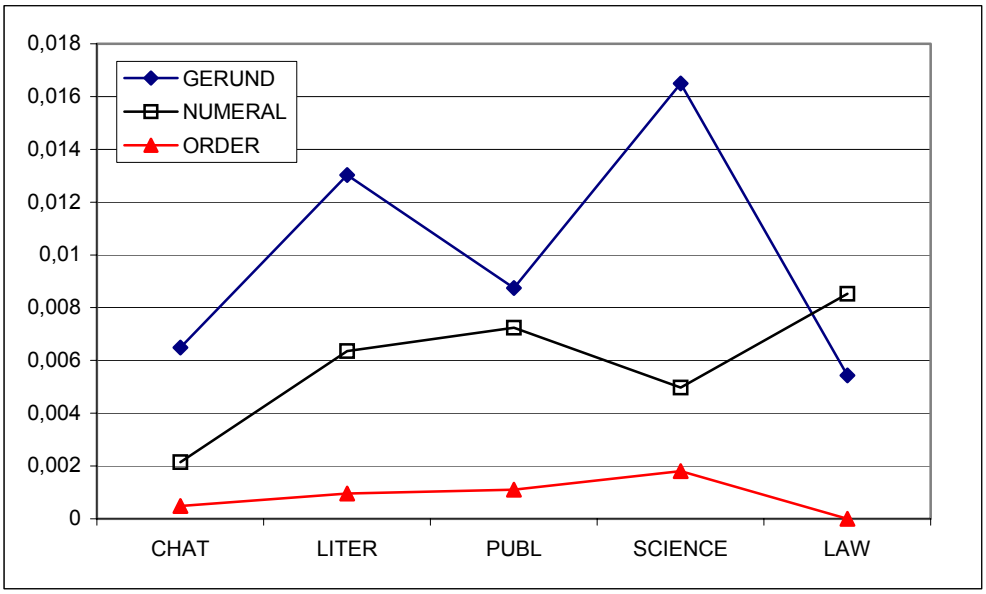

Рис. 2.3. Средние значения параметров GERUND, NUMERAL и ORDER по классам

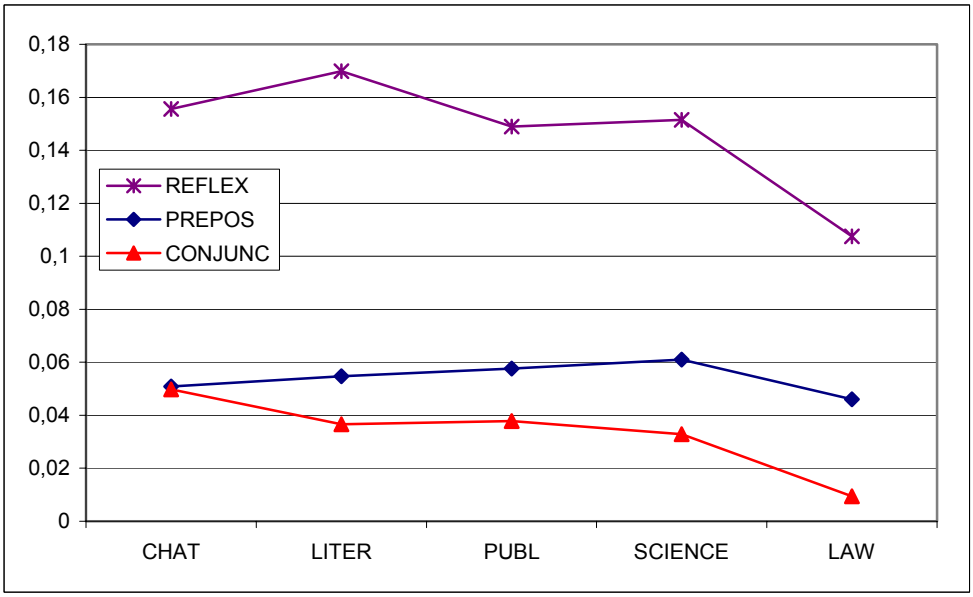

Рис. 2.4. Средние значения параметров REFLEX, PREPOS и CONJUNC по классам

Таким образом, из 31 параметра первичного набора на основании эвристических критериев мы отсеяли 10: NUMERAL, PREPOS, CONJUNC, GERUND, REFLEX, ACRONYM, WE, SUBORD, ORDER, BY, и оставили 21 параметр: VERBS, NOUNS, ADJECT, PRONOUNS, I, YOU, PARTICLE, NU_VOT, INTERJECT, ADVERB, PARTICIP, WORD_LEN, NEUTRUM,

GENITIVE, SENT_LEN, SC_TERM, SC_PREF, DOCS, SMILES, FORMUL, EXPRESS.

Анализ матрицы корреляции (табл. 2.3) позволяет выделить две группы взаимосвязанных параметров. Первую можно назвать «Морфологический строй»: VERBS, NOUNS, ADJECT, PRONOUNS, PARTICLE, PARTICIP, ADVERB, WORD_LEN. Параметры второй группы проявляются синхронно в текстах разговорного стиля: EXPRESS, INTERJECT, SMILES, I, YOU, NU_VOT. Коэффициенты корреляции параметров первой группы имеют разные знаки, второй – положительны. Это вполне закономерный результат[1].

Наличие групп взаимно коррелированных параметров говорит о возможности перехода к новому набору с меньшим числом параметров, полученных агрегированием параметров начального набора. Новый набор описывал бы объекты более лаконично, причем без потерь качества классификации. Такой переход можно осуществить с помощью процедур факторного анализа (метода главных компонент).

2.6. Классификация опытного массива документов

2.6.1. Условия применения процедур дискриминантного анализа

Процедуры дискриминантного анализа (ДА) основываются на некоторых достаточно сильных предположениях. Во-первых, предполагается, что закон распределения для каждого класса является многомерным нормальным, т.е. каждая переменная имеет нормальное распределение при фиксированных остальных переменных. Второе предположение состоит в том, что ковариационные матрицы наблюдаемых переменных равны между собой для разных классов.

---

[1] Как уже отмечалось выше, соотношение глаголов и существительных – фундаментальная морфологическая характеристика текста. Текст с преобладанием существительных нацелен на статическое описание объектов и отношений между ними. Большая доля глаголов характерна для текстов, описывающих действия, динамику, изменчивость. Каждое употребление существительного – это «повод» определить его прилагательным; действие, выраженное глаголом, можно охарактеризовать наречием; функция местоимения – заменять именные части речи и т.д.

Как видно из табл. 2.4, большинство параметров не отвечает первому предположению. И хотя значение М-критерия Бокса[1] (3808,12, что соответствует ничтожно малой ошибке первого рода) говорит о незначительном различии между элементами матриц ковариации в разных классах, относиться к этому результату надо скептически, т.к. критерий очень чувствителен к отклонениям от нормального закона.

Однако нарушение этих предположений не является препятствием для применения процедур ДА к полученным данным. В литературе указывается [9, 36, 74], что ДА является достаточно устойчивым методом, допускающим отклонение от этих предположений. Невыполнение указанных предположений приводит лишь к некоторым потерям в качестве классификации и точности оценок. В такой ситуации основная задача состоит в интерпретации результатов. Одним из главных аргументов в пользу применения метода является процент правильно классифицированных случаев.

2.6.2. Первая классификация документов опытного массива

В первом эксперименте по классификации опытного массива мы использовали 14 параметров из 21. Семь параметров[2] не могли быть включены в модель, т.к. они имеют нулевые дисперсии в одном или нескольких классах (стилях) (см. табл. 2.2).

В результате выполнения процедур ДА была получена первая дискриминантная функция:

Таблица 2.5

Коэффициенты первой дискриминантной функции

	CHAT	LITER	PUBL	SCIENCE	LAW
VERBS	597,25	629,61	589,98	544,13	518,61
NOUNS	551,01	564,89	586,38	561,07	640,89
ADJECT	-61,64	-45,94	-40,47	-90,92	16,39
PRONOUNS	822,09	785,51	749,94	734,44	742,13
PARTICLE	824,26	764,17	766,45	741,76	761,11

[1] В вычислении критерия участвуют параметры, стандартные отклонения которых отличны от нуля во всех классах.

[2] Параметры, не включенные в модель: I, YOU, SMILES, FORMUL, NU_VOT, EXPRESS, INTERJECT.

ADVERB	583,83	614,24	590,23	452,70	463,46
PARTICIP	665,60	692,28	733,18	811,63	775,83
NEUTRUM	-15,10	-4,03	6,14	33,28	41,21
GENITIVE	-28,96	-32,92	-32,95	-26,86	-29,72
WORD_LEN	56,85	57,43	60,25	67,46	58,84
SENT_LEN	1,45	1,82	1,88	1,36	1,85
SC_TERM	193,31	163,55	198,34	570,13	299,05
SC_PREF	-630,80	-644,93	-747,07	-201,63	-787,09
DOCS	-346,03	-364,83	-313,17	-292,69	55,63
Константа	-422,26	-425,05	-443,09	-482,43	-476,67

Полученная дискриминантная функция хорошо работает на документах опытного массива – ошибки составляют менее 10% (табл. 2.6). Хуже всего классифицируются документы художественного стиля (примерно четверть текстов классифицирована неправильно). Лучше всего классифицируются противоположности – документы разговорного и официально-делового стилей (ок. 2% ошибок).

Таблица 2.6

Классификационная матрица (первая классификация)

Стиль	разг.	худож.	публ.	Научн.	Оф.-дел.	Классифицировано правильно, %
Разговорный	60	1	0	0	0	98,36
Художественный	8	59	12	0	0	74,68
Публицистический	0	4	57	0	0	93,44
Научный	0	0	2	50	2	92,59
Оф. –деловой	0	0	1	0	49	98,00
Всего	68	64	72	50	51	90,16

Применение методов канонического дискриминантного анализа позволяет выявить геометрическую структуру классов (рис. 2.5). Два первых канонических направления объясняют 94% дисперсии (табл. 2.7). Первое каноническое направление (Root 1) отражает бóльшую часть стилистического разнообразия документов. Хотя четкие границы между стилями отсутствуют, пучки, соответствующие документам публицистического, художественного и разговорного стилей, располагаются последовательно вдоль первого канонического направления. Второе каноническое направление (Root 2) хорошо разделяет документы научного и официально-делового стилей, а вариации документов остальных стилей вдоль этого направления незначительны (см. табл. 2.8). Два других канонических направления (соответствуют меньшим собственным значениям) можно считать малоинформативными.

Таблица 2.7

Канонические направления

	Root 1	Root 2
VERBS	9,14	0,42
NOUNS	-5,34	-11,61
ADJECT	-2,55	-16,39
PRONOUNS	7,47	-1,15
PARTICLE	4,99	-2,56
ADVERB	15,52	-8,48
PARTICIP	-12,75	7,92
NEUTRUM	-5,33	0,10
GENITIVE	-0,26	0,72
WORD_LEN	-0,62	1,44
SENT_LEN	0,00	-0,09
SC_TERM	-26,89	52,90
SC_PREF	-12,58	100,72
DOCS	-28,36	-43,16
Константа	5,15	-2,47
Собст. значение	15,44	3,55
Накопленная доля дисперсии	0,77	0,94

Таблица 2.8

Средние значения канонических направлений по стилям

	Root 1	Root 2
CHAT	4,65	0,31
LITER	2,96	-0,28
PUBL	0,00	-0,78
SCIENCE	-4,30	3,45
LAW	-5,71	-2,72

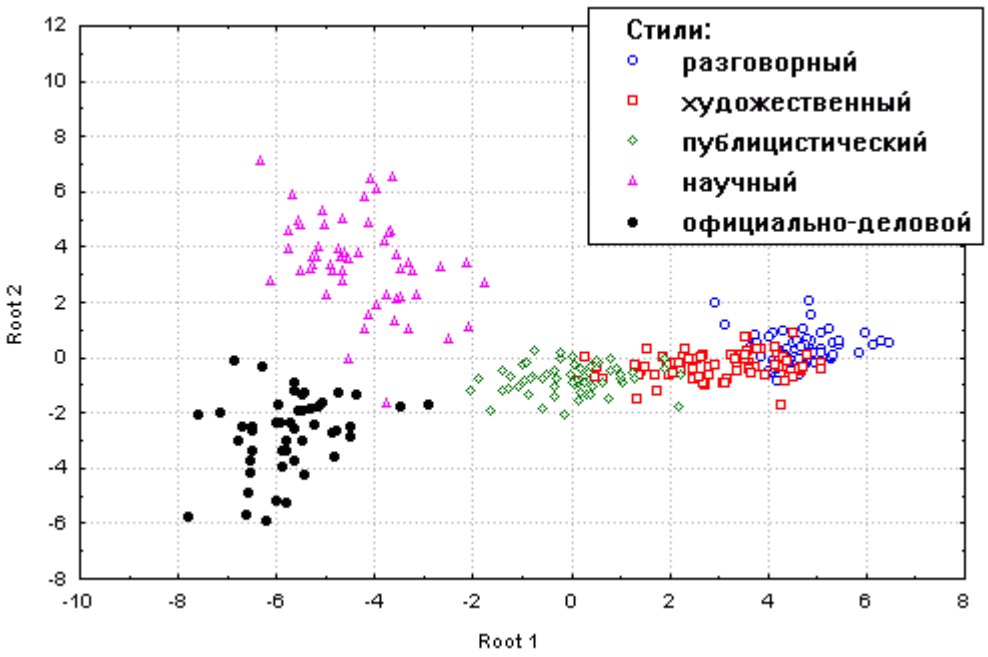

Рис. 2.5. Объекты опытного массива в канонических координатах (Root 1 – первое каноническое направление, Root 2 – второе каноническое направление)

Коэффициенты дискриминантной функции (табл. 2.5) говорят о неодинаковом вкладе параметров в процедуру классификации. Так, наименьший вклад вносят параметры NEUTRUM и GENITIVE (с учетом масштаба).

Использование пошаговых процедур ДА позволяет получить дискриминантные функции с меньшим числом переменных. Например, практически то же качество классификации опытного массива в целом можно получить, используя набор всего из семи параметров: VERBS, ADVERB, WORD_LEN, SENT_LEN, SC_TERM, SC_PREF, DOCS (табл. 2.9). Доля объектов опытного массива, правильно классифицированных с помощью этой функции, составляет 89,84%. Т.е. по сравнению с первой функцией (табл. 2.5) – на одну ошибку больше (при несколько другой «структуре ошибок» по стилям-классам).

Однако такое сокращение практически не уменьшает вычислительную сложность ни процедуры классификации, ни процедуры вычисления параметров. Процедура классификации является линейной, а число морфологических параметров не влияет на сложность их вычисления с помощью модуля

LINGUIST (вычисление одного морфологического параметра так же «затратно», как и всего списка).

Таблица 2.9

Коэффициенты оптимизированной дискриминантной функции

	CHAT	LITER	PUBL	SCIENCE	LAW
VERBS	458,75	471,56	408,92	367,12	287,77
ADVERB	318,61	313,31	275,27	173,79	122,34
WORD_LEN	37,47	40,60	44,98	50,59	48,41
SENT_LEN	0,09	0,25	0,29	0,11	0,40
SC_TERM	8,95	23,62	67,34	436,48	190,42
SC_PREF	-252,01	-292,64	-393,37	157,17	-423,40
DOCS	-238,23	-244,43	-197,99	-201,34	160,86
Константа	-139,79	-158,25	-173,39	-211,02	-192,68

Анализ результатов (рис. 2.5, табл. 2.7 и 2.8) позволяет сделать вывод, что первое каноническое направление можно рассматривать в качестве показателя «стилистической информативности» документа.

2.6.3. Главные компоненты – новые параметры классификации

Как указывалось выше, наличие групп сильно взаимосвязанных переменных делает целесообразным применение метода главных компонент [10, 34, 51].

Главными компонентами многомерного признака называют систему ортонормированных линейных комбинаций исходных признаков. При этом первая главная компонента обладает наибольшей дисперсией. Геометрически это означает, что первая главная компонента ориентирована вдоль направления наибольшей вытянутости эллипсоида рассеивания исследуемой совокупности. Вторая главная компонента ортогональна первой, ее направление выбирается по тому же принципу максимизации дисперсии. Она представляет собой проекцию на направление наибольшей вытянутости наблюдений в гиперплоскости, перпендикулярной первой главной компоненте. Третья главная компонента ортогональна первой и второй, и т.д. Задача нахождения

главных компонент сводится к вычислению собственных значений и собственных векторов матрицы ковариации исследуемой совокупности [51].

Переход к новым переменным приводит к более лаконичному описанию объектов, а также может рассматриваться как метод проецирования многомерных данных в пространство меньшей размерности. Новые переменные можно использовать в качестве новых параметров классификации. Причем при естественности классов (наличии их «в природе») качество классификации не ухудшается. В нашем случае агрегирование параметров к тому же позволяет неявно использовать параметры с нулевой дисперсией в одном или нескольких классах в процедуре дискриминантного анализа.

Исходя из геометрической структуры классов (стилей), полученной на предыдущем этапе с использованием канонического дискриминантного анализа, мы можем ограничиться двумя первыми главными компонентами. Такое решение согласуется с эвристическими критериями для определения числа факторов (критерием «каменистой осыпи» и критерием Кайзера [9, 36]).

Полученные нами главные компоненты соответствуют собственным значениям 11,9 и 1,8; накопленная доля общей дисперсии – 65,2%. В дальнейшем мы будем называть их *факторами* и обозначать Factor 1 и Factor 2.

По значениям факторных нагрузок (табл. 2.10, рис. 2.6) видно, что первый фактор (Factor 1) формируется в основном за счет положительного вклада параметров VERBS, PRONOUNS, INTERJEC, PARTICLE, ADVERB, EXPRESS, I, YOU, NU_VOT (т.е. тех, которые отражают «разговорность» текста) и отрицательного вклада параметров NOUNS, ADJECT, PARTICIP, WORD_LEN, SENT_LEN (соответствуют «информативности», «строгости» текста). Разнонаправленный вклад в формирование второго фактора (Factor 2) вносят параметры GENITIVE, SENT_LEN, DOCS (параметры, характерные для официально-делового стиля) – с одной стороны, и SC_TERM, SC_PREF, FORMUL (признаки научного стиля) – с другой.

Таблица 2.10

Факторные нагрузки

	Factor 1	Factor 2
VERBS	0,87	-0,20
NOUNS	-0,92	0,17
ADJECT	-0,89	0,18
PRONOUNS	0,90	0,01
INTERJEC	0,70	0,21
PARTICLE	0,89	0,02
ADVERB	0,80	-0,22
PARTICIP	-0,84	-0,04
NEUTRUM	-0,62	-0,29
GENITIVE	-0,66	0,44
WORD_LEN	-0,93	-0,03
SENT_LEN	-0,70	0,42
EXPRESS	0,88	0,15
FORMUL	-0,17	-0,46
SMILES	0,66	0,21
I	0,85	0,07
YOU	0,82	0,18
NU_VOT	0,80	0,21
SC_TERM	-0,45	-0,61
SC_PREF	-0,28	-0,42
DOCS	-0,59	0,46
Собств. знач.	11,90	1,79
Пропорц. доля дисп.	0,57	0,09

Средние значения первого фактора монотонно уменьшаются от разговорного к официально-деловому стилю (рис. 2.7). Средние значения второго фактора мало варьируются в разговорном, художественном и публицистическом стилях, но существенно различаются в официально-деловом и научном (рис. 2.8).

Геометрическая структура классов (стилей), которая выявляется на диаграмме рассеяния объектов опытного массива в новых координатах (рис. 2.9), аналогична полученной при помощи канонического ДА (рис. 2.5). Это еще раз подтверждает, что различия между объектами (документами) обусловлены в основном различиями между классами (стилями). Однако в данном случае границы между публицистическим и художественным, художественным и разговорным стилями более четкие. Это достигается за счет использования всего набора параметров (включая параметры, характерные для разговорного стиля: I, YOU, SMILES, NU_VOT, EXPRESS, INTERJECT).

Анализ результатов (рис. 2.6, 2.7 и 2.9, табл. 2.10) позволяет сделать вывод, что первый фактор может наравне с первым каноническим направлением претендовать на роль меры «стилистической информативности» документа.

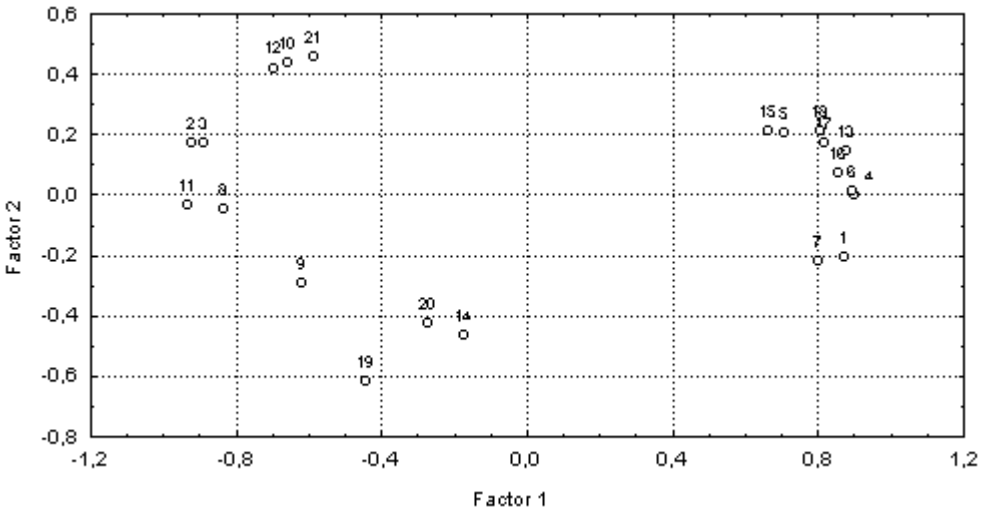

Рис. 2.6. Диаграмма факторных нагрузок.
1 - VERBS; 2 - NOUNS; 3 - ADJECT; 4 - PRONOUNS; 5 - INTERJEC; 6 - PARTICLE; 7 - ADVERB; 8 - PARTICIP; 9 - NEUTRUM; 10 - GENITIVE; 11 - WORD_LEN; 12 - SENT_LEN; 13 - EXPRESS; 14 - FORMUL; 15 - SMILES; 16 - I; 17 - YOU; 18 - NU_VOT; 19 - SC_TERM; 20 - SC_PREF; 21 - DOCS

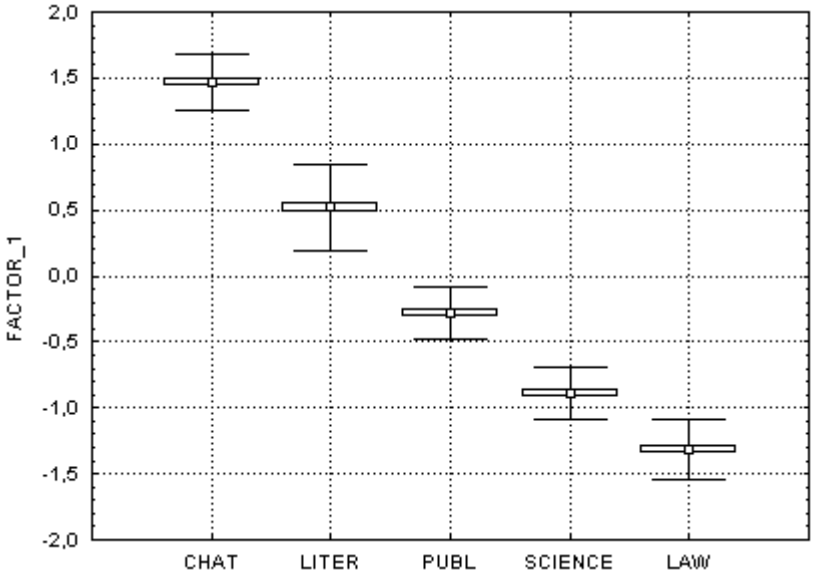

Рис. 2.7. Диаграмма первого фактора (Factor 1) по стилям (классам).
Точка – среднее, «ящик» - стандартная ошибка, «усы» - стандартное отклонение

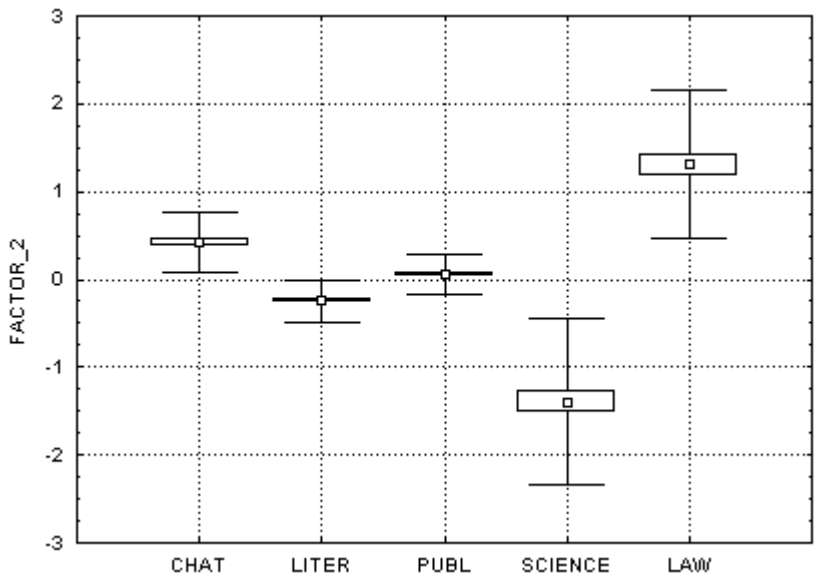

Рис. 2.8. Диаграмма второго фактора (Factor 2) по стилям (классам). Точка – среднее, «ящик» - стандартная ошибка, «усы» - стандартное отклонение

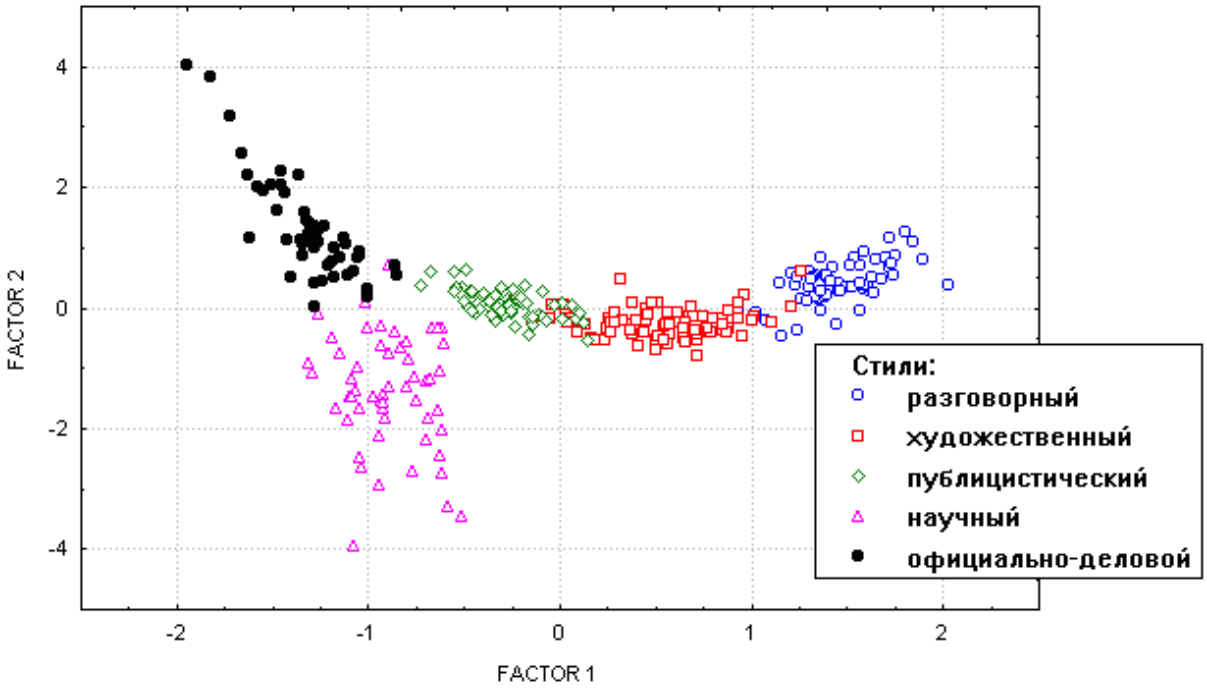

Рис. 2.9. Документы опытного массива в факторных координатах

Так же, как и первичный набор параметров, новые факторы не отвечают строгим предположениям дискриминантного анализа (табл. 2.11).

Используя полученные факторы в качестве новых параметров классификации, мы получили вторую дискриминантную функцию (табл. 2.12).

Таблица 2.11

Результаты теста Шапиро-Уилкса на нормальность распределения

(W – значение статистики, α – соответствующий уровень значимости)

	CHAT(61)		LITER(79)		PUBL(61)		SCIENCE(54)		LAW(50)		TOTAL(305)	
	W	α	W	α	W	α	W	α	W	α	W	α
Factor 1	0, 9910	0,98	0,9827	0,72	0,9766	0,53	0,9565	0,09	0,9739	0,51	0,9311	0,00
Factor 2	0,9816	0,74	0,9653	0,11	0,9809	0,71	0,9750	0,51	0,8915	0,00	0,9385	0,00

Таблица 2.12

Коэффициенты второй дискриминантной функции

	CHAT	LITER	PUBL	SCIENCE	LAW
FACTOR 1	26,11	8,67	-4,77	-17,48	-20,85
FACTOR 2	3,53	-0,02	-0,22	-5,79	2,25
Константа	-21,51	-3,61	-2,27	-13,54	-16,95

Качество классификации опытного массива в целом несколько повысилось, однако ухудшилось, например, качество классификации документов научного стиля: 92,6% в первом опыте и 87% – во втором (табл. 2.13).

Таблица 2.13

Классификационная матрица (вторая классификация)

Стиль	разг.	худож.	публ.	научн.	оф.-дел.	Классифицировано правильно, %
разговорный	59	2	0	0	0	96,72
художественный	3	65	11	0	0	82,28
публицистический	0	2	59	0	0	96,72
научный	0	0	4	47	3	87,04
оф. -деловой	0	0	0	0	50	100,00
всего	62	69	74	47	53	91,80

Канонические направления, построенные по новым данным, практически совпадают (с точностью до знака) с первыми главными компонентами. Это видно из структуры корреляции между факторами и каноническими направлениями:

Таблица 2.14

	Root 1	Root 2
FACTOR_1	-0,97	0,24
FACTOR_2	-0,04	-1,00

Этот результат еще раз подтверждает, что дисперсия объясняется в основном межклассовыми различиями объектов.

2.7. Результаты и выводы

В процессе решения задачи автоматической классификации текстовых документов по стилям были получены следующие результаты.

1. Сформулированы подходы к решению задачи. За основу принята функционально-стилевая концепция и соответствующая ей система пяти функциональных стилей. В задачах информационного поиска мы предложили использовать шкалу «стилистическая информативность». В качестве метода построения классификации был выбран дискриминантный анализ (ДА). Сформирован опытный массив из 305 документов («обучающая выборка»). Сформирован первичный набор параметров классификации (31 параметр), определена методика вычисления их значений.

2. Получены значения параметров первичного набора для документов опытного массива, проведена их первичная статистическая обработка. На основании анализа полученных данных с помощью эвристических критериев из первичного набора были исключены 10 параметров.

3. В результате применения процедур дискриминантного анализа и метода главных компонент к выборочным данным получены дискриминантные функции с хорошим качеством классификации. Выявлена геометрическая структура классов-стилей. На основании параметров первичного набора получен агрегированный параметр «стилистическая информативность документа».

4. Разработана программа, реализующая на уровне макета предложенные методы. Программа обрабатывает документы в формате plain text или HTML (кодировка Windows), относит их к одному из пяти функциональных стилей и рассчитывает значение «стилистической информативности» (на основе первой дискриминантной функции и первого канонического направления).

3. Расширение запроса с помощью тезауруса

3.1. Описание метода

Из предыдущего обзора (п. 1.3.2) видно, что тезаурусы, несмотря на богатую традицию использования в информационном поиске, находят лишь ограниченное применение в современных универсальных ИПС Internet. Во-первых, это объясняется тем, что задача экономии памяти и вычислительных ресурсов сегодня не так актуальна, как 30 лет назад. Во-вторых, чрезвычайно трудно построить тезаурус, который соответствовал бы тематическому разнообразию информации, индексируемой универсальной МП. Кроме того, в традиционных ИПС тезаурусы использовались в основном как средство повышения *полноты* поиска (объединение близких терминов по «ИЛИ») [58]. Сегодняшняя проблема Internet – это информационная перегрузка (*information overload*), и как следствие – низкая *точность* поиска.

На наш взгляд, тезаурус может стать эффективным инструментом формирования запросов к универсальным МП Internet и существенно повысить качество информационного поиска в специализированной тематической области. Для этого должны выполняться следующие условия:

- тезаурус отражает терминологию достаточно узкой предметной области;
- в тезаурусе используется набор сильно дифференцированных семантических отношений;
- тезаурус располагается на стороне пользователя (клиента), “до” программы просмотра (см. рис. 3.1, ср. рис. 1.1).

Задача построения универсального тезауруса является очень трудоемкой. Построение тезауруса узкой предметной или научной области осуществить намного проще. С одной стороны, задача намного меньше по объему. С другой стороны, в данном случае мы имеем дело не с *общеупотребительной лексикой*, а с *терминами*.

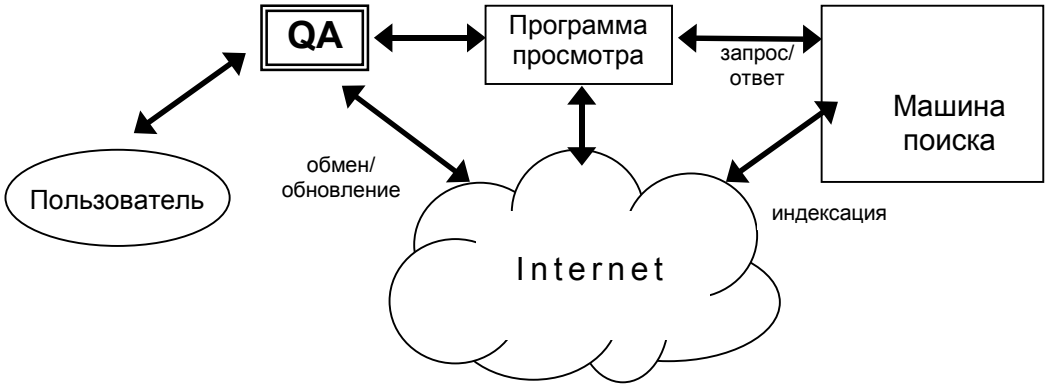

Рис. 3.1. Место ассистента формирования запросов (QA) в общей структуре информационного поиска

По определению [35], термин – это “слово или словесный комплекс, соотносящийся с понятием определенной организованной области познаний (науки, техники), вступающее в системные отношения с другими словами и словесными комплексами и образующие вместе с ними в любом отдельном случае и в определенное время замкнутую систему, отличающуюся высокой информативностью, однозначностью, точностью и экспрессивной нейтральностью.”

Образования более высокого уровня, о которых идет речь в определении выше, называются *терминосистемами*. Терминосистема – это “сложная динамическая устойчивая система, элементами которой являются ... лексические единицы какого-нибудь естественного языка, структура которой изоморфна структуре логических связей между понятиями специальной области знаний и деятельности, а функция состоит в том, чтобы служить знаковой (языковой) моделью этой области знаний и деятельности” [43].

Свойства терминов – системность, устойчивость и регулярность связей, отсутствие экспрессии, установка на объективность описания, – делают возможным их адекватное описание с помощью тезаурусов.

Особенно точно описать терминологию можно при помощи тезауруса с набором *сильно дифференцированных семантических отношений* [46, 47]. Основная идея такого описания - использование не только универсальных (например, “род-вид”, “часть-целое” и т.д.), но и специфических для конкретной предметной области отношений. Таким образом, каждый *тип отно-*

шения сам по себе несет значительную смысловую нагрузку, определяет различные аспекты семантики термина.

В прил. 5 содержится описание тезауруса по компьютерной лингвистике, в котором используется набор из 25 типов семантических отношений.

На основе набора семантических отношений тезауруса можно ввести понятие *стратегии поиска по тезаурусу*. Стратегия – это шаблон с указанием связки (“И”, “ИЛИ”, “НЕ”) и веса для каждого типа семантического отношения. Выбрав термин и применив к нему стратегию, мы получаем запрос, в котором выбранный («опорный») термин объединен со своими “соседями” в соответствии с маской-стратегией. Стратегии могут быть направлены на повышение точности или полноты поиска, выделение определенных понятийных сфер термина. Например, для выделения *сферы процессуальности* в стратегию можно включить отношения “процесс/операция”, “связь процессов”; и т.д. (см. прил. 5). Сформировав стратегию, ее можно применять последовательно к различным опорным терминам. Стратегии могут служить подсказкой начинающему пользователю, позволяют унифицировать поиск, сделать его в большей степени автоматизированным.

Средство формирования запросов на основе тезауруса позволяет тонко управлять как *полнотой*, так и *точностью* поиска.

Разнообразие, специфичность и динамика тематических интересов и информационных запросов пользователей ставит под вопрос эффективность централизованной разработки тезаурусов и расположения их на МП. Тезаурусный ассистент формирования запросов (QA) целесообразно разместить на стороне пользователя (рис. 3.1).

Заметим, что Internet в этой схеме выступает не только как хранилище информации, но и как среда для коммуникации и объединения усилий разработчиков и пользователей.

Пользователями средства расширения запросов на основе тезауруса могли бы стать как непосредственные потребители информации (специалисты-предметники), так и те, кто занимается структурированием информаци-

онного сырья Internet: аналитические службы; специалисты, занимающиеся наполнением Internet-каталогов (*information brokers*) и т.п.

Такой инструмент может также пригодится тем, кто приступает к изучению новой области знаний, причем эффект возникает не только при изучении результатов поиска, но и при работе с самим тезаурусом – его терминами, отношениями и определениями.

В следующем пункте мы излагаем первичную формализацию модели тезауруса и процедуры формирования запроса. За основу взят синтаксис запросов системы “Яндекс” (см. табл. 1.1).

3.2. Модель тезауруса и процедура формирования запросов

Следуя [25], определим модель тезауруса как упорядоченную тройку:

$$T = \langle A, R, \mathfrak{R} \rangle, \qquad (2)$$

где

A – непустое множество терминов (носитель модели),

R – множество типов (символов) бинарных отношений (сигнатура модели),

$\mathfrak{R}$ – множество бинарных отношений на множестве A, причем имеется отображение множества R в множество $\mathfrak{R}$: $r \in R \Rightarrow \rho(r) \in \mathfrak{R}$ (интерпретация сигнатуры).

Из семантических ограничений следует, что все отношения $\mathfrak{R}$ нерефлексивны (термин не связан с самим собой).

Множество типов отношений должно оптимально соответствовать терминосистеме, откуда следует, что термин не может быть связан с другим более чем одним типом отношения:

$$\bigcup_{\substack{\rho_1 \neq \rho_2 \\ \{\rho_1, \rho_2\} \in \mathfrak{R}}} (\rho_1 \cap \rho_2) = \varnothing.$$

Пример 1. В качестве примера приведем фрагмент тезауруса, соответствующего предметной области «Диагностика и ремонт автомобилей».

A = {автомобиль, ‘транспортное средство’, двигатель, мотор, трансмиссия, диагностика, ‘диагностический протокол’, ‘коробка передач’}

R = {род, вид, часть, целое, синоним, процесс/действие, ‘объект процесса/действия’, ‘результат процесса/действия’, ‘непосредственная связь’}

Множество отношений $\Re$ состоит из следующих элементов:

РОД = {(автомобиль, ‘транспортное средство’)}

ВИД = {(‘транспортное средство’, автомобиль)}

НЕПОСРЕДСТВЕННАЯ СВЯЗЬ = {(двигатель, трансмиссия), (трансмиссия, двигатель)}

ЧАСТЬ = {(автомобиль, трансмиссия), (автомобиль, двигатель), (трансмиссия, ‘коробка передач’)}

ЦЕЛОЕ = {(трансмиссия, автомобиль), (двигатель, автомобиль), (‘коробка передач’, трансмиссия)}

ПРОЦЕСС/ДЕЙСТВИЕ = {(автомобиль, диагностика), (‘диагностический протокол’, диагностика)}

РЕЗУЛЬТАТ ПРОЦЕССА/ДЕЙСТВИЯ = {(диагностика, ‘диагностический протокол’)}

ОБЪЕКТ ПРОЦЕССА/ДЕЙСТВИЯ = {(диагностика, автомобиль)}

СИНОНИМ = {(мотор, двигатель), (двигатель, мотор)}

Наглядно термины и семантические связи между ними можно представить в виде графа (рис. 3.2).

Определим стратегию поиска по тезаурусу как упорядоченную четверку:

$$S = \langle w, R_s, \varphi_s, f_s \rangle, \qquad (3)$$

где $w \in \boldsymbol{N}$ – вес опорного термина;

$R_s \subseteq R$ – множество типов связей, участвующих в стратегии;

$\varphi_s : R_s \to \{\&, |, \sim\}$ – функция, ставящая в соответствие каждому типу отношения из R_s тип связки;

$f_s : R_s \to \boldsymbol{N}$ – функция, ставящая в соответствие каждому типу отношения из R_s вес.

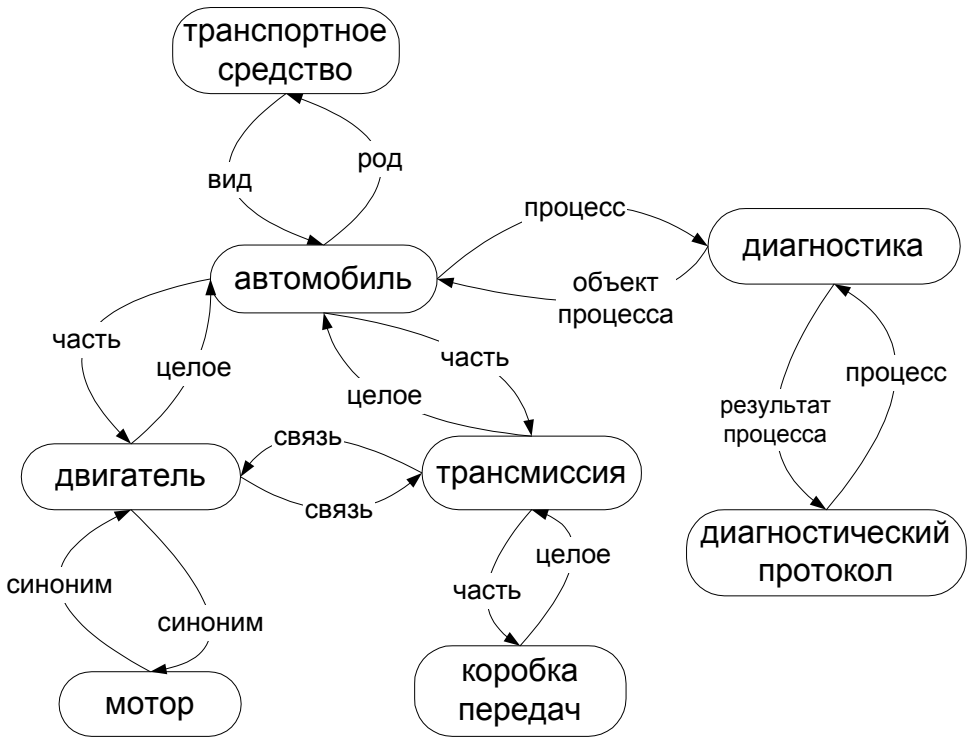

Рис. 3.2. Граф тезауруса из примера 1

Пример 2. Стратегию для тезауруса из примера 2 можно задать таблицей:

w	R_s	φ_s	f_s
10	часть	&	5
	связь	&	5
	синоним	|	8

Алфавит ассистента формирования запросов содержит термины тезауруса, натуральные числа, символы связок &, |, ~, а также символы двоеточия : и скобок (,).

Слово в алфавите ассистента формирования запросов является правильным запросом, если оно удовлетворяет следующему определению:

1. Любой термин тезауруса – правильный запрос. (Если $x \in$ A, то x – правильный запрос.)
2. Любой термин тезауруса с указанием веса – правильный запрос. (Если $x \in$ A, $w \in \boldsymbol{N}$, то $x{:}w$ – правильный запрос.)
3. Если Q1 и Q2 – правильные запросы, то (Q1&Q2), (Q1|Q2), (Q1~Q2) – правильные запросы.

Для упрощения записи можно опускать внешние скобки запроса, а также те пары скобок, без которых старшинство связок правильно определяет порядок обработки запроса.

Пример 3. Правильные запросы на основе терминов тезауруса из примера 1:

- (автомобиль|диагностика)&трансмиссия
- (трансмиссия:10|‘коробка передач’:8)&(двигатель:6|мотор:4)

По аналогии со знаком суммы **Σ** в арифметике обозначим через $\underset{P}{\mathbf{OR}}\, x$, $\underset{P}{\mathbf{AND}}\, x$, $\underset{P}{\mathbf{NOT}}\, x$ серии запросов на основе тезауруса x, удовлетворяющих условию P и объединенных связками |, &, ~ соответственно.

Пример 4. Объединение всех «соседей» термина *автомобиль* связкой | можно обозначить следующим образом:

$$\underset{\substack{(\text{автомобиль},\, x) \in \rho \\ \rho \in \Re}}{\mathbf{OR}\, x} = \text{‘транспортное средство’|диагностика|двигатель|трансмиссия}$$

Тогда запрос Q, построенный как результат применения стратегии S (3) к опорному термину a*, выглядит следующим образом:

$$Q(\text{a*}, \text{S}) = \text{a*}{:}w \mid \underset{\substack{(\text{a*},\, \text{a}) \in \rho(\text{r}) \\ \varphi_s(\text{r}) = |}}{\mathbf{OR}\ \text{a}{:}f_s(r)} \ \& \ \underset{\substack{(\text{a*},\, \text{a}) \in \rho(\text{r}) \\ \varphi_s(\text{r}) = \&}}{\mathbf{AND}\ \text{a}{:}f_s(r)} \sim \underset{\substack{(\text{a*},\, \text{a}) \in \rho(\text{r}) \\ \varphi_s(\text{r}) = \sim}}{\mathbf{NOT}\ \text{a}{:}f_s(r)}. \quad (4)$$

Пример 5. Запрос, полученный в результате применения стратегии из примера 2 к опорному термину *автомобиль* тезауруса из примера 1, выглядит следующим образом:

автомобиль:10&двигатель:5&трансмиссия:5.

Та же стратегия, примененная к термину *двигатель*, дает следующий результат:

двигатель:10|мотор:8&трансмиссия:5.

Две стратегии S_1 и S_2 назовем совместимыми, если либо они не пересекаются по задействованным в них типам связей ($R_{S1} \cap R_{S2} = \varnothing$), либо значения функций φ_{S1}, φ_{S2} совпадают на этом пересечении ($\varphi_{S1}(x) = \varphi_{S2}(x)$, $\forall x \in R_{S1} \cap R_{S2}$).

Для двух совместимых стратегий S_1 и S_2 следующим образом можно определить объединение S:

$$S = S_1 \oplus S_2 = \langle w_1 + w_2,\ R_{S1} \cup R_{S2},\ \varphi_s, f_s \rangle, \quad (5)$$

где функции φ_s и f_s определяются так:

$$\varphi_s(x) = \begin{cases} \varphi_{s_1}(x), x \in R_{s_1} / R_{s_2} \\ \varphi_{s_2}(x), x \in R_{s_2} \end{cases},$$

$$f_s(x) = \begin{cases} f_{s_1}(x), x \in R_{s_1} / R_{s_2} \\ f_{s_1}(x) + f_{s_2}(x), x \in R_{s_1} \cap R_{s_2} \\ f_{s_2}(x), x \in R_{s_2} / R_{s_1} \end{cases}.$$

Пример 6. Объединение стратегии из примера 2 со стратегией:

w	R_s	φ_s	f_s
2	связь	&	1
	целое	|	2

выглядит следующим образом:

w	R_s	φ_s	f_s
12	часть	&	5
	целое	|	2
	связь	&	6
	синоним	|	8

&-стратегией назовем стратегию, в которой $\varphi_s \equiv$& на всем множестве R_s. Аналогично определяются |- и ~-стратегии. &-, |- и ~-стратегии будем называть монотонными.

Очевидно, что любую стратегию ($w \geq 3$) можно представить в виде объединения монотонных стратегий.

Операция (5) позволяет сформировать набор элементарных стратегий, а новые получать в виде объединения элементарных. Операция объединения стратегий может быть продуктивна при интеграции нескольких тезаурусов.

3.3. Программная реализация

Идеи, изложенные выше, реализованы на уровне макета в программе *ProThes Q*. Программа разработана в среде Delphi.

Интерфейс программы *ProThes Q* содержит три окна: “Термин”, “Запрос”, “Стратегия”.

Окно «Термин» (рис. 3.3) предназначено для навигации по тезаурусу и выбора «опорного» термина. В окне отображаются: алфавитный список терминов; дефиниция термина (если есть) и его «соседи».

Перемещаться по терминам тезауруса можно:

1) через поиск терминов;
2) по алфавитному списку;
3) по семантическим связям между терминами.

Окно “Стратегия” (рис. 3.4) предназначено для формирования стратегий поиска. Каждому типу отношения назначается тип связки: «И», «ИЛИ», «НЕ» (или «пусто», если тип отношения не участвует в стратегии). Указывается вес опорного термина и каждого типа отношения.

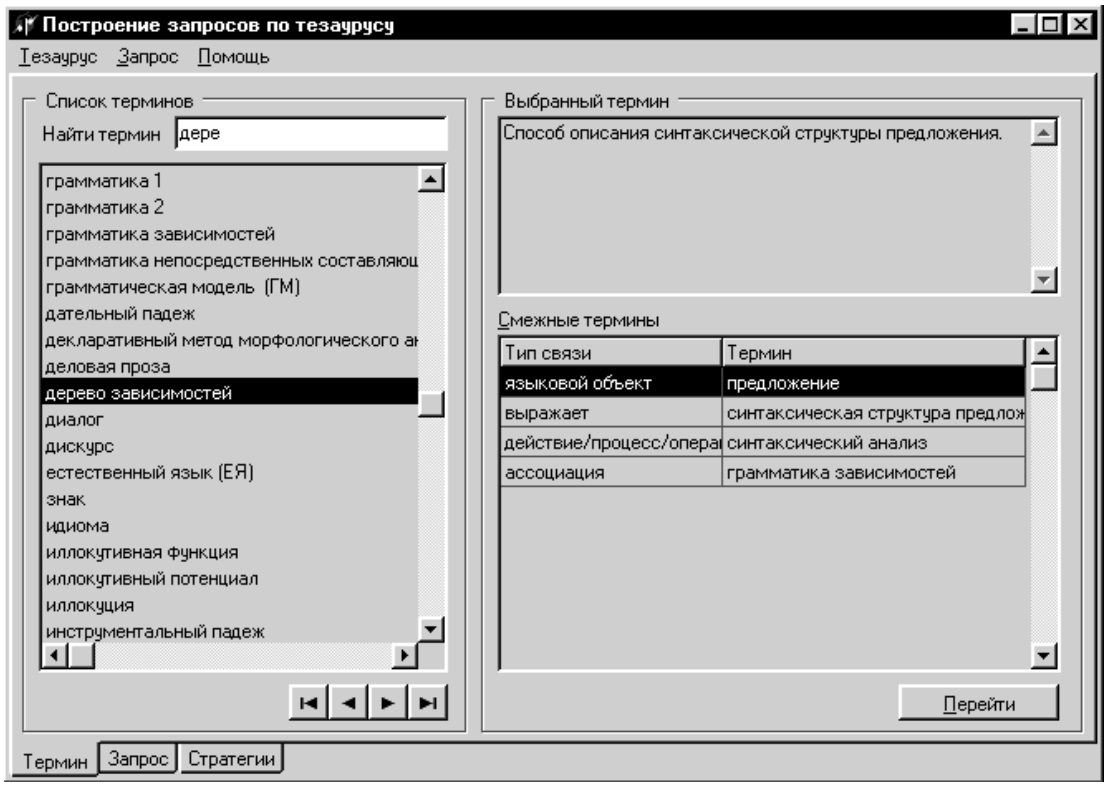

Рис. 3.3. Окно "Термин" программы *ProThes Q*

Окно “Запрос” (рис. 3.5) предназначено для работы с поисковыми запросами. Запрос формируется на основе опорного термина (выбирается в окне «Термин») и стратегии (окно «Стратегия»). В этом окне пользователь может выбрать и скорректировать стратегию, сформировать, отредактировать и

сохранить запрос. Передача запроса в программу просмотра осуществляется через буфер обмена.

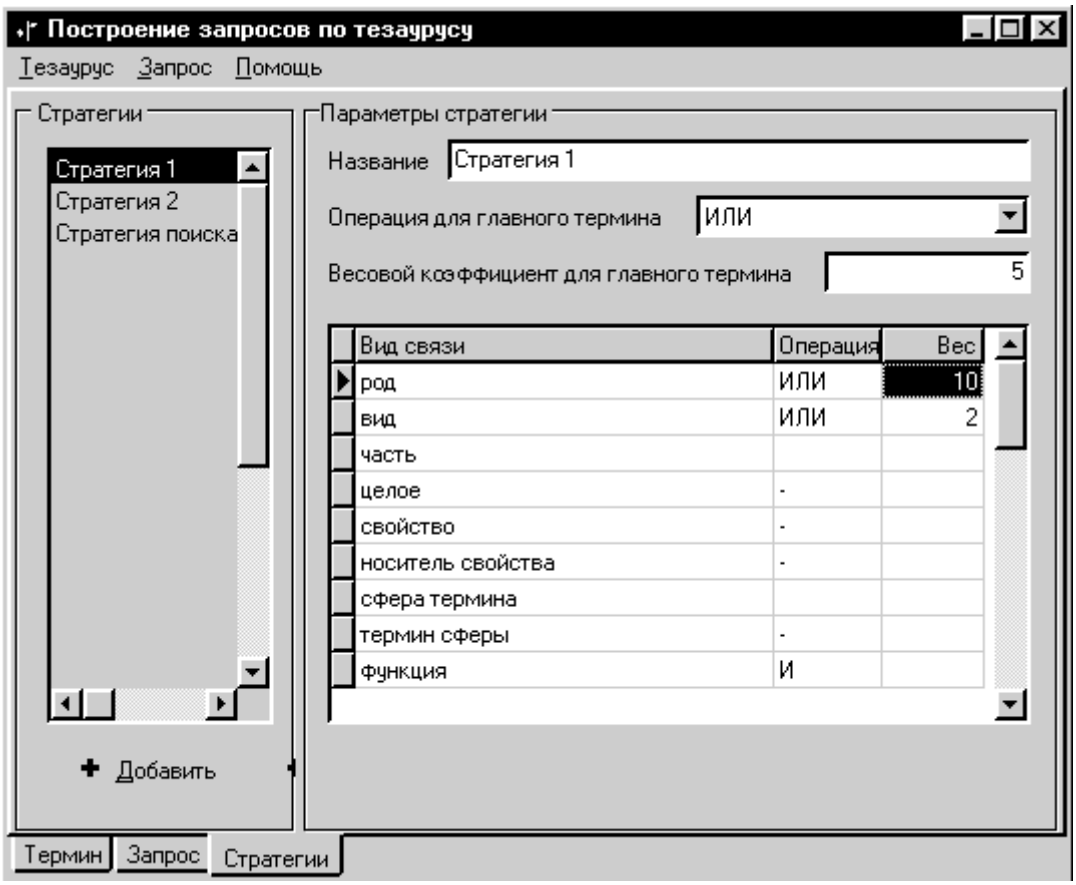

Рис. 3.4. Окно "Стратегия" программы *ProThes Q*

3.4. Результаты и выводы

Итак, мы сформулировали собственный взгляд на условия эффективного расширения запроса к универсальной МП Internet с помощью тезауруса.

Описана архитектура, процедура и модель расширения запроса с помощью тезауруса с сильно дифференцированными семантическими отношениями. Предложенная модель включает в себя в следующие компоненты: тезаурус, стратегия, запрос, процедура формирования запроса по стратегии.

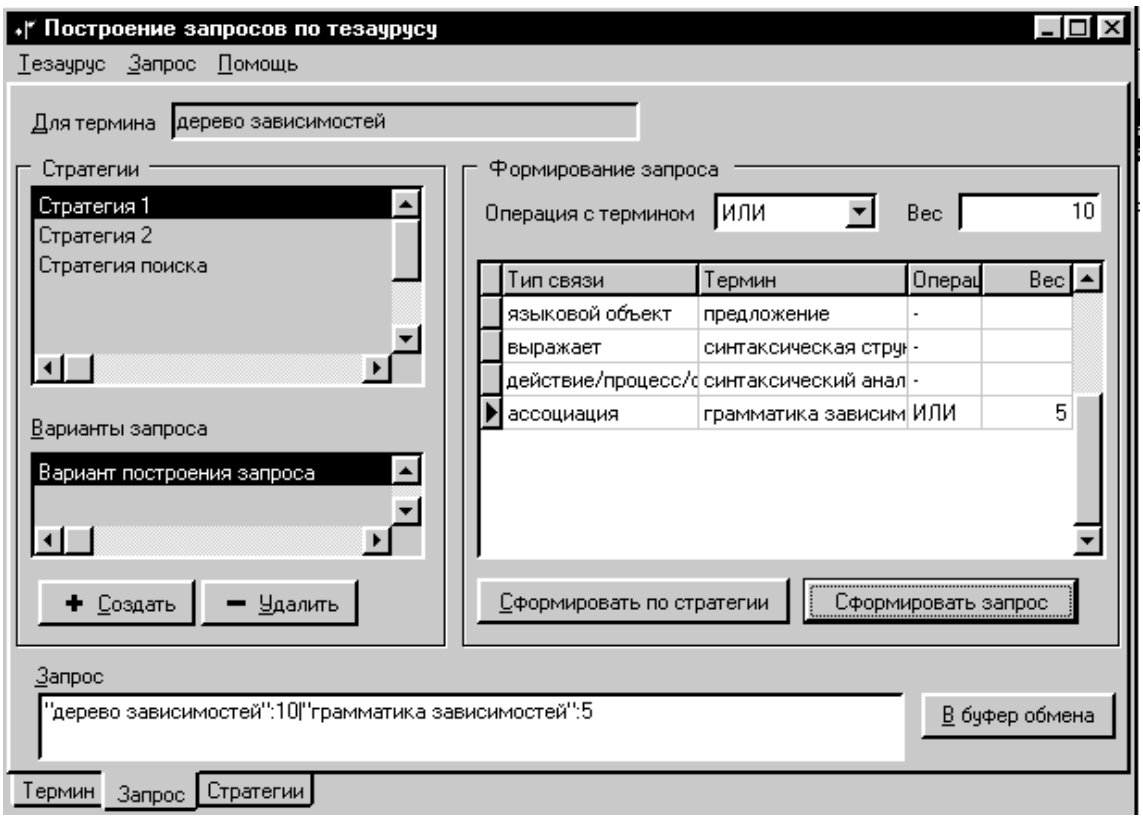

Рис. 3.5. Окно "Запрос" программы *ProThes Q*

Разработан интерфейс программы-ассистента формирования запросов. Разработана программа, которая реализует интерфейс и процедуру на уровне макета.

Направлениями дальнейшего развития формализма тезаурусов применительно к информационному поиску могли бы стать:

- формализация процедуры формирования запроса по стратегии глубиной больше единицы (в запрос включаются не только ближайшие соседи опорного термина);
- формализация процедуры интеграция различных тезаурусов.

4. Проверка результатов. Практические рекомендации

4.1. Стилистическая классификация документов

4.1.1. Классификация тестового массива документов

Для проверки полученных результатов мы сформировали тестовый массив документов. В массив вошел 71 документ, ссылки на которые выдала машина поиска «Яндекс» (www.yandex.ru) в ответ на запрос '*радикал отношение*'[1]. В массив вошли документы в форматах TXT и HTML, которые содержат достаточный объем текста для вычисления значимых параметров. Полный список документов с указанием источников приведен в прил. 4.

Основная часть документов тестового набора принадлежит научному и публицистическому стилям, в соответствии со значениями слова запроса '*радикал*'. Можно предположить, что в целом тестовый массив лучше отражает стилистическую гамму текстов Internet, чем опытный массив.

Проблема оценки качества классификации (составления классификационной матрицы) состоит в том, что некоторые документы тестового массива нельзя однозначно отнести к одному из пяти функциональных стилей.

Например, формально опус С.М.Печкина «Начальный курс квенья[2]» (№ 24 в табл. 4.1) можно было бы отнести к научно-учебному стилю, однако знакомство со *стилем* предисловия удерживает нас от такого решения:

> …Не стану обольщать вас по поводу достоверности всего, изложенного тут. Это, конечно, никакой не учебник… Так что для поклонников научной истины и реализма назовем мы это все опытом конструирования языка по некоторым имеющимся у нас данным. С тем и будем это кушать и из этого будем исходить.
>
> … Работа над учебником начата мною где-то зимой 1993 года, закончена же весной 1995. К работе меня сподвигла не только сама интересность за-

[1] Впоследствии выяснилось, что в некоторых документах, выданных в ответ на запрос, не встречается ни одно из слов запроса (ни в основном тексте, ни в полях разметки). Однако это обстоятельство не влияло на включение документа в тестовый массив.

[2] Квенья – «древнеэльфийский» язык, придуманный Дж.Р.Р. Толкиеном.

дачи, но и, на мой взгляд, существующая потребность в ней у широких толкиенистических и толкиенизирующих(ся) масс, а также необходимость единения, о которой я многократно выступал и буду продолжать выступать.

А стиль произведения «Евангелие от Афрания» Кирилла Еськова (№ 28 в табл. 4.1), которое имеет подзаголовок «Священная история как предмет для детективного расследования», определить вообще затруднительно.

Мы ввели следующие обозначения для промежуточных стилей: PUBL A – аналитическая публицистика, PUBL C – публицистика, близкая разговорному стилю, SCIENCE H – текст, относящийся к гуманитарным наукам, SCIENCE P – научно-популярный текст.

Результаты двух опытов классификации с помощью первой и второй дискриминантной функции (см. п. 2.6.2 и 2.6.3 соответственно) приведены в табл. 4.1. Звездочкой (*) отмечены неправильно классифицированные случаи.

В той же таблице приведены значения первого канонического направления (R1) и первой главной компоненты (F1) для документов тестового массива. Параметры нормированы таким образом, что принимают значения из отрезка [0, 1], причем бóльшие значения соответствуют бóльшей стилистической информативности.

Таблица 4.1

Результаты классификации тестового массива

	Документ	Оценка	R1	F1	Опыт 1	Опыт 2
1.	Доктрина. Архив	CHAT	0,28	0,33	*LITER	*LITER
2.	Сан-Антонио «Причесывая жирафу»	LITER	0,30	0,42	LITER	LITER
3.	Люди прошлого и социализм	PUBL	0,47	0,58	PUBL	PUBL
4.	Черный Л., Павлов Н. МАНИФЕСТ – 93	PUBL	0,58	0,64	PUBL	PUBL
5.	Троцкий Л. Решающий этап	PUBL	0,40	0,54	PUBL	PUBL
6.	Лукьянов А. На пороге третьего тысячелетия	PUBL	0,52	0,61	PUBL	PUBL
7.	Павлов А., Александров Б. Деньги, пахнущие подлостью	PUBL	0,47	0,56	PUBL	PUBL

Продолжение табл. 4.1

Документ	Оценка	R1	F1	Опыт 1	Опыт 2

8. Журнал "Спортивная Жизнь России", раздел «Наши консультации»	PUBL	0,40	0,55	PUBL	PUBL
9. Капитонов К. Ближний восток: в ожидании омоложения и перемен	PUBL	0,46	0,57	PUBL	PUBL
10. Вишневский Б. "Железный дровосек", не получивший сердца...	PUBL	0,41	0,57	PUBL	PUBL
11. Московская прокуратура признала законным решение московского правительства о запрете проведения съезда РНЕ	PUBL	0,72	0,74	*LAW	*LAW
12. Калашников В. Кто там шагает правой?	PUBL	0,52	0,62	PUBL	PUBL
13. Вестник «Сербский крест», – 1999. – №12 (сентябрь).	PUBL	0,41	0,51	PUBL	PUBL
14. Левашов В. Двадцать лет спустя	PUBL	0,57	0,66	PUBL	PUBL
15. Верховский А. Итоги парламентских выборов 1999 года для радикальных национал-патриотов	PUBL	0,48	0,60	PUBL	PUBL
16. ЕС показывает зубы	PUBL	0,47	0,58	PUBL	PUBL
17. Шерман С. Совет Международных Отношений (CFR)	PUBL	0,58	0,67	PUBL	PUBL
18. Кузнецов А. Предвыборный марафон под знаком "зимней войны"	PUBL	0,63	0,71	PUBL	*LAW
19. Е. Щекатихин "Безобидный" сионизм	PUBL	0,40	0,49	PUBL	PUBL
20. Медведев А. Не умножать зла	PUBL	0,52	0,62	PUBL	PUBL
21. Казеннов С., Кумачев В. Либералы: “большие маневры” в преддверии больших перемен	PUBL	0,48	0,60	PUBL	PUBL
22. Мелле Шт. Правые радикалы в немецкой армии	PUBL	0,54	0,65	PUBL	PUBL
23. Антимилитарист Он-Лайн (Информационный бюллетень) – 1999. – №18 (июнь).	PUBL	0,66	0,70	PUBL	PUBL
24. Печкин С.М. Начальный курс квенья	PUBL	0,37	0,52	PUBL	PUBL
25. Любой конфликт заканчивается переговорами. Беседа Максима Шевченко с председателем Совета муфтиев России шейхом Равилем Гайнутдином	PUBL	0,40	0,53	PUBL	PUBL
26. Норта Д. Марксизм и профсоюзы	PUBL	0,58	0,68	PUBL	PUBL

Продолжение табл. 4.1

Документ	Оценка	R1	F1	Опыт 1	Опыт 2
27. Капитонов К. Ближний Восток: в ожидании омоложения и перемен	PUBL	0,46	0,57	PUBL	PUBL
28. Еськов К. Евангелие от Афрания	PUBL (?)	0,39	0,53	PUBL	PUBL
29. Коган-Ясный В.В. Политический аспект отношений федеральных органов власти Российской Федерации с Чеченской республикой в 1990–1994 гг.	PUBL A	0,53	0,65	PUBL	PUBL
30. Березовский В., Червяков В. Хасбулатов и Руцкой: портрет политического альянса	PUBL A	0,68	0,74	*LAW	*LAW
31. Ипполитов К.Х. Время выбора решений	PUBL A	0,61	0,69	PUBL	PUBL
32. Сиван Э. Радикальный ислам: причины и последствия террористического насилия	PUBL A	0,58	0,67	PUBL	PUBL
33. Малашенко А. Исламский фактор в Прикаспийском регионе	PUBL A	0,59	0,68	PUBL	PUBL
34. Группы и фракции – Аграрная партия	PUBL A	0,61	0,68	PUBL	PUBL
35. Нодия Г. Спорные Границы на Кавказе	PUBL A	0,47	0,62	PUBL	PUBL
36. Малашенко А. Исламский фактор в прикаспийском регионе	PUBL A	0,60	0,68	PUBL	PUBL
37. Прейгер Д., Телушкин Дж. Глава из книги "Почему евреи?"	PUBL A	0,43	0,58	PUBL	PUBL
38. России после кризиса:новая расстановка сил, новые сценарии политической борьбы	PUBL A	0,60	0,69	PUBL	PUBL
39. Шляпентох В. От любви к ненависти	PUBL A	0,52	0,65	PUBL	PUBL
40. Акишев А. Политическая элита Казахстана: история, современность, перспективы	PUBL A	0,53	0,63	PUBL	PUBL
41. Второе полугодие 1994 г. (господствующие тенденции развития политической ситуации)	PUBL A	0,58	0,66	PUBL	PUBL
42. Мухин Ю.И. Товарищ брянский волк	PUBL C	0,19	0,38	*LITER	*LITER
43. Фадеев Ю. Россия требует к себе важительного подхода, как к евроазиатской исламско-христианской стране (ответ на вопросы "Русского Журнала").	PUBL C	0,29	0,46	*LITER	*LITER
44. Мельников А. Егор, Вы не совсем правы	PUBL C	0,35	0,43	PUBL	*LITER
45. Лейбов Р. «Бессрочная ссылка» от	PUBL C	0,32	0,46	PUBL	*LITER

20.04.98.					

Продолжение табл. 4.1

Документ	Оценка	R1	F1	Опыт 1	Опыт 2
46. Абакумов А. В. ХОЧУ ВОЗРАЗИТЬ!? Сергею Васильцову - историку, Сергею Обухову – экономисту	PUBL C	0,51	0,61	PUBL	PUBL
47. Экспериментальные методы изучения радикальных реакций	SCIENCE	0,61	0,69	SCIENCE	SCIENCE
48. Балаболкин М.И., Клебанова Е.М., Креминская В.М. Патогенез ангиопатий при сахарном диабете	SCIENCE	0,64	0,71	SCIENCE	SCIENCE
49. АЗАКТАМ (описание лекарственного препарата)	SCIENCE	0,73	0,75	SCIENCE	SCIENCE
50. Тарабанько В. Е. Исследование кислотности радикальных интермедиатов окисления лигнина квантово-химическими методами	SCIENCE	0,75	0,73	SCIENCE	SCIENCE
51. Переходцев Г.Д., Дубинский А.А., Лебедев Я.С. Ферромагнитный обмен и аномальные знаки D в радикальных парах	SCIENCE	0,64	0,68	SCIENCE	SCIENCE
52. Борщевская М.И., Васильева С.М. Развитие представлений о биохимии и фармакологии меланиновых пигментов	SCIENCE	0,66	0,72	SCIENCE	SCIENCE
53. Фролков Г.Д. Влияние природной механоактивации на выбросоопасность углей	SCIENCE	0,76	0,79	SCIENCE	*LAW
54. Воздействие радиации на ткани живого организма	SCIENCE	0,80	0,75	SCIENCE	SCIENCE
55. Романченко А. Удивительный гормон – мелатонин	SCIENCE	0,65	0,73	SCIENCE	SCIENCE
56. Гудзь О.В. Итоги и перспективы клинического применения дезинфекционных средств из группы четвертичных аммониевых соединений	SCIENCE	0,76	0,80	SCIENCE	SCIENCE
57. Письменский А. В. Влияние условий окисления и структуры ароматических диаминов на механизм и эффективность их ингибирующего действия»	SCIENCE	0,86	0,84	SCIENCE	SCIENCE
58. Автономный электронный стимулятор желудочно-кишечного тракта и слизистых оболочек "Электронный нормализатор"	SCIENCE	0,70	0,74	SCIENCE	SCIENCE

Продолжение табл. 4.1

Документ	Оценка	R1	F1	Опыт 1	Опыт 2
59. Строение гидроксильной группы	SCIENCE	0,69	0,71	SCIENCE	SCIENCE
60. Левицкий Е.Л. и др. Коррекция поражений ядерного генома антиоксидантами в условиях токсического повреждения печени	SCIENCE	0,75	0,75	SCIENCE	SCIENCE
61. Соловьев М. В. Виртуальное антистарение	SCIENCE	0,65	0,71	SCIENCE	SCIENCE
62. Вартанян Л.С., Гуревич С.М. Влияние ионола на метаболизм супероксидных радикалов в печени мышей	SCIENCE	0,81	0,79	SCIENCE	SCIENCE
63. Словецкий Д.И. Механизмы плазмохимического травления материалов	SCIENCE	0,73	0,76	SCIENCE	*LAW
64. Ким Сун-чжон, Пиков Г.Г. Жан Кальвин и некоторые проблемы швейцарской Реформации (Материалы к спецкурсу)	SCIENCE H	0,57	0,64	*PUBL	*PUBL
65. Чудинов А.В. Размышления англичан о Французской революции: Э.Берк, Дж.Макинтош, У.Годвин	SCIENCE H	0,64	0,69	SCIENCE	*PUBL
66. Русское население республик Северного Кавказа	SCIENCE H	0,66	0,72	*PUBL	*LAW
67. Никольский С.А. "Аграрная реформа" 1991 – 1995 годов и проблема модернизации российской деревни	SCIENCE H	0,62	0,67	*PUBL	*PUBL
68. Калашников З.М. Политические и правовые идеи в странах Западной Европы и Америки в новое и новейшее время	SCIENCE H	0,64	0,69	*PUBL	*PUBL
69. Вуколов С. Кислоты // Брокгауз Ф.А., Ефрон И.А. Энциклопедический словарь	SCIENCE P	0,46	0,57	*PUBL	*PUBL
70. Мелатонин	SCIENCE P	0,68	0,73	SCIENCE	SCIENCE
71. “MIRACLE 7”- лучший антиоксидантный состав	SCIENCE P	0,62	0,68	SCIENCE	SCIENCE

Результаты двух опытов классификации с помощью первой/второй дискриминантных функций сведены в табл. 4.2:

Таблица 4.2

Классификационная матрица для тестового массива

	CHAT	LITER	PUBL	SCIENCE	LAW	Классифицировано правильно
CHAT	0/0	1/1	0/0	0/0	0/0	0/0
LITER	0/0	1/1	0/0	0/0	0/0	1/1
PUBL	0/0	2/4	40/37	0/0	2/3	0,91/0,84
SCIENCE	0/0	0/0	5/5	20/17	0/3	0,80/0,68
Всего	0/0	4/6	45/42	20/17	2/6	0,86/0,77

Заметим, что ошибки классификации закономерны на «промежуточных» текстах – т.е. тех текстах, которые сложно однозначно отнести к одному из стилей. Поэтому не все ошибки классификации имеют «одинаковый статус».

При формировании части опытного массива, репрезентирующей научный стиль речи, мы ориентировались на естественнонаучные тексты. В этих текстах характерные особенности стиля выражены наиболее рельефно. Стиль многих гуманитарных научных работ (особенно описательного характера), весьма близок среднему по эмоциональности газетно-журнальному стилю.

Например, документ «Жан Кальвин и некоторые проблемы швейцарской Реформации. (Материалы к спецкурсу)» (№ 64 в табл. 4.1) был в обоих опытах отнесен к публицистическому стилю:

> О кальвинизме же массовый читатель мог получить самое общее и далеко не всегда верное представление, которое сводилось, по сути, лишь к кальвинисткой теории предопределения. Более подробную, но зато и намного менее объективную информацию можно было получить из пособий для разного рода духовных заведений, с помощью которых студенты семинарий и академий знакомились с вероучением и историей неправославных христианских вероисповеданий, которые в России было (и остается до сих пор!) принято называть "западными", хотя их приверженцев можно во множестве встретить в любом регионе земного шара.

Статья «Кислоты» из энциклопедии Брокгауза и Ефрона начала века (№ 69 в табл. 4.1) – пример научно-популярного стиля, также была отнесена к классу PUBL:

> Сделать характеристику их (кислот. – *П.Б.*) в немногих словах является крайне затруднительным и только обращаясь к историческому изложению того, как слагались взгляды на К., можно дать некоторое понятие об них. В самую раннюю эпоху развития химических знаний под К. подразумевались вещества, кислые на вкус. Типичным представителем их мог считаться уксус (acetum), название которого, вероятно, и послужило впоследствии для наименования этой группы (acidum).

Отнесение этих документов к публицистическому стилю в зависимости от исходных предпосылок и задач классификации может рассматриваться и как ошибка, и как правильное решение.

Заметим, что ни один из документов не был ошибочно отнесен к научному стилю. Т.е. можно предположить, что для научных текстов *точность поиска* будет существенно повышена.

Нормированные параметры F1 и R1 могут служить количественными показателями шкалы «стилистической информативности». На наш взгляд, параметр R1 лучше отражает информативность текстов тестового набора. Заметим, что для публицистики, склонной к аналитичности (PUBL A) или просто более ровной эмоционально, значения этого параметра ниже среднего. Обратная ситуация – для публицистики, тяготеющей к разговорному стилю (PUBL C).

4.1.2. Возможные реализации

Мы предлагаем два варианта практической реализации метода стилистической классификации для повышения эффективности поиска информации в Internet.

Целенаправленный поиск научных документов. Опросы пользователей показывают, что научная и образовательная информация принадлежит к наиболее популярным и востребованным ресурсам Internet. С другой сто-

роны, недостаток этой информации пользователи ощущают особенно остро (см. п. 1.1.3). Можно предположить, что с точки зрения поиска информации в Internet наибольшей ценностью в системе функциональных стилей обладает научный стиль речи.

Технически этой может быть просто дополнительный интерфейс к универсальной машине поиска[1], ориентированный на специфическую группу пользователей – «искателей научной информации». Полученная нами дискриминантная функция может быть несколько модифицирована с учетом особенностей текстов, относящихся к гуманитарным наукам.

Стилистическое ранжирование документов. Введенная нами шкала «стилистическая информативность» позволяет линейно упорядочивать найденные документы (по возрастанию или убыванию). Можно также предоставить пользователю возможность задавать интервал параметра, например, (0;0,5) или (0,5;0,8). Тогда будут выданы только те документы, для которых значение параметра попадает в заданный интервал.

4.2. Расширение запроса с помощью тезауруса

4.2.1. Запросы к ИПС «Яндекс»

С помощью тезауруса по компьютерной лингвистике (прил. 5) мы провели ряд экспериментов по формированию запросов к ИПС «Яндекс» (табл. 4.3).

[1] Уже сегодня у поисковой машины Яндекс есть два интерфейса: «яркий» (www.yandex.ru) и «спартанский» (www.ya.ru).

Таблица 4.3

Отклики ИПС «Яндекс» на запросы, сформированные при помощи программы *ProThes Q* и тезауруса по компьютерной лингвистике

№	Запрос	Кол-во ссылок
1.	*«формальная грамматика»*	274
2.	*(формальная|контекстно-зависимая|контекстно-свободная| сетевая| трансформационная|зависимостей| «непосредственных составляющих») /1 грамматика*[*)]	549
3.	*«лингвистическая трансляция»*	3
4.	*«расширенная сеть переходов»*	4
5.	*иллокуция перлокуция*	4
6.	*слово:10 предложение морфема словосочетание словоформа лексема аффикс*	68
7.	*анафора&&антецедент*	3
8.	*фрейм:10 «искусственный интеллект»*	15

[*)] Запрос был отредактирован вручную для исключения нескольких вхождений слова *грамматика*.

Первые два запроса (табл. 4.3) демонстрируют возможность управления полнотой поиска. Если первый запрос состоит из одного термина – *формальная грамматика*, то второй объединяет по «ИЛИ» семь близких терминов. Анализ начала списка ссылок в обоих случаях позволяет сделать заключение о приемлемой точности поиска (см., рис. 4.1, 4.2).

Запрос: «формальная грамматика»

Результат поиска: страниц - 274, серверов – 61 (формальная: 180217, грамматика: 94224)

1. Problems for "informatics" | Показать найденные слова
   http://www.isu.ru/~slava/teach/art/problems.htm - 1К - 20.11.1998 - совпадение фразы
   Похожие документы | Все с сервера не менее 2 док.

2. http://www.etu.ru/misc/edu/section1/part121.htm | Показать найденные слова
   Пред.Страница След.Страница Раздел Содержание 1.2. Определение формальной грамматики и языка Изучение предмета начнем с определения первичных понятий.
   http://www.etu.ru/misc/edu/section1/part121.htm - 9К - совпадение фразы
   Похожие документы | Все с сервера не менее 12 док.

3. Формальные грамматики дают естественный способ упорядочения языков с помощью огр .. | Показать найденные слова
   Классификация Хомского Формальные грамматики дают естественный способ упорядочения языков с помощью ограничений, которые можно ввести в продукции (правила переписывания). Перечислим их в порядке ...
   http://utc.uni-dubna.ru/~mazny/kurses/loi/10_.htm - 25К - 14.01.1999 - совпадение фразы
   Похожие документы | Все с сервера не менее 7 док.

Рис. 4.1. Отклик «Яндекс» на запрос *«формальная грамматика»*

4. ФПМ 1995, ТОМ 1, ВЫПУСК 3, СТР. 729-751 | Показать найденные слова
In Russian, KOI-8. M.R.Pentus. Lambek calculus and formal grammars. Abstract (HTML), article (PostScript). Journal "Fundamental'naya i prikladnaya matematika" ("Fundamental and Applied Mathematics")
http://math.msu.su/~fpm/rus/95/953/95311.htm - 2K - 14.12.1998 - совпадение фразы
Похожие документы | Все с сервера не менее 5 док.

5. Mono_2_2 | Показать найденные слова
2.2. Некоторые традиционные конструктивные формальные системы 2.2.1. Формальные грамматики (ТКФС ФГ) Формальные грамматики имеют наиболее простые механизмы, поэтому будут рассмотрены среди ...
http://www.monograph.epi-system.unir.ru/Mono_2_2.htm - 15K - 23.12.1998 - совпадение фразы
Похожие документы | Все с сервера не менее 10 док.

6. Язык и грамматика | Показать найденные слова
http://www.citforum.ru, C++ Бархатный путь, Марченко А.Л. Избранные главы
http://icc.jamal.ru/iso-8859.5/citforum/programming/cpp_march/cpp_003.shtml - 20K - совпадение фразы
Похожие документы | Все с сервера не менее 2 док.

7. Lectures on Computer Science 1999-2000 | Показать найденные слова
Программирование Программа лекций первого потока второго курса 1999-2000 учебный год Лектор В. В. Борисенко Основы языка Java. Типы, операции. Классы, члены классов, методы, константы. Статические ...
http://mars.math.msu.su/2course/Borisenko/prog2000.html - 4K - 26.04.2000 - совпадение фразы
Похожие документы | Все с сервера не менее 5 док.

8. Parsing Glossary | Показать найденные слова
parsing, context-free grammars, terminal and nonterminal tokens, lexical scanner, etc.
http://infoscope.nm.ru/languages/parsing/glossary/index.html - 102K - 03.07.2000 - совпадение фразы
Похожие документы | Все с сервера не менее 4 док.

9. http://www.rsuh.ru/win/whoiswho/4/gladki.html | Показать найденные слова
ГЛАДКИЙ Алексей Всеволодович - доктор физико-математических наук (1966), профессор (1968) кафедры логики и интеллектуальных систем факультета теоретической и прикладной лингвистики (ФТиПЛ). Специалист в
http://www.rsuh.ru/win/whoiswho/4/gladki.html - 1K - 10.10.1998 - совпадение фразы
Похожие документы | Все с сервера не менее 6 док.

10. Язык и грамматика | Показать найденные слова
http://www.citforum.ru, C++ Бархатный путь, Марченко А.Л. Избранные главы
http://techlib.mipt.ru/books/citforum/programming/cpp_march/cpp_003.shtml - 18K - совпадение фразы
Похожие документы | Все с сервера не менее 2 док.

Страницы: **1** | 2 | 3 | 4 | 5 | 6 | 7 | Еще >>

Окончание рис. 4.1.

Запрос: (формальная|контекстно-зависимая|контекстно-свободная| сетевая| трансформационная| зависимостей | "непосредственных составляющих") /1 грамматика

Результат поиска: страниц - 549, серверов – 76 (формальная: 180217, контекстно: 4803, зависимая: 87532, свободная: 1356245, сетевая: 1735795, трансформационная: 2599, зависимостей: 700801, непосредственных: 701990, составляющих: 2919000, грамматика: 94224)

1. Контекстно-свободные грамматики | Показать найденные слова
   Контекстно-свободные грамматики Общий алгоритм разбора Метод рекурсивного спуска Алгоритм разбора для LL(1)-грамматик
   http://bspu.secna.ru/Department/WMiP/shen/node74.html - 1K - 10.01.1999 - строгое соответствие
   Похожие документы | Все с сервера не менее 2 док.

2. ABC Referats. Show File Info. | Показать найденные слова
   Рефераты | Рецепты | Цветочки | Книга ЖиП | Знакомства | Домой | E-Mail Полная информация о файле: 10/240-0732.zip Поиск на сервере рефераты книги видео
   http://abc.osu.ru/cgi-bin/refs.cgi?462=10/240-0732.zip - 8K - строгое соответствие
   Похожие документы | Все с сервера не менее 2 док.

3. YACC(1) | Показать найденные слова
   Сервер содержит море(!) аналитической информации CIT Forum CD-ROM
   http://doc.trecom.tomsk.su/citforum.ru/win/operating_systems/manpages/YACC.1.sht .. - 10K - строгое соответствие
   Похожие документы | Все с сервера не менее 1 док.

4. http://vt.miem.edu.ru/main/documentation/algorithm/book/g13w.shtml | Показать найденные слова
   Глава 13. Контекстно-свободные грамматики 13.1. Общий алгоритм разбора Чтобы определить то, что называют контекстно-свободной грамматикой (КС-грамматикой), надо: (а) указать конечное множество A, ...
   http://vt.miem.edu.ru/main/documentation/algorithm/book/g13w.shtml - 41K - строгое соответствие
   Похожие документы | Все с сервера не менее 1 док.

5. Формальные грамматики дают естественный способ упорядочения языков с помощью огр .. | Показать найденные слова
   Классификация Хомского Формальные грамматики дают естественный способ упорядочения языков с помощью ограничений, которые можно ввести в продукции (правила переписывания). Перечислим их в порядке ...
   http://utc.uni-dubna.ru/~mazny/kurses/loi/10_.htm - 25K - 14.01.1999 - строгое соответствие
   Похожие документы | Все с сервера не менее 5 док.

6. Спецификация языка ЯВА. Грамматики. | Показать найденные слова
   Содержание | Предыдущая | Следующая ГЛАВА 2 Грамматики В этой главе описываются контекстно-свободные грамматики, используемые в данной спецификации для определения лексической и синтаксической ...
   http://www.uni-vologda.ac.ru/java/jls/2-doc.htm - 11K - 15.11.1999 - строгое соответствие
   Похожие документы | Все с сервера не менее 1 док.

7. http://www.etu.ru/misc/edu/section1/part121.htm | Показать найденные слова
   Пред.Страница След.Страница Раздел Содержание 1.2. Определение формальной грамматики и языка Изучение предмета начнем с определения первичных понятий.
   http://www.etu.ru/misc/edu/section1/part121.htm - 9K - строгое соответствие
   Похожие документы | Все с сервера не менее 12 док.

Рис. 4.2. Отклик «Яндекс» на расширенный запрос *«формальная грамматика»*

8. История языкознания. Глава 10 | Показать найденные слова
И.П. Сусов. История языкознания Оглавление учебника Библиография Глава 9 Глава 10 10.1. Генеративное (порождающее) языкознание 10.2. Современные исследования в области функциональной лингвистики
http://homepages.tversu.ru/~susov/Hist_10.htm - 71K - 15.05.2000 - строгое соответствие
Похожие документы | Все с сервера не менее 19 док.

9. Problems for "informatics" | Показать найденные слова
http://www.isu.ru/~slava/teach/art/problems.htm - 1K - 20.11.1998 - строгое соответствие
Похожие документы | Все с сервера не менее 2 док.

10. Джон Гриндер, Ричард Бэндлер. Структура магии (том 1) | Показать найденные слова
Джон Гриндер, Ричард Бэндлер. Структура магии (том 1) ТОМ 1 ПРЕДИСЛОВИЕ Из глубины веков доходят до нас песни и легенды о чудесной власти магов и кудесников. Обычного человека всегда захватывала ...
http://lib.ru/NLP/magianlp1.txt - 448K - 28.04.2000 - строгое соответствие
Похожие документы | Все с сервера не менее 3 док.

Страницы: **1** | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Еще >>

Окончание рис. 4.2

Высокую точность поиска доставляют запросы *«лингвистическая трансляция», «расширенная сеть переходов»* и *‘иллокуция перлокуция’*, сформированные с помощью тезауруса по компьютерной лингвистике (№№ 3 – 5, табл. 4.3). В ответ на первый запрос «Яндекс» выдал три ссылки на один и тот же документ (рис. 4.3), в ответ на второй и третий – по четыре ссылки (рис. 4.4, 4.5). Все ссылки являются релевантными.

Запрос: «лингвистическая трансляция»

Результат поиска: страниц - 3, серверов - 3 (лингвистическая: 46281, трансляция: 103766)

1. Magazine Mathematical Machines and System | Показать найденные слова
Математичні машини і системи. 1997 N2 РЕФЕРАТИ ОБЧИСЛЮВАЛЬНІ СИСТЕМИ УДК 681.3 Ситуацъйнъ центри - основа управлъння органъзацъйними системами великоъ розмърностъ /
http://immsp.kiev.ua/immsp/win/publ/MMS97N2.html - 26K - совпадение фразы
Похожие документы | Все с сервера не менее 1 док.

2. Magazine Mathematical Machines and System | Показать найденные слова
Математичні машини і системи. 1997 N2 РЕФЕРАТИ ОБЧИСЛЮВАЛЬНІ СИСТЕМИ УДК 681.3 Ситуаційні центри - основа управління організаційними системами велико розмірності /
http://opinion.kiev.ua/immsp/koi/publ/MMS97N2.html - 26K - совпадение фразы
Похожие документы | Все с сервера не менее 1 док.

3. Magazine Mathematical Machines and System | Показать найденные слова
Математичні машини і системи. 1997 N2 РЕФЕРАТИ ОБЧИСЛЮВАЛЬНІ СИСТЕМИ УДК 681.3 Ситуацъйнъ центри - основа управлъння органъзацъйними системами великоъ розмърностъ /
http://www.immsp.kiev.ua/immsp/win/publ/MMS97N2.html - 26K - совпадение фразы
Похожие документы | Все с сервера не менее 1 док.

Рис. 4.3. Отклик «Яндекс» на запрос *"лингвистическая трансляция"*

Запрос: «расширенная сеть переходов»

Результат поиска: страниц - 4, серверов – 4 (расширенная: 543660, сеть: 1786557, переходов: 373246)

1. Меня зовут Моцарт -- Полный ПК - самый народный журнал! | Показать найденные слова
   Меня зовут Моцарт
   http://www.osp.ru/fullpc/06/05.htm - 14К - строгое соответствие
   Похожие документы | Все с сервера не менее 1 док.

2. Parsing Glossary | Показать найденные слова
   parsing, context-free grammars, terminal and nonterminal tokens, lexical scanner, etc.
   http://infoscope.nm.ru/languages/parsing/glossary/index.html - 102К - 03.07.2000 - строгое соответствие
   Похожие документы | Все с сервера не менее 1 док.

3. Р. Бертон. Семантическая грамматика | Показать найденные слова
   Р. Бертон Семантическая грамматика СОДЕРЖАНИЕ 1 Введение а) отличительные особенности семантических грамматик б) стандартные грамматики и семантические грамматики 2 Преимущества семантических ...
   http://www.arnika.sci-nnov.ru/kn/info/progr/intellect/semant_grammar.htm - 9К - 30.04.2000 - строгое соответствие
   Похожие документы | Все с сервера не менее 1 док.

4. ПРОМТ: история машинного перевода | Показать найденные слова
   Соколова Светлана кандидат технических наук президент компании ПРОМТ Как переводит компьютер Машинный перевод - это такая специфическая область применения компьютеров, в проблемах которой почти ...
   http://www.promt.ru/rus1/mtw/history.htm - 18К - 08.08.1999 - строгое соответствие
   Похожие документы | Все с сервера не менее 1 док.

Рис. 4.4. Отклик "Яндекс" на запрос "*расширенная сеть переходов*"

Пример уточнения семантики термина за счет «соседей» демонстрирует следующий запрос (№ 6, табл. 4.3). В нем многозначное и общеупотребительное *слово* были уточнено за счет терминов *предложение* (также является многозначным словом), *морфема*, *словосочетание*, *словоформа*, *лексема*, *аффикс*. В выдаче содержатся ссылки на 68 документов, все документы начала списка являются релевантными (рис. 4.6).

Однако даже при достаточно специфических запросах, которые формируются из терминов тезауруса, выданные документы могут отличаться по стилю.

Так, на запрос *‘фрейм:10 "искусственный интеллект"’* было выдано 15 ссылок (рис. 4.7). Дайджест книги Дугласа Хофштадтера «Гедель, Эшер, Бах: вечная золотая цепь» (№ 1), равно как и опус Михаила Сухарева «Движение порядка в природе» (№№ 5, 6) формально являются релевантным откликом,

так как в нем действительно идет речь об искусственном интеллекте и фреймах.

Запрос: иллокуция перлокуция

Результат поиска: страниц - 5, серверов - 4 (иллокуция: 28, перлокуция: 8)

1. Основные понятия лингвистики | Показать найденные слова
   Основные понятия лингвистики И.В.Утехин 1. Языковедческие дисциплины: их соотношение и задачи. Понятие о языке. Функции языка. Язык, речь, речевая деятельность. Компоненты и уровни языка и речевой ...
   http://www.eu.spb.ru/ethno/ethno104.htm - 5К - 09.06.2000 - совпадение фразы
   Похожие документы | Все с сервера не менее 1 док.

2. Семиотика | Показать найденные слова
   Уголок студента, специализирующегося по направлению "ЛИНГВИСТИКА И НОВЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ" 14. Семиотика Извлечения из Госстандарта - 2000 Рекомендации по содержанию общепрофессиональной ...
   http://homepages.tversu.ru/~susov/Applied14.htm - 15К - 01.08.2000 - строгое соответствие
   Похожие документы | Все с сервера не менее 2 док.

3. УРАО - это... | Показать найденные слова
   Е. В. Клюев. Программа курса "Лингвистическая прагматика" (с контрольными вопросами) Программа предназначена прежде всего для студентов факультета журналистики, но может использоваться также ...
   http://www.urao.edu/facultet/jour/subjects/linguist.html - 31К - 17.03.2000 - строгое соответствие
   Похожие документы | Все с сервера не менее 1 док.

4. Иван Павлович Сусов. История языкознания | Показать найденные слова
   Иван Павлович Сусов. История языкознания Email: susov@tversu.ru WWW: http://homepages.tversu.ru/~susov/ Иван Павлович Сусов, доктор филологических наук, профессор, Заслуженный деятель науки РФ ...
   http://icc.jamal.ru/library/koi/TEXTBOOKS/yazykoznanie.txt - 613К - строгое соответствие
   Похожие документы | Все с сервера не менее 1 док.

Рис. 4.5. Отклик "Яндекс" на запрос '*иллокуция перлокуция*'

Запрос: слово:10 предложение морфема словосочетание словоформа лексема аффикс
Результат поиска: страниц - 68, серверов - 35 (слово: 4799984, предложение: 2872037, морфема: 1853, словосочетание: 34381, словоформа: 12222, лексема: 7518, аффикс: 948)

1. Описание языка в формальном аспекте | Показать найденные слова
   Профессор И.П. Сусов. Введение в теоретическое языкознание Модуль 1. Язык как объект языкознания Описание языка в формальном аспекте Оглавление учебника Содержание модуля 1
   http://homepages.tversu.ru/~susov/1_07.htm - 22К - 15.08.2000 - нестрогое соответствие
   Похожие документы | Все с сервера не менее 14 док.

2. Вопросы 16-20 | Показать найденные слова
   1-5 6-10 11-15 16-20 21-25 26-30 31-35 36-41 16. Фонологическая система. Понятие корреляции. 17. Понятие морфемы. Варианты морфем. Виды морфем (корни, аффиксы).
   http://csd.newmail.ru/brothers/linguistics/16.htm - 10К - 20.02.2000 - нестрогое соответствие
   Похожие документы | Все с сервера не менее 4 док.

Рис. 4.6. Отклик «Яндекс» на расширенный запрос с опорным термином *слово*

3. Ольга Лаэдэль. Ладар'атэль - язык планеты Атэа. | Показать найденные слова
Описание грамматики ladar'atel - языка цивилизации лемле с планеты Атэа. (Часть утопии о цивилизации лемле)
http://www.chat.ru/~laedel/lingua.htm - 32К - 18.08.2000 - нестрогое соответствие
Похожие документы | Все с сервера не менее 1 док.

4. Ольга Лаэдэль. Ладар'атэль - язык планеты Атэа. | Показать найденные слова
Описание грамматики ladar'atel - языка цивилизации лемле с планеты Атэа. (Часть утопии о цивилизации лемле)
http://laedel.euro.ru/lingua.htm - 32К - 24.08.2000 - нестрогое соответствие
Похожие документы | Все с сервера не менее 1 док.

5. конференция | Показать найденные слова
Д.Р.Копосов (Казань) ПРОБЛЕМА МОРФЕМЫ В ИССЛЕДОВАНИЯХ В.М.МАРКОВА Данное сообщение имеет целью показать те тенденции в формировании русской морфологической системы, которые иллюстрируются наиболее ...
http://www.kcn.ru/tat_ru/science/news/rus_lang/t7.htm - 6К - нестрогое соответствие
Похожие документы | Все с сервера не менее 7 док.

6. Zubkova L.G. | Показать найденные слова
Л.Г.Зубкова ФОНЕТИКА И ГРАММАТИКА: К ОБОСНОВАНИЮ ГРАММАТИЧЕСКОЙ МОТИВИРОВАННОСТИ РУССКОГО ЗВУКОВОГО СТРОЯ По типу морфологической структуры слова русский язык обычно квалифицируется как ...
http://logos.philol.msu.ru/rus/gorn/arso/zubkova.htm - 18К - нестрогое соответствие
Похожие документы | Все с сервера не менее 5 док.

7. http://axis.novsu.ac.ru/uni/vestnik.nsf/All/96413FE25DD5B5E7C3256727002E7BA1 | Показать найденные слова
Назад 1998 ВЕСТНИК НОВГОРОДСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА 9 Жуков А.В. Развитие лингвистических идей в.п.жукова: единство теории, методологии и практики
http://axis.novsu.ac.ru/uni/vestnik.nsf/All/96413FE25DD5B5E7C3256727002E7BA1 - 29К - нестрогое соответствие
Похожие документы | Все с сервера не менее 2 док.

8. Common characteristic | Показать найденные слова
ПОЛИКАРПОВ Анатолий Анатольевич. ЦИКЛИЧЕСКИЕ ПРОЦЕССЫ В СТАНОВЛЕНИИ ЛЕКСИЧЕСКОЙ СИСТЕМЫ ЯЗЫКА: МОДЕЛИРОВАНИЕ И ЭКСПЕРИМЕНТ 1. Объект и основные задачи исследования
http://humlang.newmail.ru/modela/comchar.htm - 30К - 28.04.2000 - нестрогое соответствие
Похожие документы | Все с сервера не менее 2 док.

9. Кафедра русского языка и литературы | Показать найденные слова

http://biro.ufanet.ru/biro/path_5.htm - 44К - 19.05.2000 - нестрогое соответствие
Похожие документы | Все с сервера не менее 1 док.

10. Т.А. Кабанова. К интерпретации простых глаголов в кетском языке | Показать найденные слова
Т.А. Кабанова К интерпретации простых глаголов в кетском языке Т.А.Кабанова, 1998. В своей монографии "Кетский язык" (Томск, 1968) А.П. Дульзон выделил 38 простых глаголов на основе следующих ...
http://arctogaia.krasu.ru/laboratory/p2/kabanova.htm - 3К - 16.05.2000 - нестрогое соответствие
Похожие документы | Все с сервера не менее 1 док.

Страницы: **1** | 2 | 3 | 4 | Еще >>

Окончание рис. 4.6

Две цитаты из этих документов могут дать представление об их содержании и стиле изложения.

Дайджест книги Хофштадтера: «Диалог переключает дискуссию на вопрос о том, как знание представляется на разных уровнях контекста. Это приводит к современным идеям *искусственного интеллекта – фреймам* (курсив наш. – *П.Б.*). Для конкретности представлен организованный фреймо-подобным образом набор головоломок на визуальные образы. Затем обсуждаются глубокие последствия взаимодействия концепций вообще, которое завершается некоторыми рассуждениями о творчестве. Глава заканчивается набором личных "Вопросов и размышлений" по поводу ИИ и разума вообще.»

Сухарев: «Дальнейшее совершенствование средств производства оказывается связано с проблемой создания *искусственного интеллекта* (курсив наш. – *П.Б.*), а та, в свою очередь и неизбежно – с основным вопросом философии в форме отношения материи и сознания. Впервые развитие материального базиса общества оказывается в такой непосредственной связи с философией, и это знаменательный факт.»

Результат обработки запроса *‘анафора && антецедент’* содержит ссылку на документ «Цицероны филфака» – собрание смешных изречений преподавателей филологического факультета МГУ. Слова запроса встречаются в этом документе в следующих фразах: «*сермяжная теория анафоры*» и «*консеквент получили – антецедент обратно присобачили*».

Обработав тексты этих «спорных» документов с помощью программы стилистического анализа, мы получили следующие результаты:

Таблица 4.4

Документ	Результат классификации	R1
Гедель, Эшер, Бах	PUBL	0,57
Движение порядка в природе	PUBL	0,37
Цицероны филфака	LITER	0,23

Результаты говорят о возможности совместного использовании методов, описываемых в данной работе.

Запрос: фрейм:10 "искусственный интеллект"

Результат поиска: страниц - 15, серверов - 14 (фрейм: 107478, искусственный: 270807, интеллект: 154075)

1. http://www.paco.net/~indexless/wwg/russian/obzor.htm | Показать найденные слова
   О книге Дугласа Хофштадтера 'Гедель, Эшер, Бах: вечная золотая цепь'
   http://www.paco.net/~indexless/wwg/russian/obzor.htm - 31К - строгое соответствие
   Похожие документы | Все с сервера не менее 1 док.

2. SMIsC: July'97 | Показать найденные слова
   СМИСК - июль"97 Система СМИСК - SMIsC, System of Meaning Integrities structural Creation, "Система структурного создания смысловых целостностей" - работает с большими совокупностями элементов ...
   http://www.HINTECH.ru/smiscr/smiju97r.htm - 18К - 04.11.1998 - строгое соответствие
   Похожие документы | Все с сервера не менее 1 док.

3. ПРОДУКТЫ -- ОБЪЕКТНО-ОРИЕНТИРОВАННЫЕ СИСТЕМЫ: Состояние и перспективы | Показать найденные слова
   ЧАСТЬ С: ПРОДУКТЫ 1. СОЗДАНИЕ ПРИКЛАДНЫХ ПРОГРАММ 1.1. Группы продуктов В этой главе рассматривается развитие объектно-ориентированной технологии на пути занятия важного места в разработке ПО в 90-х ...
   http://www.math.rsu.ru/smalltalk/obzornew-3.ru.html - 118К - 09.02.2000 - строгое соответствие
   Похожие документы | Все с сервера не менее 1 док.

4. Frame Oriented Programming | Показать найденные слова
   Pisoft Group Исследовательские программы на уровне коммерчеких продуктов [Содержание | Главная | Новости | Программные продукты] Фреймовый подход от Pisoft Group
   http://www.videoscan.ru/pisoft/frame.htm - 6К - 05.06.2000 - строгое соответствие
   Похожие документы | Все с сервера не менее 1 док.

5. Движение порядка в природе | Показать найденные слова
   Движение порядка в природе Михаил Сухарев (Проект препринта, утвержденный к печати в 1986 г. но не напечатанный, ибо "это не марксизм". Рецензенты были правы - это не марксизм, хотя и замаскировано ...
   http://www.krc.karelia.ru/krcras/ies/personnel/suharev/Documents/OldpISO.htm - 101К - 02.06.2000 - строгое соответствие
   Похожие документы | Все с сервера не менее 1 док.

6. Движение порядка в природе | Показать найденные слова
   Движение порядка в природе Михаил Сухарев (Проект препринта, утвержденный к печати в 1986 г. но не напечатанный, ибо "это не марксизм". Рецензенты были правы - это не марксизм, хотя и замаскировано ...
   http://suharev.narod.ru/Documents/Oldpages.htm - 99К - 11.06.2000 - строгое соответствие
   Похожие документы | Все с сервера не менее 1 док.

7. 210300 вечерний | Показать найденные слова
   Специальность "Роботы и робототехнические системы" (210300) ГОСУДАРСТВЕННЫЙ ОБРАЗОВАТЕЛЬНЫЙ СТАНДАРТ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ Государственные требования к минимуму содержания и уровню ...
   http://www.mirea.ru/2103eve.htm - 46К - 04.12.1998 - строгое соответствие
   Похожие документы | Все с сервера не менее 2 док.

Рис. 4.7. Отклик «Яндекс» на запрос '*фрейм:10 "искусственный интеллект"*'

8. Типовая программа по базам данных и экспертным системам | Показать найденные слова
Типовая Программа д и с ц и п л и н ы "БАЗЫ ДАННЫХ И ЭКСПЕРТНЫЕ СИСТЕМЫ" Содержание СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ (СУБД) ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ЭКСПЕРТНЫЕ СИСТЕМЫ
http://sv.uven.ru/dep/facultet/mehmat/data/m/tpm/tbdes.htm - 6К - 22.03.1999 - строгое соответствие
Похожие документы | Все с сервера не менее 1 док.

9. ПРЕДСТАВЛЕНИЕ ЗНАНИЙ В СИСТЕМЕ РАСПРЕДЕЛЕННЫХ БАЗ ЗНАНИЙ И ДАННЫХ В INTERNET / I .. | Показать найденные слова
дизайн, полиграфия, баннеры, разработка логотипов, разработка баннеров, оформление, оформление плакатов, анимация, мультипликация, карикатуры, консультации, экспертные системы, аналитические системы, ...
http://amfora.lvs.ru/freeware/articles/novt97-2.htm - 14К - 18.04.2000 - строгое соответствие
Похожие документы | Все с сервера не менее 1 док.

10. Вестник АмГУ | Показать найденные слова
vestnik:/ все выпуски / выпуск 7 С. В. Агеев О РОЛИ ФРЕЙМОВ ЗНАНИЙ В ИНТЕРПРЕТАЦИИ МЕТАФОРИЧЕСКИХ ВЫРАЖЕНИЙ
http://www.amursu.ru:8100/vestnik/7/6_7_99.html - 15К - строгое соответствие
Похожие документы | Все с сервера не менее 1 док.

Страницы: **1** | 2 | Еще >>

Окончание рис. 4.7

Запрос: анафора && антецедент

Результат поиска: страниц - 3, серверов - 3 (анафора: 471, антецедент: 198)

1. И.Е. Ким. Выражение рефлексии в русском языке (значения 'контроль' и 'сопричастн .. | Показать найденные слова
И.Е. Ким Выражение рефлексии в русском языке (значения 'контроль' и 'сопричастность') 1. Говоря о релевантных в языковом отношении классификациях "явлений действительности", можно обнаружить ...
http://res.krasu.ru/paradigma/3/kim.htm - 42К - 17.12.1999 - строгое соответствие
Похожие документы | Все с сервера не менее 1 док.

2. ЛИНГВИСТИКА. ЧАСТНЫЕ АСПЕКТЫ ЯЗЫКОВОГО АНАЛИЗА. | Показать найденные слова
ЛИНГВИСТИКА ЧАСТНЫЕ АСПЕКТЫ ЯЗЫКОВОГО АНАЛИЗА Л.В.Александрова О классификации слов в узусных грамматиках XVII века Начавшийся в XVI в. процесс нормализации, упорядочения французского языка получил ...
http://www.pomorsu.ru/ScientificLife/Library/Sbornic2/article4.htm - 20К - 29.10.1999 - строгое соответствие
Похожие документы | Все с сервера не менее 1 док.

3. Народ.Ру: ЦИЦЕРОНЫ ФИЛФАКА | Показать найденные слова
Аналогичная страничка Юли Евтеевой Родная кафедра ЦИЦЕРОНЫ ФИЛФАКА ОГЛАВЛЕНИЕ Прохоров и Ильюшина Баранов
http://mitrius.narod.ru/prepody.html - 26К - 17.04.2000 - строгое соответствие
Похожие документы | Все с сервера не менее 1 док.

Рис. 4.8. Отклик "Яндекс" на запрос '*анафора && антецедент*'

4.2.2. Замечание о реализации

Делая ставку на специфичность тезауруса и размещая его на стороне клиента (пользователя), мы можем придти к проблеме (пока только гипотетической) разработки и распространения готовых тезаурусов. Поэтому выше (п. 3.1) мы уже указывали на то, что Internet должен быть не только пространством поиска информации, но и средой для коммуникации и объединения усилий разработчиков и пользователей тезаурусов (рис. 3.1).

Для этого необходимо создать оболочку (собственно ассистента), стандартизировать формат представления тезаурусов и организовать сайт для обмена тезаурусами и в последующем – возможно, – их интеграции. Формат должен отвечать требованиям открытости, переносимости, расширяемости.

Примерно таким образом поступила компания ABBYY (www.abbyy.ru) – разработчик семейства словарей LINGVO. На сайте www.lingvo.ru выкладываются разработанные пользователями подключаемые словари, проводятся конкурсы и ведутся дискуссии. Internet знает множество примеров синергетических эффектов подобного рода[1].

4.3. Результаты и выводы

На основании проверки разработанных процедур и методов можно сделать следующие выводы.

1. Полученные на предыдущем этапе дискриминантные функции и параметры «стилистическая информативность» хорошо работают на документах тестового массива.

2. Для реализации процедуры стилистической классификации документов при поиске информации в Internet предложены следующие варианты:

а) целенаправленный поиск научных документов,

---

[1] В качестве примера обычно приводят Linux и проект GNU. Еще один пример - проект SETI @ home. Международное движение SETI (Search for Extraterrestrial Intelligence, Поиск внеземного разума) планирует объединить персональные компьютеры добровольных участников во всем мире в один большой суперкомпьютер для обработки данных, поступающих от радиотелескопа, расположенного в Пуэрто-Рико. Проект lib.ru Максима Мошкова на таком фоне смотрится скромно, однако и он – типичный пример “самоорганизующейся системы” в Internet.

b) стилистическое ранжирование документов.

3. Результаты, полученные при использовании программы *ProThes Q* и тезауруса по компьютерной лингвистике, позволяют сделать вывод об эффективности предложенного метода расширения запросов. Метод позволяет эффективно управлять полнотой, точностью поиска, а также устранять неоднозначность слов за счет указания семантического окружения.

4. Проверка показала, что возможно совместное использование средств повышения эффективности поиска научной информации, обсуждаемых в работе.

Заключение

В результате проведенных исследований предложены методы повышения эффективности поиска научной информации в Internet.

Основные результаты работы заключаются в следующем:

1. Обосновано использование стилистических характеристик документов для повышения эффективности поиска научной информации в Internet. Использование стилистических параметров может расширить, с одной стороны, выразительные возможности языка запросов, а с другой – методы ранжирования списка выдачи.
2. Сформулированы подходы к задаче автоматической классификации документов по стилям, критерии для выбора параметров и метода классификации. Сформирован набор параметров классификации, разработана методика их вычисления. В качестве метода классификации выбран дискриминантный анализ.
3. С помощью процедур дискриминантного анализа и метода главных компонент получены дискриминантные функции с хорошим качеством классификации. Проведена оптимизация набора параметров классификации. Выявлена геометрическая структура классов-стилей.
4. Введен параметр «стилистическая информативность документа», который может использоваться при поиске информации.
5. Разработана программа, реализующая на уровне макета метод стилистической классификации и вычисление «стилистической информативности документа».
6. На тестовом массиве документов проведена проверка полученных процедур стилистической классификации. Проверка доказала надежность метода и его практическую значимость.

7. Предложены варианты практической реализации метода стилистической классификации: целенаправленный поиск научных документов, деление документов на «информативные» и «образные», ранжирование документов с использованием шкалы «стилистичская информативность».
8. Определены условия для эффективного применения процедуры расширения запросов с помощью тезауруса для поиска научной информации в Internet:
   - тезаурус отражает терминологию достаточно узкой научной области;
   - в тезаурусе используется набор сильно дифференцированных семантических отношений;
   - тезаурус располагается на стороне пользователя (клиента).
9. Описана архитектура, процедура и модель расширения запроса с помощью тезауруса с сильно дифференцированными семантическими отношениями.
10. Разработан интерфейс процедуры расширения запроса, а так же программа, которая реализует эту процедуру на уровне макета.
11. С использованием тезауруса по компьютерной лингвистике проведена проверка процедуры расширения запроса. Результаты позволяют сделать вывод об эффективности предложенного метода. Метод позволяет эффективно управлять полнотой, точностью поиска, а также устранять неоднозначность слов за счет указания семантического окружения. Проверка показала, что возможно совместное использование средств повышения эффективности поиска информации, обсуждаемых в работе.
12. Результаты работы прошли испытания и внедрены в компании “Конвекс” (Екатеринбург), Свердловской областной универсальной научной библиотеке им. В.Г.Белинского, НИИ ЦветМет (Екатерин-

бург). Результаты работы используются в научных исследованиях и учебном процессе на кафедре риторики и стилистики русского языка Уральского государственного университета и на кафедре вычислительной техники Уральского государственного технического университета.

В заключение мы хотим поблагодарить д-ра филол. наук, профессора Тамару Вячеславовну Матвееву и канд. физ.-мат. наук, доцента Юрия Борисовича Мельникова за участие в обсуждении результатов и содержания работы; компанию «Агама» (www.agama.com) – за предоставленный модуль морфологического анализа; Михаила Щекотилова и Илью Бирюкова – за помощь в создании программ.

Библиографический список

1. Абраменко А. Ненайденный документ — потерянный документ. – Электронный офис. – 1998. – №2. – В работе использовалась электронная версия: http://www.vest.msk.ru/EO/articles/2-98_04.htm .
2. Айвазян С. А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Основы моделирования и первичная обработка данных: Справ. изд. – М.: Финансы и статистика, 1983. – 471 с.
3. Андреев Н.Д. Статистико-комбинаторные методы в теоретическом и прикладном языковедении. – Л.: Наука, 1967. – 403 с.
4. Андрющенко В.М. Концепция и архитектура Машинного фонда русского языка. – М.: Наука, 1989. – 196 с.
5. Ашманов И. Информация и знания: невидимая грань // Электронный офис. – 1998. – №2. – В работе использовалась электронная версия: http://www.vest.msk.ru/EO/articles/2-98_02.htm .
6. Ашманов И., Руссова Н. Поисковые системы русскоязычных документов // Электронный офис. – 1997. – Май/Июнь. – В работе использовалась электронная версия: http://www.vest.msk.ru/EO/articles/may_july97_5.htm
7. Ашманов И. Автоматический поиск документов: осознанная необходимость // Электронный офис. – 1996. – Октябрь. – В работе использовалась электронная версия: http://www.vest.msk.ru/EO/articles/eo_oktober_4.htm .
8. Барлас Л.Г. Русский язык. Стилистика. Пособие для учителей. – М.: Просвещение, 1978. – 256 с.
9. Боровиков В.П. Популярное введение в программу STATISTICA. – М.: КомпьютерПресс, 1998. – 267 с.
10. Боровиков В.П., Боровиков И.П. STATISTICA – статистический анализ и обработка данных в среде Windows. – 2-е изд., стереотип. – М.: Инф.-изд. дом “Филинъ”, 1998. – 608 с.

11. Браславский П.И. Автоматическая классификация документов Internet по стилям: реализация макета: Доклад V рабочего совещания по электронным публикациям – EL-PUB-2000, Новосибирск, Академгородок, ИВТ СО РАН, 21-23 июня 2000 г. – http://www.ict.nsc.ru/ws/el-pub-2000/29/
12. Браславский П.И. Пути повышения эффективности поиска научной информации в Internet: Тезисы доклада IV рабочего совещания по электронным публикациям – EL-PUB-99, Новосибирск, Академгородок, ИВТ СО РАН, 21-23 апреля 1999 г. – http://www.ict.nsc.ru/ws/show_abstract.dhtml?2+44
13. Браславский П.И. Распознавание стилей речи применительно к информационному поиску: постановка задачи // Математические структуры и моделирование: Сб. научн. тр., Вып. 3. / Под ред. А.К.Гуца. – Омск: Омск. гос. ун-т, 1999. – С. 134-140.
14. Браславский П.И. Расширение поискового запроса с помощью тезауруса с сильно дифференцированными связями: Тезисы доклада рабочего совещания "Новые Интернет-технологии", Петрозаводск, КарНЦ РАН, 25-28 июня 2000 г. http://www.krc.karelia.ru/structure/math/conf/nit/papers/paper.ru.phtml?file=braslavsky.html
15. Браславский П.И. Стилистическая классификация документов в задачах информационного поиска // Проблемы теоретической и прикладной математики: Тезисы докладов 30-ой Региональной молодежной конференции. – Екатеринбург: ИММ УрО РАН, 1999. – С. 79-80.
16. Браславский П.И. Тезаурусное описание терминологии: принципы, технология, инструментальные средства // Материалы XXXV Международной научной студенческой конференции «Студент и научно-технический прогресс»: Информационные технологии. – Новосибирск: НГУ, 1997. – С. 18.
17. Браславский П.И., Гольдштейн С.Л., Ткаченко Т.Я. Тезаурус как средство описания систем знаний// Научно-техническая информация. Сер.2, – 1997. – №11. – С.16-21.

18.Браславский П.И., Ткаченко Т.Я. Тезаурус как инструмент описания терминологии и представления знаний. Принципы, технология, инструментальные средства // Информационные технологии, системы управления и электроника: Тезисы докладов всероссийской научно-технической конференции. – Екатеринбург: УГТУ, 1997. – С. 74.

19.Васильева А.Н. Курс лекций по стилистике русского языка. Научный стиль речи. – М.: Русский язык, 1976. – 189 с.

20.Виноградов В.В. Проблема авторства и теория стилей. – М.: Гос. изд. худ. литературы, 1961. – 614 с.

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

22.Головин Б.Н. Язык и статистика. – М.: Просвещение, 1970. – 190 с.

23.Гринберг И., Гарбер Л. Разработка новых технологий информационного поиска // Открытие системы, – 1999. – №9-10(41-42). – С 28-30.

24.Едемский М. Бродяга //Мир Internet. – 1998. – №1(16). – С. 68-74.

25.Ершов Ю.Л., Палютин Е.А. Математическая логика: Учеб. пособие для вузов. – 2-е изд., испр. и доп. – М.: Наука, Гл. ред. физ.-мат. лит., 1987. – 336 с.

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

27.Изменения в языке научной прозы/ О.Б.Сиротинина, С.А.Бах, В.А. Богданова и др. // Вопросы стилистики. Вып. 3. – Саратов: Изд-во Сарат. ун-та, 1969. – С. 37-55.

28.Изменения в языке публицистики (на материале международных обзоров) / О.Б.Сиротинина, С.А.Бах, В.А.Богданова и др. // Вопросы стилистики. Вып. 3. – Саратов: Изд-во Сарат. ун-та, 1969. – С. 5-36.

29.Каменнова М. Управление электронными документами: технологии и решения // Открытие системы, – 1995. – №4. – С. 38-44.

30.Капустин В. Поиск информации в Интернете // Мир Internet, – 1998. – №9. – С. 54-59.

31.Карташева Е. Интеллектуальные поисковые системы Excalibur // Сети, – 1997. – №6. – С. 98-105.

32.Кауфман С.И. Из курса лекций по статистической стилистике. – М.: МОПИ, 1970. – 319 с.

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

34.Ким Дж.-О., Мьюллер Ч.У. Факторный анализ: статистические методы и практические вопросы // Факторный, дискриминантный и кластерный анализ: Пер. с англ. – М.: Финансы и статистика, 1989. – С. 5-77.

35.Кияк Т.Р. Лингвистические аспекты терминоведения. – К.: УМК ВО, 1989. – 104 с.

36.Клекка У.Р. Дискриминантный анализ // Факторный, дискриминантный и кластерный анализ: Пер. с англ. – М.: Финансы и статистика, 1989. – С. 78-138.

37.Клочкова Э.А. О влиянии формы разговорной речи на распределение классов слов // Русская разговорная речь. Сб. науч. трудов. – Саратов: Изд-во СГУ, 1970. – С. 126-134.

38.Колмановская Е. Что и как найти в русском Internet? //Мир ПК – 1999. – №10. – С. 70-73.

39.Кожина М.Н. К основаниям функциональной стилистики. – Пермь: Б. и., 1968. – 251 с.

40.Кожина М.Н. О речевой системности научного стиля сравнительно с некоторыми другими. – Пермь: Б. и., 1972. – 395 с.

41.Кожина М.Н. Стилистика русского языка. – М.: Просвещение, 1977.

42.Кожина М.Н. Стилистика русского языка. 3-е изд., перераб. и доп. – М.: Просвещение, 1993. – 221 с.

43.Лейчик В.М. Терминология информатики: теоретические и практические вопросы// Информатика: Итоги науки и техники. Т.2. – М., 1977. – С.40-53.

44.Михайлов А.И., Черный А.И., Гилиревский Р.С. Основы информатики. 2-е изд., перераб. и доп. – М.: Наука, 1968. – 756 с.

45.Налимов В.В. Вероятностная модель языка. О соотношении естественных и искусственных языков. 2-е изд., перераб. и доп. – М.: Наука, 1979. – 303 с.

46.Никитина С.Е. Семантический анализ языка науки. (На материале лингвистики.) – М.: Наука, 1987. – 141 с.

47.Никитина С.Е. Тезаурус по теоретической и прикладной лингвистике. – М.: Наука, 1978. – 375 с.

48.Осиновский Д.Э., Ровный А.С., Новицкий Д.В. Компьютеризованный текстологический анализ исторических документов: возможности программы НТА // Информационный бюллетень ассоциации "История и компьютер" – 1998. – № 22 (январь). – В работе использовалась электронная версия: http://kleio.dcn-asu.ru/aik/bullet/22/23.shtml

49.Пархоменко В.Ф. Работа с русскоязычными полнотекстовыми базами данных в ИПС АРТЕФАКТ // Научно-техническая информация. Сер.2. – 1998. – №1. – С. 25-29.

50.Плешко В.В. Визуализация массивов полнотекстовых документов в информационных системах // Материалы конференции "Анализ систем на рубеже XXI века". – М., 1997. – В работе использовалась электронная версия: http://is.park.ru/park/websom/ansys97.zip .

51.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. – 606 с.

52.Применение статистических методов для интеллектуальной компьютерной обработки текстов/ И.Ашманов, С.Григорьев, В.Гусев и др. // Труды Международного семинара по компьютерной лингвистике и ее приложениям “Диалог 97”, 1997. – В работе использовалась электронная версия: http://www.medialindua.ru/mlingua/News/Dialog97/ashmanov.htm (26.08.98).

53.Россия в Интернете. Всероссийский опрос IV’99./Агентство monitoring.ru. – http://www.monitoring.ru/internet/archive/1999/IV/1/ .

54.Россия в Интернете. Февраль 2000: Обзор отчета /Агентство monitoring.ru. – http://www.monitoring.ru/internet/summarizing.html .

55.Русская разговорная речь. Фонетика. Морфология. Лексика. Жест. – М.: Наука, 1983. – 238 с.

56.Садовничий В. А., Васенин В. А., Мокроусов А. А., Тутубалин А. В. Российский Интернет в цифрах и фактах. – М.: Изд-во МГУ, 1999. – 148 с.

57.Сиротинина О.Б. Современная разговорная речь и ее особенности. – М., 1974. – 144 с.

58.Солтон Дж. Динамические библиотечно-информационные системы. – Пер. с англ. – М.: Мир, 1979. – 558 с.

59.Степанов Ю.С. Стиль // Языкознане. Большой энциклопедический словарь /Гл. ред. В.Н.Ярцева. – 2-е изд. – М.: Большая Российская энциклопедия, 1998. – С. 494-495.

60.Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления: ГОСТ 7.25-80. – Введ. 01.01.82. – М.: Изд-во стандартов, 1981. – 15 с.

61.Тезаурус научно-технических терминов / Под. ред. Шемакина Ю.И. – М.: Воениздат, 1972. – 671 с.

62.Тулдава Ю. Опыт классификации текстов с помощью кластер-анализа // Актуальные проблемы квантитативной лингвистики и автоматического анализа текстов.– Тарту, 1981. – С. 136-157. – (Труды по лингвостатистике / Тартуский гос. ун-т; Кн. VII.)

63.Хмелев Д. Лингвоанализатор. – http://www.rusf.ru/books/analysis/ .

64.Храмцов П. Информационно-поисковые системы Internet // Открытые системы, – 1996. – №3(17). – С. 46-49.

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

66.Храмцов П.Б. Лабиринт Internet. Практическое руководство. – М.: “ЭЛЕКТРОИНФОРМ”, 1996. – 256 с.

67.Частотный словарь общенаучной лексики. Под общ. ред. Е.М. Степановой. – М.: Изд-во МГУ, 1970. – 87 с.

68.Частотный словарь русского языка. Под. ред. Л.Н.Засориной. – М.: Русс. яз., 1977. – 934 с.

69.Черный А.И. Общая методика построения тезаурусов // Научно-техническая информация. Сер.2, – 1968. – № 5. – С.17-33.

70.Чибисов А. Поисковые возможности Excalibur RetrievalWare // Открытие системы, – 1996. – №5(19). – С 49-53.

71.Шемакин Ю.И. Тезаурус в автоматизированных системах управления и обработки информации. – М.: Воениздат, 1974. – 188 с.

72.Шехтман Н.А. Об одном принципе расширения и углубления тезауруса // Научно-техническая информация. Сер.2, – 1981. – №7. – С.10-12.

73.Шрейдер Ю.А. Тезаурус в информатике и теоретической лингвистике // Научно-техническая информация. Сер.2, – 1971. – №3. – С. 21-24.

74.Электронный учебник по статистике / StatSoft, Inc. – 1999. – http://www.statsoft.ru/home/textbook/default.htm.

75.Aitchison J. et al. Thesaurus construction and use: a practical manual. 3rd edn. – London: Aslib, 1997. – 186 p.

76.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. – P. 544-552.

77.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, – 1996. – Vol. 7, No. 1. – P. 88-102.

78.Ellis D. Progress and Problems in Information Retrieval. – London: Library Association Publishing, 1996. – 456 p.

79.Hert C.A. Understanding information retrieval interactions: theoretical and practical implementations. – London: Ablex Publishing Corp., 1997. – 326 p.

80. Karlgren J. Stylistic Variation in an Information Retrieval Experiment // Proc. 2nd International Conference on New Methods in Language Processing (NeMLaP 2). – Ankara: Bilkent University, 1996. – В работе использовалась электронная версия: http://www.sics.se/jussi/Artiklar/1996_NeMLaP_Bilkent/nemlap.ps .
81. Karlgren J., Cutting D. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis // Proc. 15th International Conference on Computational Linguistics (COLING). – Kyoto, 1994. – Vol. 2. – P. 1071-1075.
82. Kirriemuir J.W., Willet P. Use of Cluster Analysis Methods for Analysing the Outputs of Multiple-Database Searches //Electronic Library and Visual Information Research. Proc. 2nd ELVIRA Conference. – London, 1995. – P. 117-126.
83. Schwarz C. Web Search Engines // Journal of the American Society for Information Science. – 1998. – №49(11). – P. 973-982.

Приложения

П1. Опытный массив документов

П1.1. Официально-деловой стиль

Федеральные законы Российской федерации (50)	
Номер закона	Название
N 15	О связи
N 24	Об информации, информатизации и защите информации
N 128	О внесении дополнений в закон российской федерации "О валютном регулировании и валютном контроле"
N 140	О внесении дополнения в федеральный закон "О некоммерческих организациях"
N 141	О налоге на отдельные виды транспортных средств
N 143	Об административной ответственности юридических лиц (организаций) и индивидуальных предпринимателей за правонарушения в области производства и оборота этилового спирта, алкогольной и спиртосодержащей продукции
N 144	О реструктуризации кредитных организаций
N 151	О внесении изменений и дополнений в закон РСФСР "О государственной налоговой службе РСФСР"
N 155	О внесении изменений и дополнений в федеральный закон "О введении в действие части первой налогового кодекса Российской Федерации"
N 157	О внесении изменений и дополнений в статью 238 уголовного кодекса Российской федерации
N 158	О внесении дополнений в уголовный кодекс Российской Федерации и дополнений и изменения в статью 126 уголовно-процессуального кодекса РСФСР
N 159	О введении в действие бюджетного кодекса Российской Федерации
N 160	Об иностранных инвестициях в Российской Федерации
N 161	О материальной ответственности военнослужащих

N 162	О порядке перевода государственных ценных бумаг СССР и сертификатов сберегательного банка СССР в целевые долговые обязательства Российской Федерации
N 163	О внесении изменения в статью 5 закона Российской Федерации "О государственных пенсиях в Российской Федерации"
N 165	Об основах обязательного социального страхования
N 168	О внесении изменений и дополнений в закон российской федерации "О налогах на имущество физических лиц"
N 169	О внесении изменения и дополнения в статью 3 закона Российской Федерации "О статусе судей в Российской Федерации"
N 171	О внесении изменения в статью 16 федерального закона "О государственных пособиях гражданам, имеющим детей"
N 172	О внесении изменений и дополнения в федеральный закон "О социальной защите инвалидов в российской федерации" и закон Российской Федерации "О государственных пенсиях в Российской Федерации"
N 174	О финансировании государственного оборонного заказа для стратегических ядерных сил Российской Федерации
N 175	О внесении изменений и дополнений в закон российской федерации "О занятости населения в Российской Федерации"
N 176	О почтовой связи
N 178	О государственной социальной помощи
N 181	Об основах охраны труда в Российской Федерации
N 183	Об экспортном контроле
N 184	Об общих принципах организации законодательных (представительных) и исполнительных органов государственной власти субъектов Российской Федерации
N 185	О внесении изменений и дополнений в федеральный закон "О порядке опубликования и вступления в силу федеральных конституционных законов, федеральных законов, актов палат федерального собрания"

N 189	О внесении дополнения в федеральный закон "О тарифах страховых взносов в пенсионный фонд российской федерации, фонд социального страхования российской федерации, государственный фонд занятости населения Российской Федерации и в фонды обязательного медицинского страхования на 1999 год"
N 190	О ратификации Европейской конвенции о выдаче, дополнительного протокола и второго дополнительного протокола к ней
N 193	О ратификации Европейской конвенции о взаимной правовой помощи по уголовным делам и дополнительного протокола к ней
N 194	О внесении изменений в статьи 10 и 15 федерального закона "Об оружии"
N 196	О внесении изменений и дополнений в федеральный закон "О проведении эксперимента по налогообложению недвижимости в городах Новгороде и Твери"
N 197	О тарифах страховых взносов в пенсионный фонд Российской Федерации, фонд социального страхования Российской Федерации, государственный фонд занятости населения Российской Федерации и в фонды обязательного медицинского страхования на 2000 год
N 198	Об участках недр, право пользования которыми может быть предоставлено на условиях раздела продукции (участке недр "северные территории")
N 199	Об участках недр, право пользования которыми может быть предоставлено на условиях раздела продукции (приобском (северном лицензионном участке) нефтяном месторождении)
N 200	О внесении дополнений в федеральный закон "О федеральном бюджете на 1999 год"
N 201	О потребительской корзине в целом по Российской Федерации
N 204	О внесении изменений и дополнений в закон Российской Федерации "Об организации страхового дела в Российской Федерации"
N 207	О внесении изменений и дополнений в закон Российской Федерации "О подоходном налоге с физических лиц"
N 208	О внесении изменений и дополнений в федеральный закон «О федеральной фельдъегерской связи»
N 209	О внесении изменения и дополнений в статью 10 закона Российской Федерации "О милиции"

N 210	Об административной ответственности юридических лиц за нарушение законодательства Российской Федерации о выборах и референдумах
N 211	Об общих принципах организации и деятельности ассоциаций экономического взаимодействия субъектов российской федерации
N 212	О внесении изменений и дополнений в закон Российской Федерации "О защите прав потребителей"
N 213	О внесении изменений и дополнений в гражданский кодекс Российской Федерации
N 214	О внесении изменения в статью 60 основ законодательства Российской Федерации об охране здоровья граждан
N 215	О внесении дополнений в статью 17 федерального закона "О лицензировании отдельных видов деятельности"
N 216	О внесении дополнения в статью 17 федерального закона "О лицензировании отдельных видов деятельности"

П1.2. Научный стиль

Труды Института математики и механики УрО РАН. Том 5, 1998. http://tom.imm.intec.ru/new/stom5koi.html	
1. Белоногов В.А.	Малые взаимодействия в группах
2. Васильков В.И., Масленков Г.В.	Исследование k-дуг в плоскости Хьюза порядка 9 с помощью ЭВМ
3. Финогенов А.А.	О p-Регулярных группах
4. Величко Н.В.	О сильной топологии
5. Пыткеев Е.Г.	К теории К-аналитических пространств
6. Агеев А.Л.	О единственности и неединственности решения совокупности эквивалентных по точности уравнений 1-го рода
7. Бердышев С.В.	Относительная константа Юнга пространства
8. Густомесов В.А.	О классификации разностных линейных систем второго порядка
9. Ким, А.В. Пименов В.Г.	О применении i-гладкого анализа к разработке численных методов решения функционально-дифференциальных уравнений
10. Новиков С.И.	О разрешимости одной спектральной краевой задачи
11. Бабенко А.Г.	Точное неравенство Джексона-Стечкина в пространстве

12. Балаганский В.С.	Аппроксимированные свойства множеств с выпуклым дополнением
13. Васильев А.И.	К теореме Бернштейна о последовательности наилучших приближений в пространствах
14. Попов В.Ю.	Точное неравенство Джексона-Стечкина в пространстве L2 на гиперболоиде
15. Кругликов С.В.	О свойствах разделения априорных задач гарантированного управления и оценивания. Операторная реализация
16. Логинова Н.М., Ченцов А.Г.	К вопросу о расширении стохастических ограничений в классе конечно-аддитивных вероятностей
17. Ноймарк Б.М., Исмаил-заде А.Т., Короткий А.И. и др.	Численная реализация гидродинамической трехмерной модели формирования осадочных бассейнов
18. Цепелев И.А.	Динамическое восстановление множества параметров в краевой задаче Гурса-Дарбу
19. Стефанова А.В.	Уравнение Гамильтона-Якоби (Беллмана) в нелинейной задаче импульсного уравнения
20. Ченцов А.Г.	Универсальная ассимптотическая реализация интегральных ограничений и конструкции расширения в классе конечно-аддативных мер
21. Еремин И.И.	Сигма-кусочные функции и задачи дизъюнктивного программирования
22. Попов Л.Д.	О точности решения внутренних подзадач в методе Хестенса-Пауэлла
Статьи из научных журналов из электронной библиотеки УГТУ (11) http://www.ustu.ru/library/catalog_magaz.html	
23. Швайка О.П.	Ароматические электролиты: межионная ассоциация азиниевых солей
24. Купряжкин А. Я., Рыжков М. В., Дудоров А. Г.	Взаимодействие гелия с ионами при его растворении в кристаллах фторидв стронция
25. Безруков И.Я., Кляйн С.Э., Набойченко С.С.	Проблемы и способы переработки отработанных ванадиевых катализаторов сернокислотного производства
26. Уфимцев В.М.	Грануляция в современных технологиях складирова-

	ния дисперсионных промышленных отходов
27. Мильман И. И., Кортов В.С., Никифоров С. В.	Интерактивный процесс в механизме термостимулированной люминесценции анион-дефектных кристаллов α-Al_2O_3
28. Огородников И.Н., Поротников А.В., Кудяков С.В. и др.	Стабильное и метастабильное оптическое поглощение нелинейных кристаллов LiB_3O_5
29. Кортов В.С., Мильман И.И., Никифоров С.В.	Особенности кинетики термостимулированной люминесценции кристаллов a-Al_20_3 с дефектами
30. Вайнштейн И. А., Зацепин А. Ф., Кортов В. С.	Особенности проявления правила Урбаха в стеклообразных материалах
31. Катышев С.Ф., Дубинин Б.В., Десятник В.Н.. и др.	Плотность и поверхностное натяжение расплавов системы LiCl-KCl-UCl_4
32. Набойченко С.С., Романова В.В., Халемский О.А.	Полупромышленные испытания автоклавной очистки растворов купоросного цеха от мышьяка
33. Набойченко С. С., Карелов С. В., Мамяченков С. В., Якорнов С. А.	Комплексная переработка свинецсодержащих техногенных отходов медеплавильных предприятий урала
Вестник Вавиловского общества генетиков и селекционеров (ВОГиС) (10) http://wwwicg.bionet.nsc.ru/vogis/win/2.htm	
34. Аксенович Т.И.	Генетический анализ: этапы развития, проблемы и перспективы
35. Каракин Е.И.	Нейрогенетика дрозофиллы и проблемы нейробиологии
36. Неуймин С.И.	Уровни группового приспособления райграса, LOLIUM PERENNE L., в зависимости от сочетания генотипов на микропопуляционном уровне организации
37. Железнов А.В.	О селекции растений на устойчивость к «загрязнителям» окружающей среды
38. Каракин Е.И.	Рак простаты: о диагностической ценности *fPSA/tPSA*

39. Кочетов А.В.	Проблемы трансгенеза. Феномен "замолкания" трансгенов и устойчивость растений к вирусным инфекциям
40. Дейнеко Е.В.	О потенциальных возможностях использования Т-ДНК почвенных бактерий
41. Омельянчук Л.В.	Теломера и репродуктивное старение клетки
42. Ланкин В.С.	Развитие исследований доместикационного поведения животных
43. Матушкин Ю.Г.	Существует ли ассоциированный с прионами патогенный агент?
Электронный журнал "Исследовано в России". Секция "Естественные науки" (4) http://zhurnal.mipt.rssi.ru/articles/n.html	
44. Макарьев Д.И., Гавриляченко В.Г.	Пьезоэлектрическая анизотропия сегнетоэлектрической керамики подвергнутой циклической переполяризации
45. Осина Е.Л., Юнгман В.С., Горохов Л .Н .	Термодинамические свойства молекул трииодидов лантанидов
46. Кустов Е.Ф., Кустов Д.Е.	Энергия многоэлектронных атомов (II)
47. Наумов В.Н.	Межмолекулярные и внутримолекулярные колебания в теплоемкости $CR(C_5H_7O_2)_3$
Доклады II Уральской региональной школы-семинара молодых ученых и студентов по физике конденсированного состояния (3) http://www.usu.ru/win/usu/events/1998/school-seminar/sd.html	
48. Вольхин И.Л., Жерихова М.Я., Коротаев Н.Н. и др.	Физическое моделирование распространения света через отдельную диэлектрическую частицу
49. Берг Д.Б.	Гетерофазные флуктуации плотности в липидном лиотропном жидком кристалле вблизи главного фазового перехода
50. Берг Д.Б., Хацевич С.Г., Приезжев А.В.	Кинетика агрегации лиотропных жидкокристаллических липидных везикул

Аннотации статей журнала Российского химического общества им. Д. И. Менделеева (2)	
51. 1994, т. 38, #6 .	http://www.chem.msu.su/rus/jvho/v38.html
52. 1995, т. 39, # 2.	http://www.chem.msu.su/rus/jvho/v39(2).html
Отдельные статьи	
53. В.А. Мухин Грибы и их роль в природе и в развитии цивилизации http://virlib.eunnet.net/proceedings/N12_99/win/16.htm	
54. В.Г. Щетинин Многослойная самоорганизация нейронных сетей оптимальной сложности http://www.chat.ru/~neurolab/	

П1.3. Публицистический стиль

Автор(ы)	Название	Дата
Gazeta.Ru		
1. Александр Корнилов	«Выход» Владимира Яковлева из ОВР	18.12.1999
2. Редакционная	В Берлине завершается «большая русско-чеченская неделя»	18.12.1999
3. Василий Сергеев	Новый флаг над Макао	18.12.1999
4. Андрей Матяш	Гантамиров потерял в Грозном тысячу человек	29.12.1999
5. Александр Лиманов	«Единство» не сумело достичь согласия	29.12.1999
6. Алексей Маслов	Америка поставила на Путина	29.12.1999
7. Василий Сергеев	Сектанты возрождают «АУМ Синрикё»	29.12.1999
8. Трофим Лобачев	Путин и жизнь	05.01.2000
9. Василий Сергеев	Индия объявила Пакистану информационную войну	05.01.2000
10. Андрей Матяш	Штурм Грозного провалился	08.01.2000
11. Петр Иванов	Назад, к победе коммунизма	08.01.2000
12. Александр Лиманов	В бой идут одни беспартийные	17.01.2000
13. Андрей Степанов	В США нашли замену российским акциям	17.01.2000
14. Александр Лиманов	Новые победы федералов на пропагандистском чеченском фронте	17.01.2000
15. Редакционная	Как юрист-пианист разыграл западную прессу	18.01.2000
16. Андрей Матяш	Штурм Грозного. Вторая попытка	18.01.2000

17. Трофим Лобачев	Олбрайт и Милошевич подозреваются в убийстве	18.01.2000
18. Редакционная	Эта Дума может все	19.01.2000
19. Редакционная	Путинова победа	19.01.2000
20. Олег Семенов	Избирателей сгоняют в Чечню	19.01.2000
21. Александр Корнилов	Селезнев и «девушки»	20.01.2000
22. Дмитрий Радышевский	...чинить насилие	20.01.2000
23. Александр Чудов	Метод Ельцина живет и побеждает	20.01.2000
24. Наталья Копосова	Наркоманы, чужие попы и китайцы не пройдут!	21.01.2000
25. Александр Лиманов	ФСБ начала отсчет поляков	21.01.2000
26. Александр Лиманов	Меньшевик – это только звучит гордо	22.01.2000
27. Андрей Матяш	Боевики стали «крысами»	22.01.2000
Polit.Ru		
28. Елена Висенс	Баски возвращаются к террору	01.12.1999
29. Миша Фишман	Феномен "Единства": шизофрения, как и было сказано	10.12.1999
30. Алексей Улюкаев	Конь и осел в семье Ульяновых	24.12.1999
31. Борис Грозовский	Транспортировка азербайджанской нефти. У России опять есть шансы	29.12.1999
32. Олег Вьюгин	Основная проблема - отсутствие внешнего финансирования	13.01.2000
33. Сергей Алексашенко	Неспособность российских властей получить кредит МВФ в конце 1998 года может дорого обойтись нашей экономике	18.01.2000
Vesti.Ru		
34. Лаврентий Пугачев	Левая подпорка	16.12.1999
35. Геннадий Нечаев	Рушайло стелит соломку	17.01.2000
36. Лаврентий Пугачев	Великая правая мечта	17.01.2000
37. Лена Березницкая	Старый лис	17.12.1999
38. Владимир Петров	Губернаторы вцепились в свои кресла	17.12.1999
39. Владимир Петров	Строев метит в президенты	24.12.1999

40. Владимир Корсунский	Меж Чечней и Тамбовщиной	27.12.1999
41. Владимир Тучков	Лукашенко с человеческим лицом	28.12.1999
42. Владимир Тучков	Скуратов выкапывается	28.12.1999
43. Вера Мамонтова	Санпропускник для депутатов	28.12.1999
44. Эльза Добер	Хорошист Иванов	28.12.1999
45. Владимир Тучков	Дело Александра Никитина лопнуло	30.12.1999
46. Лев Бруни	Дед сделал свое дело	31.12.1999
47. Аркадий Дубнов	Талибы стали нам ближе	18.01.2000
48. Владимир Тучков	Коль стал трижды бывшим	19.01.2000
49. Владимир Петров	Один другого стоит	19.01.2000
50. Лаврентий Пугачев	Не было бы счастья	19.01.2000
51. Сергей Митрофанов	Интеллигенция ужаснулась	19.01.2000
52. Вера Мамонтова	Дума что дышло	20.01.2000
53. Александра Семушкина	Оптовикам скидки	20.01.2000
54. Лена Березницкая	Ястребы не тонут	20.01.2000
55. Сергей Митрофанов	Фемида парализована	20.01.2000
56. Наталья Ищенко	Спикера убрали	21.01.2000
57. Лаврентий Пугачев	Что-то Вешняков притих	21.01.2000
58. Яна Лолаева	Аграрии рвутся на волю	21.01.2000
59. Сергей Митрофанов	Тайна взрыва не раскрыта	21.01.2000
60. Владимир Тучков	Массовый отлов шпионов	21.01.2000
61. Глеб Черкасов	Последняя кампания	09.02.2000

П1.4. Художественный стиль

Рассказы, поданные на конкурс сетевой литературы «Тенета-98»

http://www.teneta.rinet.ru/1998/rasskaz/

Автор(ы)	Название
1. Алексей Александров	Лесной человек
2. Али	Случай в Варфоломеевскую ночь
3. Алиса	Знаешь...
4. Алексей Андреев	Цветок красоты
5. Сергей Ашанов	Опыт бегства

6. Николай Байтов	Искушение поэта в снежной пустыне
7. Александр Бобраков-Тимошкин	Underground
8. Владимир Боровой	Осенние фото
9. Александра Бутузова	Джой
10. Андрей Бычков	Мат и интеллигенция
11. Валка	Без названия
12. Дмитрий Власенко	Нравственное письмо Сенеки к Луцилию
13. Евгений Волошин	Этюд о вечном
14. Кузьма Востриков	Путешествие
15. Кузьма Востриков	Рассказы в письмах
16. ВьюГа	Дом восходящего солнца
17. Данила Гамлет	Живая вода
18. Игорь Герцев	С днем рождения
19. Иван Глаголев	Микророман в письмах
20. Данила Давыдов[1]	Подземные жители
21. Вадим Залива	Пес
22. Ольга Зондберг	Вечерний звон для утреннего пользования
23. Евгений Иz	ОтСчет детства
24. Индеец Юра	Снова любовь и кровь (романтическая сказка)
25. Рамон Карнизов	Конфабуляция
26. Роман Карнизов	Фотографический снимок с трехмесячной выдержкой
27. Дмитрий Коваль	Хоккеисты
28. Макс Козлов	Марина на Таганке (Стихотворения в прозе)
29. Константин Лаврентьев	Коллизия
30. Валерий Лебедев	Служил он в Таллинне при Сталине
31. Анна Левина	Коммуналка
32. Лиза	Гость
33. Лисса	Одинокое сердце
34. Виктория Львова	Тюремный священник
35. Станислав Львовский	Народное телевидение
36. Серж Мелентьев	Мандринчики

[1] Этот рассказ был исключен из опытного массива.

37. Евгений Медников	Выйти замуж за гинеколога
38. Эрик Миккивер	О праведных и неправедных жителях
39. Сергей Михайлов	По ту сторону зеркала
40. Елена Мулярова	Кимоно ее бабушки
41. Наташа	Стип и Ли
42. Мария Ордынская	Бузиненыш
43. Жорж Полежаев	Lakosta
44. Олег Попов, Владимир Белобров	Радость от ума
45. Олег Потоцкий	Муха
46. Михаил Прокопьев	Корова
47. Владимир Пузий	Возвращение
48. Владимир Пузий	...На приезжих гладиаторов
49. Владимир Севриновский	Баллада о критике
50. Валерий Сегаль	Счастливый человек Светка Иванова
51. Сергей Серебрин	Вокзал на троих
52. Полина Сиголаева	Розовый бегемот
53. Гера Скрипка	Госпиталь
54. Вадим Смоленский	Очки
55. Леонид Стомакаров	Пенетрэйшн в ж (письма Лу)
56. Валерий Суси	Дневник
57. Тихая сапа	Калитка
58. Алексей Толкачев	Бабье лето
59. Алексей Толкачев	Европа. 21 век
60. Анатолий Туровский	Фима и Сорок Вторая
61. А. Туяр	Логос
62. Виктория Угрюмова	Саксофонист
63. Максим Федосеев	Несуществующее животное
64. Виктория Фомина	Зона турбулентности
65. Дмитрий Чела	Дикая балка
66. Юрий Черняков	Хасбулат удалой...
67. Владислав Чопоров	Неформал
68. Сергей Шаманов	Тень камня
69. Александр Шамес	Чашка кофе по-турецки

70. Надежда Шахова	Мы и Макларен
71. Виктор Шнейдер	Лапины
72. Эдуард Шульман	Трубач
73. Кирилл Щербицкий	Перерождения в Лиссе
74. Кирилл Якимец	Костылита
75. Envil	Белая цапля
76. Grassy	Спреднеталитянские легенды
77. Lunatic	Холодное утро
78. Ric	Мысли...
79. Stan	Раки
80. Valium	Внутренний монолог старухи

П1.5. Разговорный стиль

На рис. П1.1 в качестве примера разговорного стиля приведен листинг чата «На плотинке» (http://www1.ekaterinburg.com/leisure/chat/) от 01.02.2000.

Дурень (19:03): **Здарова BUBU**
Madonna (19:04): **Чеширский сколько тебе лет ?**
Vlasta (19:04): **Privet!!!!**
Yri (19:04): Баг...наверное.....
Будда (19:04): **МЯУ ВСЕМ НА ПРОЩАНИЕ!!!!!!!!!!!!!!!!!!!!!!!!**
Mayson (19:04): Юр, у тя каки общие впечатления от того вечера то остались.....????
ARAHNID (19:04): *БУБУ неа савсем недобрый!!! :-((*
BUBU (19:04): *Привет, Иван :))*
Новый человек в чате: (:]shaitan[:).
Новый человек в чате: Mr big.
Чеширский_ (19:04): **Мадонна, Чеширские живут долго...лет этак 2000**
Новый человек в чате: Оттяжник.
(:]shaitan[:) (19:04): **Всем привет**
Новый человек в чате: саня.
Aleksey (19:05): **Шайтан хая**
Чеширский_ (19:05): **Мадонна, а тебе похоже 16**
Новый человек в чате: РАБ.
Madonna (19:05): **Чеширский а чем ты занимаешься ?**
.... (19:05): Мадонна, я так и думала , что он отьедет, я тоже спрашивала..... :))))))
Vlasta (19:05): **Ne-ne!**
Yri (19:05): Мэйс...ваще..полный пинцет :)))))) Правда я там диван несколько сломал....грустно :((
леди (19:05): **все будет в ажуре. Я организую.Катечка**
Чеширский_ (19:05): **Мадонна, сейчас грущу**
ЯНЕЖ (19:05): *ПРИВЕТ ШАЙТАН*
BUBU (19:05): *Арахнид, а что тогда внем плохого?*
Новый человек в чате: Гаденыш.
Новый человек в чате: ХАЧИК.
Aleksey (19:06): **Гаденыш ты же уходила?**
Madonna (19:06): **Чеширский с чего ты взял ?**
ХАЧИК (19:06): **ПРИЕВЕТ!**

Рис. П1.1 Листинг чата «На плотинке», 01.02.2000

Новый человек в чате: DELTA-1(man).
РАБ. (19:06): ЛЕДИ куда приехать то ??? :)))
(:]shaitan[:) (19:06): Vlasta привет
Yri (19:06): Гадик :))))))))
Чеширский_ (19:06): придумал
Mayson (19:06): Юр, дык я знаю.....:)))) я и напелся....и напился.....:))))
Новый человек в чате: АН-2.
Гаденыш (19:06): Привет!!!
Оттяжник (19:06): юр, я ЗОЛОТО, баг - навоз
ХАЧИК (19:06): DEKTA-1 ПИВВЕТ
Гаденыш (19:06): Привет!!!
Mayson (19:06): Оттяг!!!!:))))
(:]shaitan[:) (19:06): Гаденыш привет
BUBU (19:06): *Гади, привет*
Новый человек в чате: KIKI.
Vlasta (19:07): Ludi!
Гаденыш (19:07): Алексей, а я и ушла!!!
BUBU (19:07): *Анди :)*
леди (19:07): они не разрешают мне писать адрес
Mayson (19:07): Гади:))))
Yri (19:07): Мэйс....аналогично :))))))))
Гаденыш (19:07): Юр:))))
Оттяжник (19:07): Мэйс, тебя Манана с собаками ищет, скажи рабочий телефон
DELTA-1(man) (19:07): Хачик, привет! Только не Декта, а Дельта.
ЯНЕЖ (19:07): *Кики:)) ПРИВЕТ!!*
леди (19:07): может побьем их
АН-2 (19:07): *BUBU Привет!*
(:]shaitan[:) (19:07): Vlasta привет говорю....
Гаденыш (19:07): Шайтан, Бубу, привет
BUBU (19:07): *Власта, чего хотим?*
ХАЧИК (19:07): ДЕЛЬТА ПРИВЕТ!
Гаденыш (19:07): Мэйс. привет
Mayson (19:07): Оттяг, а она не знает что ли? (((и ты не знаешь??? (((
Aleksey (19:07): Гаденыш т.е. из другого места :) и сразу за комп :)))
DELTA-1(man) (19:08): Как жизнь молодая?
Yri (19:08): Гадик....може надумаешь.....
ЯНЕЖ (19:08): *Власта я люди;)*
леди (19:08): а потом развлечемся
Mayson (19:08): Оттяг - 222-651
Гаденыш (19:08): Алексей. ага, в гостях, пока меня не развлекают, я сама решила
Новый человек в чате: Курт.
BUBU (19:08): *Ну вот пойду я :))*
Гаденыш (19:08): Алексей. ага, в гостях, пока меня не развлекают, я сама решила
Новый человек в чате: ОПС.
Новый человек в чате: Svetik.
ХАЧИК (19:08): ВСЕ ОКЕЙ, ТОКА КУШАТЬ ОХОТА.....
Оттяжник (19:08): Мэйс, я потерял, а она меня спрашивала сегодня
Vlasta (19:08): Шайтан не хочешь написать мне сочинение? :)))
ОПС (19:08): АН
леди (19:08): рабббб
DELTA-1(man) (19:08): леди, кого бит собрались?
Гаденыш (19:08): Юр. я уже в гостях!!!
Svetik (19:08): Всем здраствуйте
Чеширский_ (19:08):
(:]shaitan[:) (19:08): Vlasta какая тема???
Курт (19:08): *леди давай ....а как ты развлекаешься?*
DELTA-1(man) (19:08): Не кормят шоли?
ЯНЕЖ (19:08): *ОПС Svetik ПРИВЕТ*

Продолжение рис. П1.1

Mayson (19:09): Оттяг, записал?
DELTA-1(man) (19:09): **Светик, приветик!**
Дурень (19:09): **....**
Madonna (19:09): **Леди ну ты и развратная!**
Svetik (19:09): **ЯНЕЖ хай**
Yri (19:09): Оттяг...так она с тобой , что ли искала Мэйса ? :))))))))))))))
Новый человек в чате: ARAHNID.
ОПС (19:09): **DELTA Я В ЧАТЕ**
Vlasta (19:09): **Шайтан Чо то про Пастернака!**
ХАЧИК (19:09): **ДА ЕЩЕ НЕ ПРИГОТОВИЛСЯ УЖИН А ПАХНЕТ ПРОСТО СУПЕРРРРРРРРР**
Чеширский_ (19:09): **Светик. Добрый вечер**
Svetik (19:09): **ДЕЛЬТА И ТЕБЕ привет**

Окончание рис. П1.1

П2. Списки для вычисления формально-семантических параметров

П2.1. Общенаучная лексика (37)

Слово	f_1 *)	f_2 **)	f_1/f_2
1. уравнение	3,01E-03	2,84E-06	1060,784
2. функция	3,72E-03	2,56E-05	145,644
3. плоскость	1,87E-03	2,27E-05	82,090
4. образовывать	8,28E-04	1,14E-05	72,846
5. формула	1,25E-03	2,08E-05	60,022
6. коэффициент	7,73E-04	1,51E-05	51,003
7. определение	1,27E-03	2,56E-05	49,591
8. преобразование	1,03E-03	2,27E-05	45,116
9. фаза	4,33E-04	1,04E-05	41,535
10. ось	1,03E-03	3,22E-05	31,924
11. измерение	5,38E-04	1,70E-05	31,545
12. величина	2,11E-03	6,72E-05	31,394
13. множество	2,41E-03	7,76E-05	31,047
14. ток	1,24E-03	4,54E-05	27,290
15. рассмотреть	7,85E-04	3,12E-05	25,129
16. единица	7,88E-04	3,41E-05	23,108
17. организм	1,51E-03	6,63E-05	22,825
18. определяться	7,15E-04	3,50E-05	20,414
19. раствор	1,54E-03	7,67E-05	20,117
20. элемент	2,73E-03	1,44E-04	18,991
21. число	4,50E-03	2,38E-04	18,929
22. определить	1,82E-03	9,66E-05	18,797
23. точка	3,69E-03	2,09E-04	17,638
24. кислота	1,09E-03	6,44E-05	16,933
25. ядро	7,95E-04	4,83E-05	16,467
26. сумма	9,55E-04	5,87E-05	16,272
27. изменение	1,64E-03	1,06E-04	15,445
28. система	3,77E-03	2,74E-04	13,771

Слово	f_1 *)	f_2 **)	f_1/f_2
29. молекула	9,30E-04	7,38E-05	12,595
30. значение	2,81E-03	2,24E-04	12,503
31. прямой	1,66E-03	1,34E-04	12,368
32. относительно	5,75E-04	4,83E-05	11,910
33. определять	7,95E-04	6,72E-05	11,829
34. выражение	9,08E-04	8,05E-05	11,278
35. данный	1,70E-03	1,52E-04	11,154
36. таблица	4,95E-04	4,64E-05	10,672
37. доказательство	4,00E-04	3,98E-05	10,061

*)Частота слова в научных текстах по *Частотному словарю общенаучной лексики*. – М., 1970.
**)Частота слова «во всех текстах» по *Частотному словарю русского языка*. – М., 1977.

П2.2. Названия официальных документов (43)

акт
декларация
директива
договор
закон
заявление
извещение
инструкция
классификатор
кодекс
комментарий
конвенция
конституция
контракт
обращение
определение
пакт
перечень
письмо
положение
порядок
послание
постановление
правила
прейскурант
приказ
программа
проект
протокол
разъяснение
распоряжение
регламент
рекомендации
решение
соглашение
сообщение
справка
требования
уведомление
указ
указания
устав
хартия

П2.3. Научные корнеаффиксы приставочного типа (70)

агро
аква
актино
анти
антропо
архи
астро
ауто
бакт
бета
био
гальвано

гекса
гекто
гелио
гемато
геми
гетеро
гигро
гипер
гипо
гисто
гомео
дендро
дермато
деци
дифф
зоо
интро
ихтио
квази
крио
крипто
лакто
лексико
лингво
макро
мега
мезо
метео
микро
невро
нейро
нитро
орнито
орто
орфо
офтальмо
палео
пневмо
псевдо
сапро
сейсмо
термо
ультра
физио
фито
хризо
хроно
цикло
цито
экви
экзо
экстра
экто
эндо
энтеро
энто
эпид
этно

П3. Диаграммы «ящики с усами» для параметров первичного набора

Точка – среднее значение, «ящик» – стандартная ошибка, «усы» – стандартное отклонение.

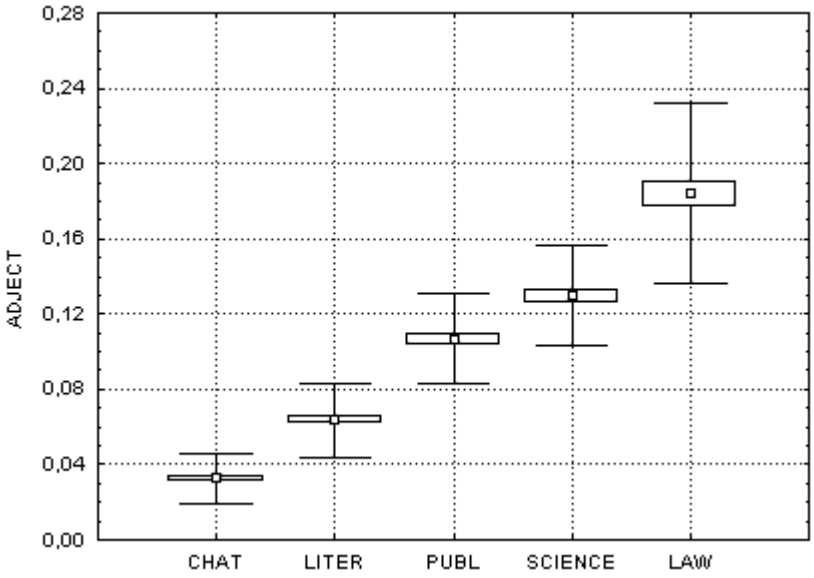

Рис. П3.1. Параметр ADJECT

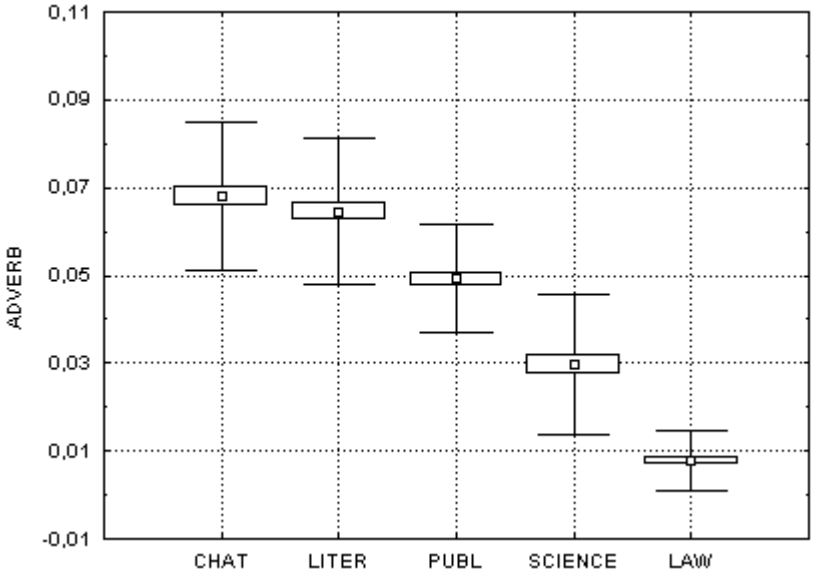

Рис. П3.2. Параметр ADVERB

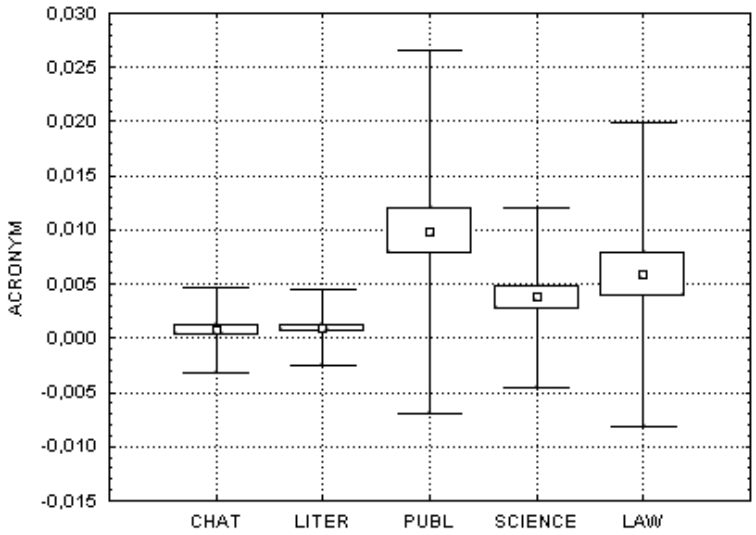

Рис. П3.3. Параметр ACRONYM

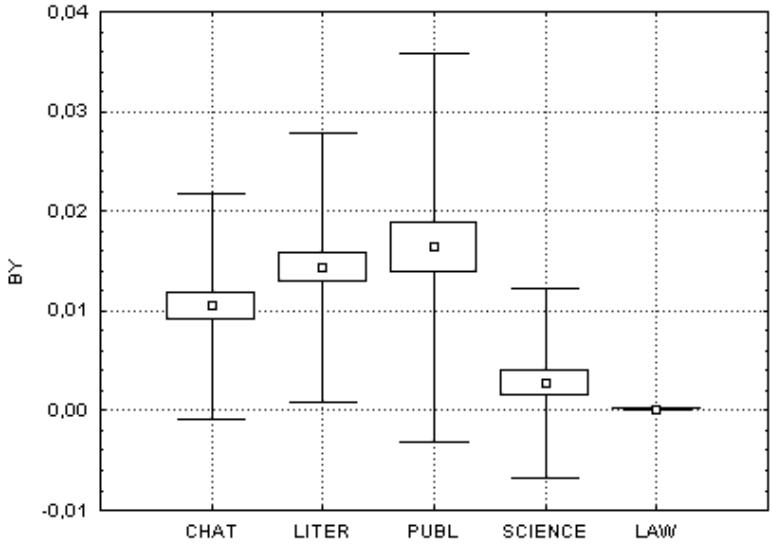

Рис. П3.4. Параметр BY

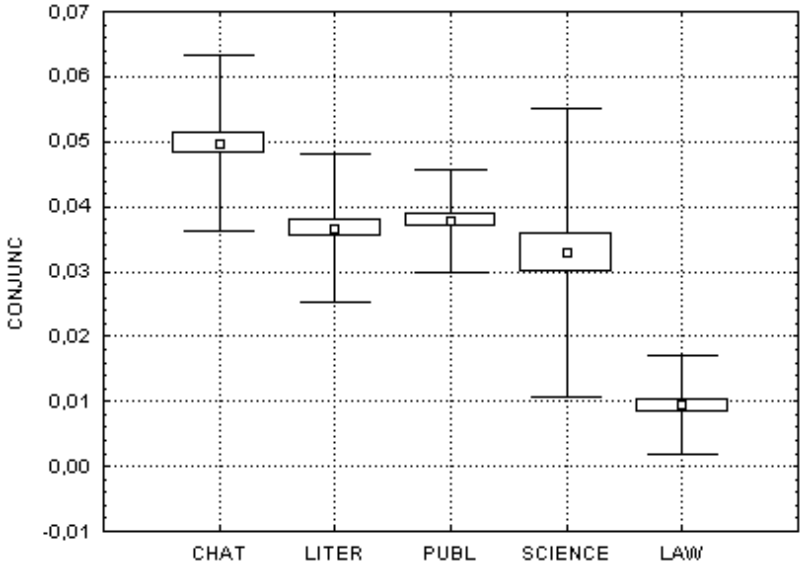

Рис. П3.5. Параметр CONJUNCT

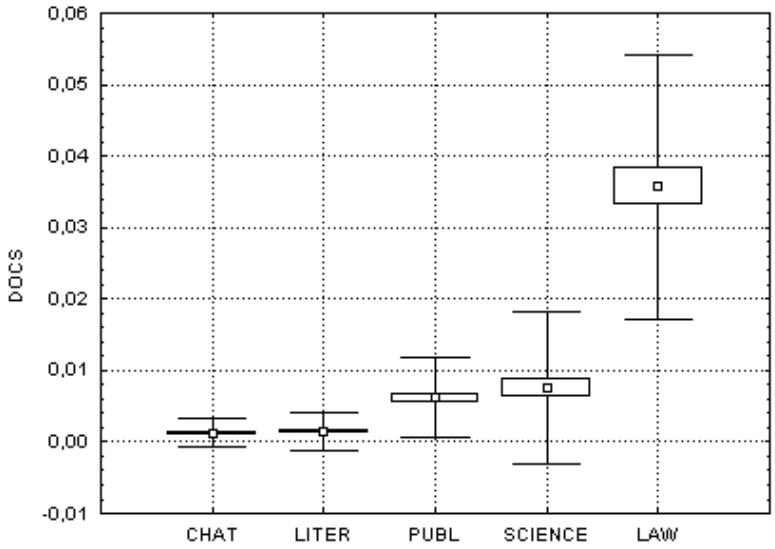

Рис. П3.6. Параметр DOCS

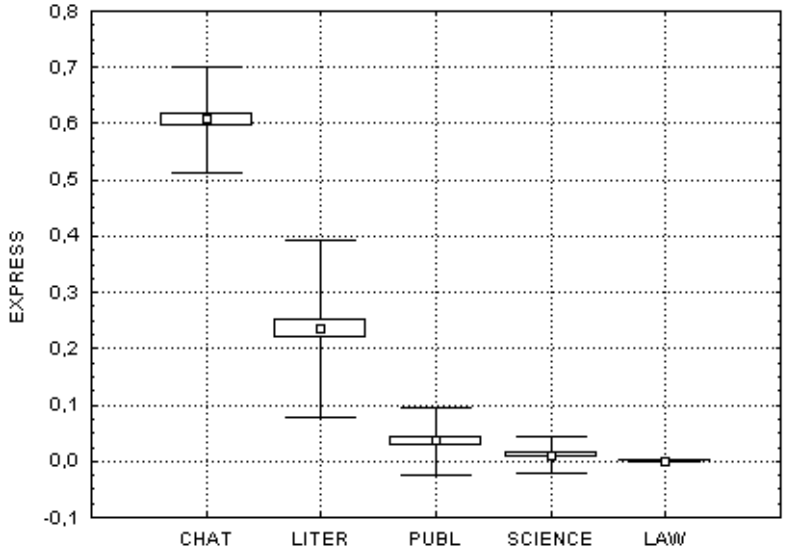

Рис. П3.7. Параметр EXPRESS

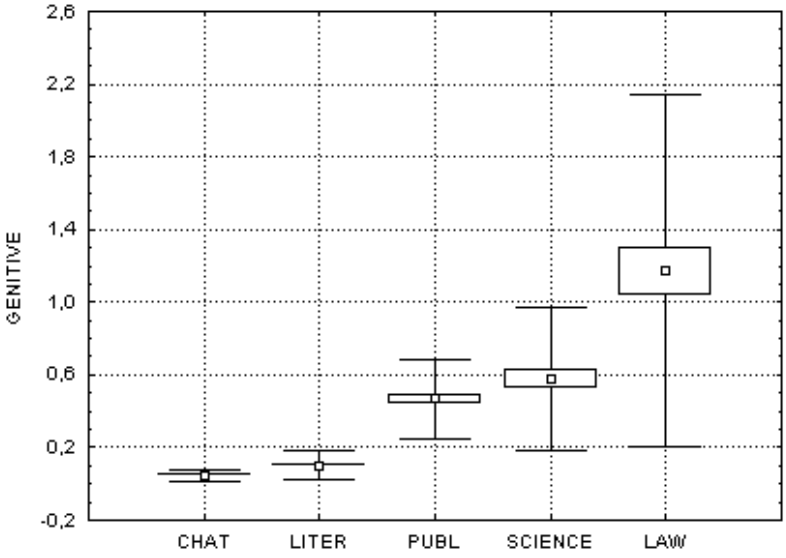

Рис. П3.8. Параметр GENITIVE

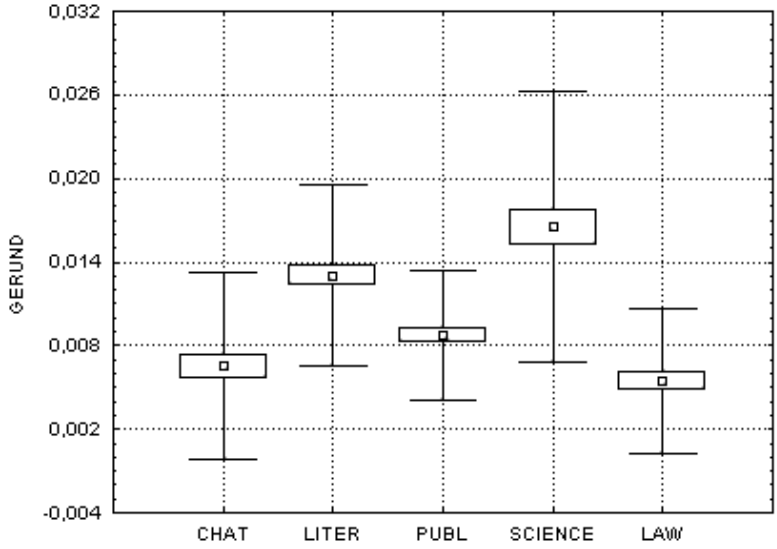

Рис. П3.9. Параметр GERUND

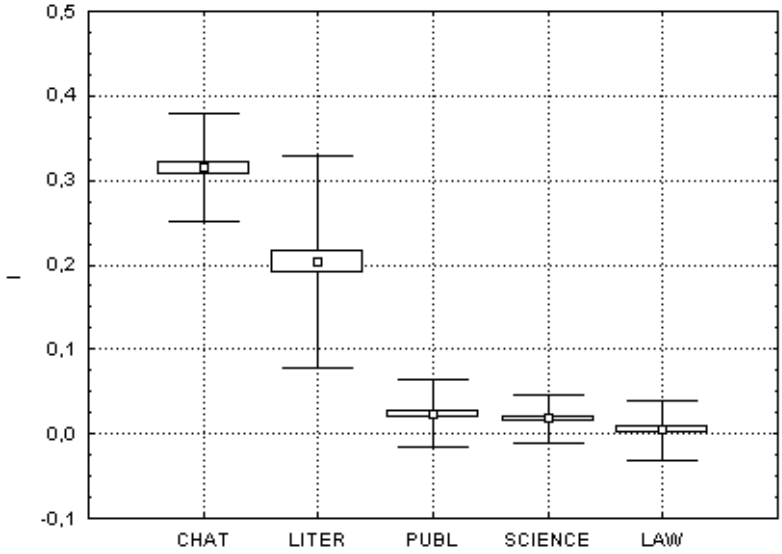

Рис. П3.10. Параметр I

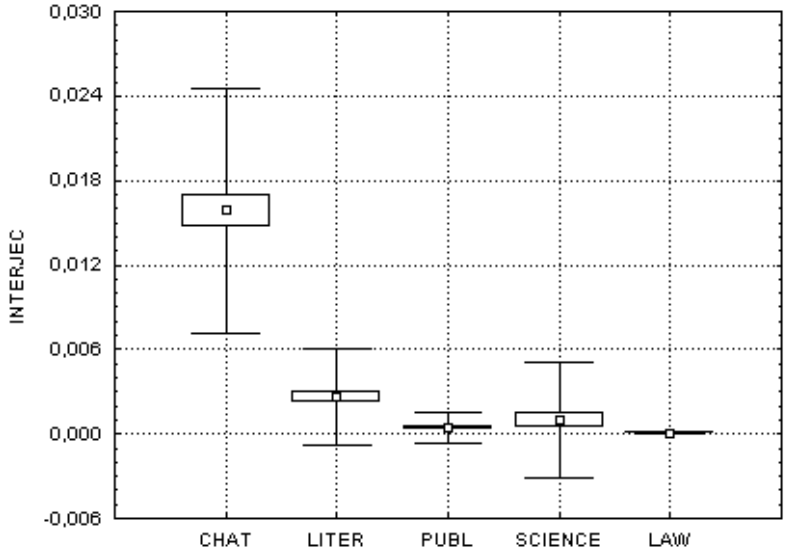

Рис. П3.11. Параметр INTERJEC

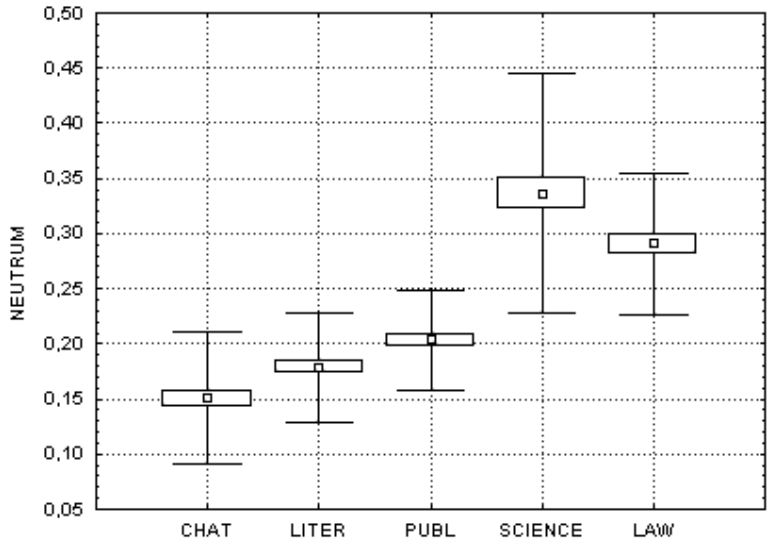

Рис. П3.12. Параметр NEUTRUM

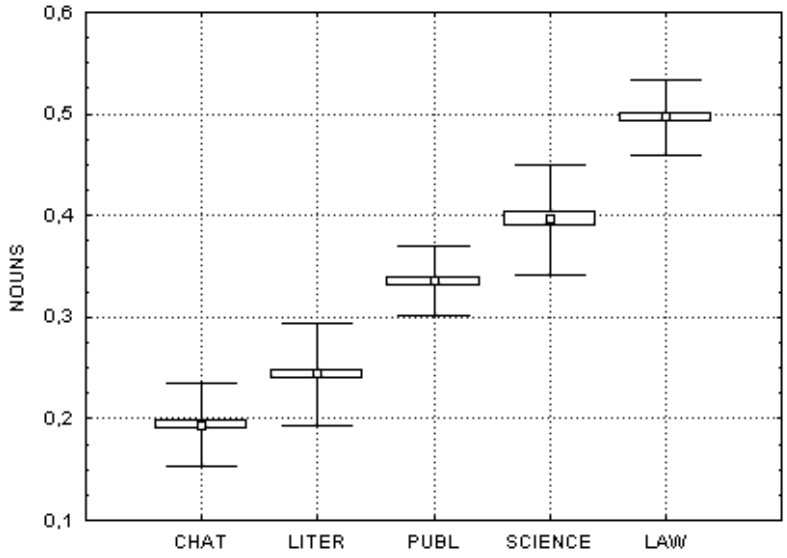

Рис. П3.13. Параметр NOUNS

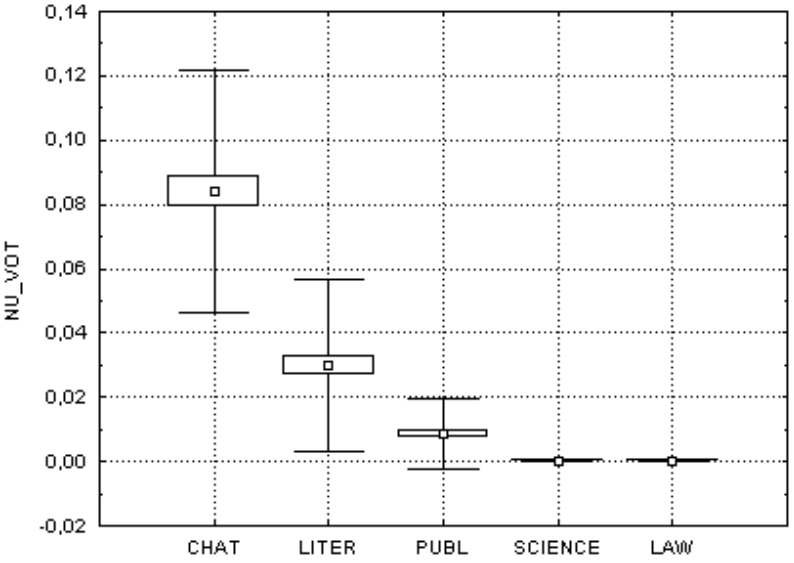

Рис. П3.14. Параметр NU_VOT

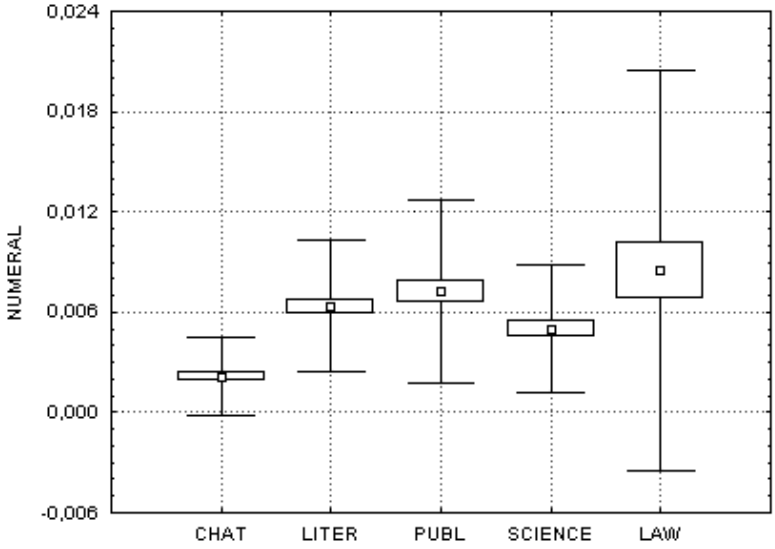

Рис. П3.15. Параметр NUMERAL

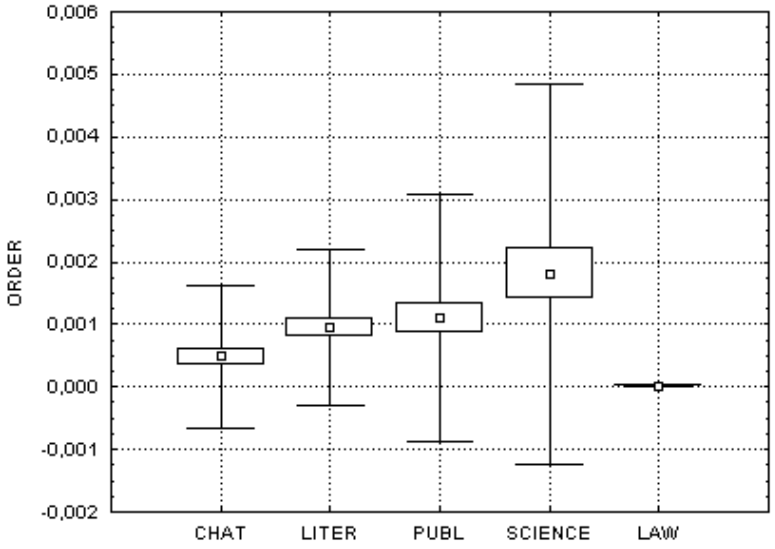

Рис. П3.16. Параметр ORDER

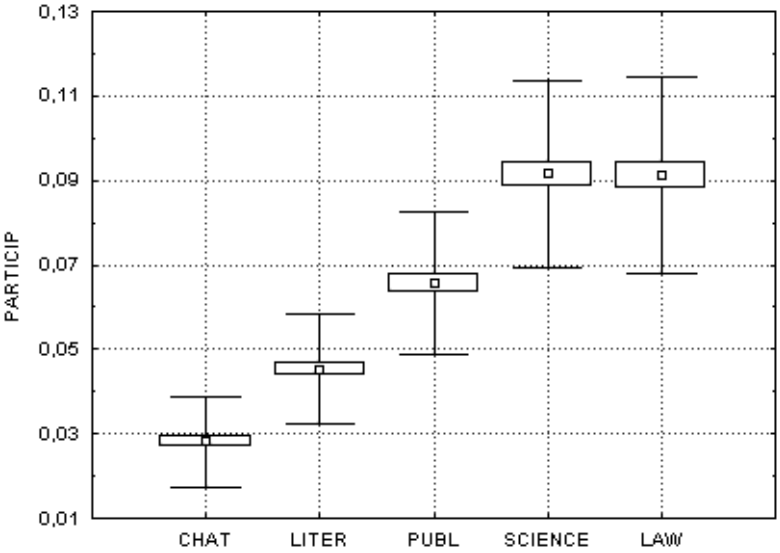

Рис. П3.17. Параметр PARTICIP

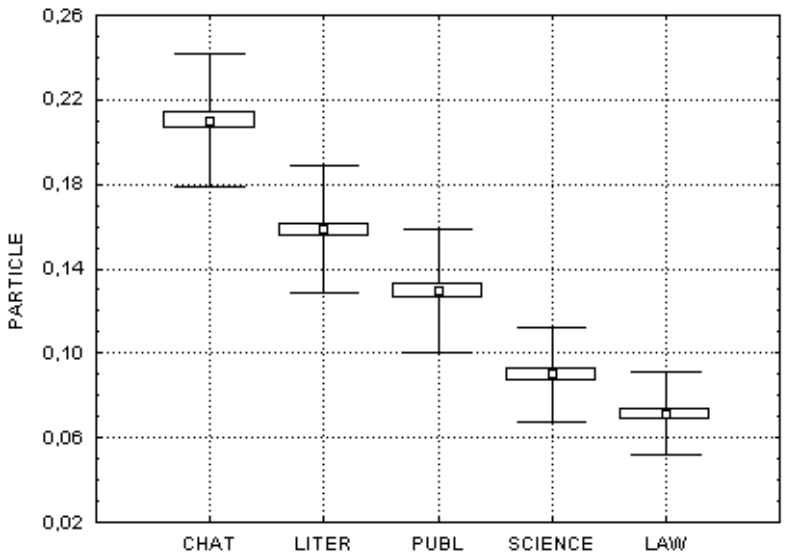

Рис. П3.18. Параметр PARTICLE

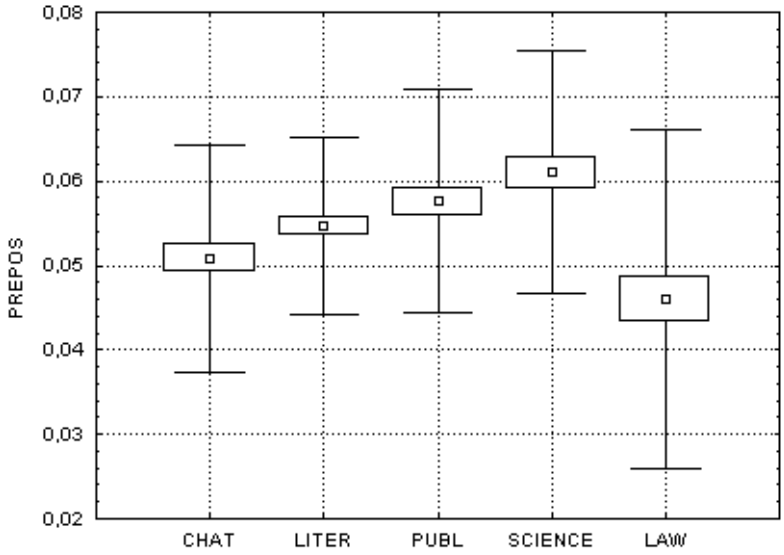

Рис. П3.19. Параметр PREPOS

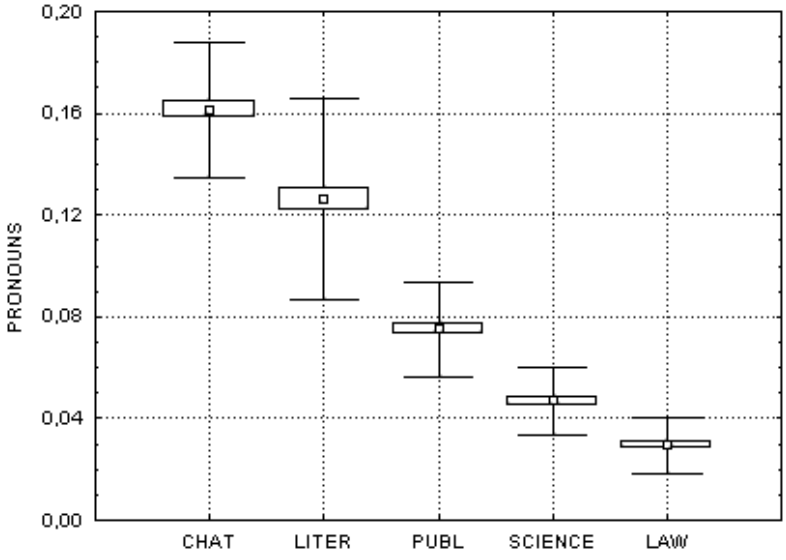

Рис. П3.20. Параметр PRONOUNS

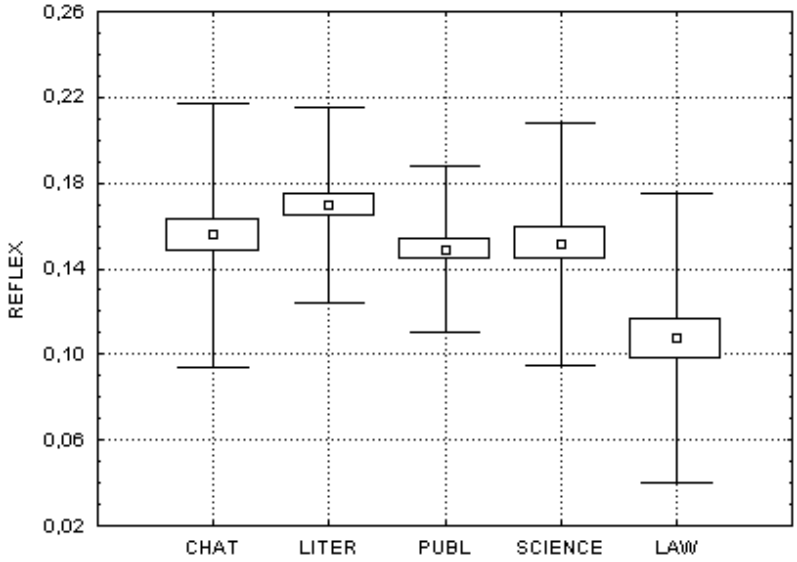

Рис. П3.21. Параметр REFLEX

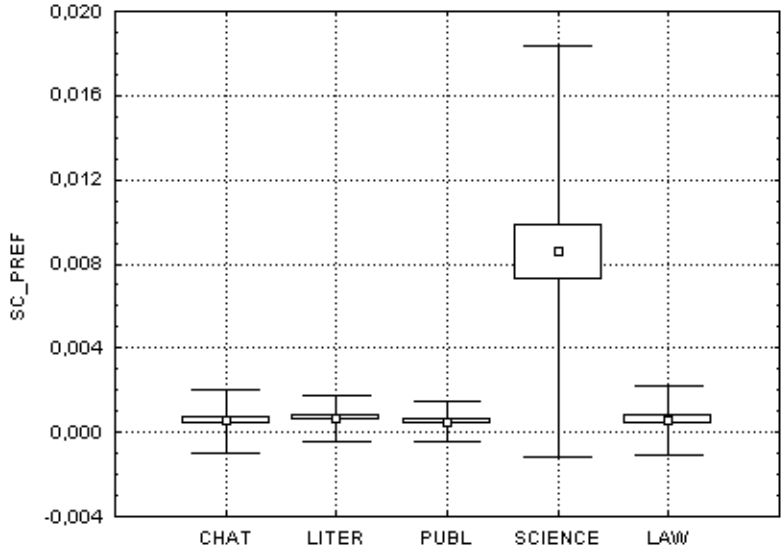

Рис. П3.22. Параметр SC_PREF

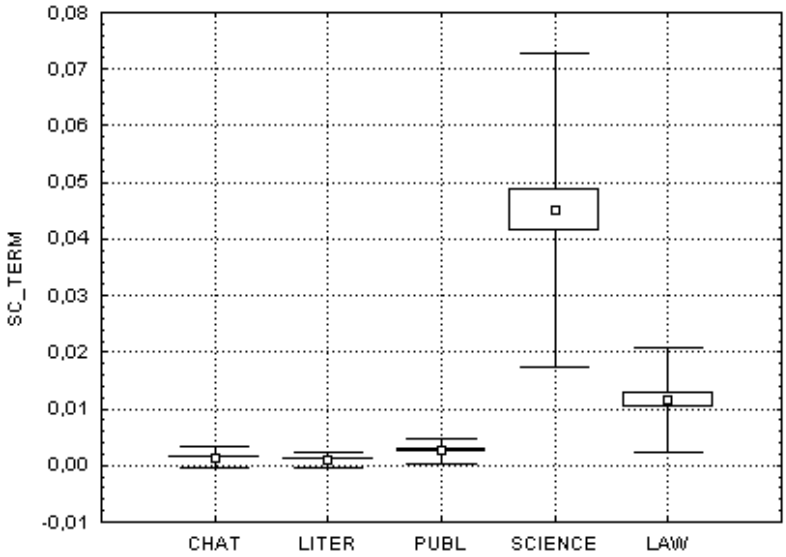

Рис. П3.23. Параметр SC_TERM

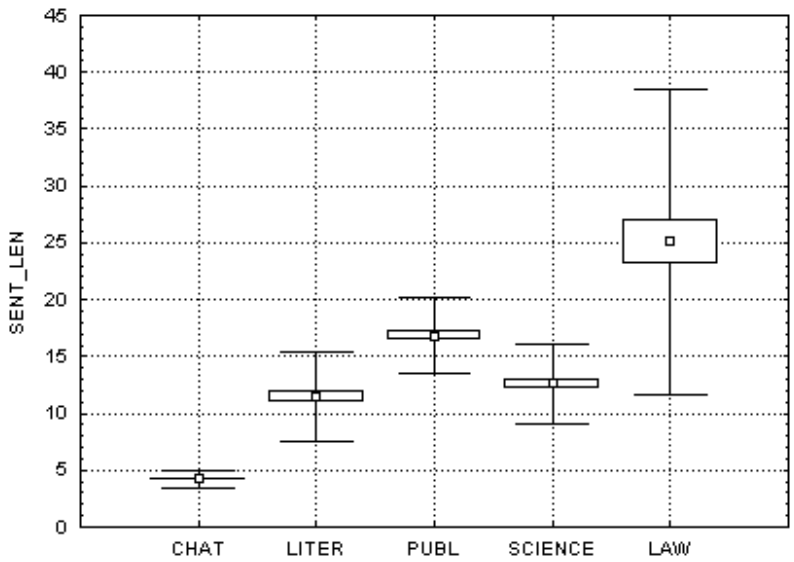

Рис. П3.24. Параметр SENT_LEN

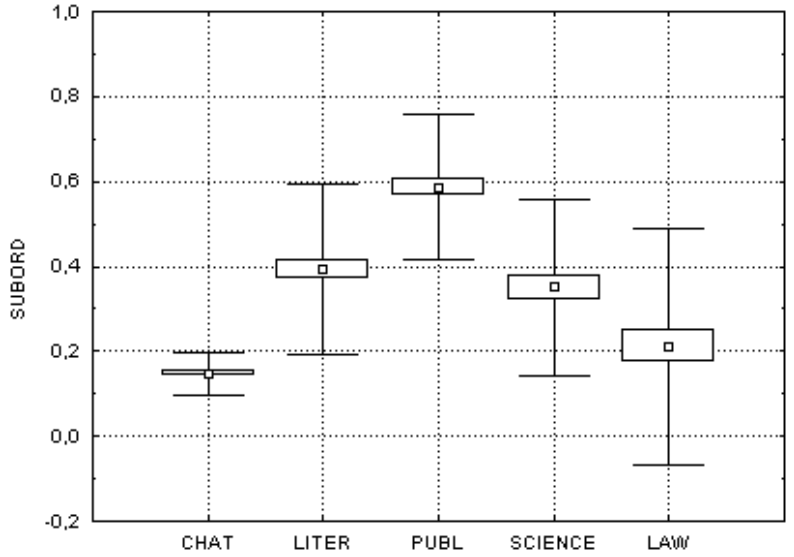

Рис. П3.25. Параметр SUBORD

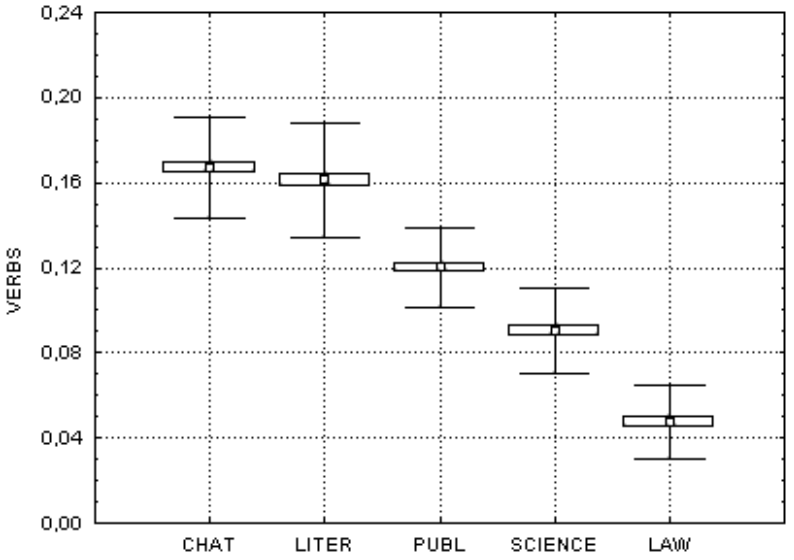

Рис. П3.26. Параметр VERBS

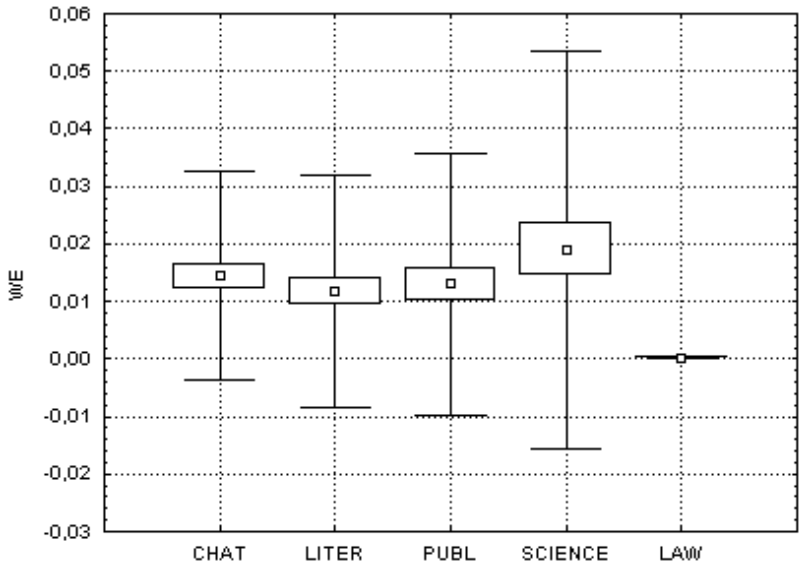

Рис. П3.27. Параметр WE

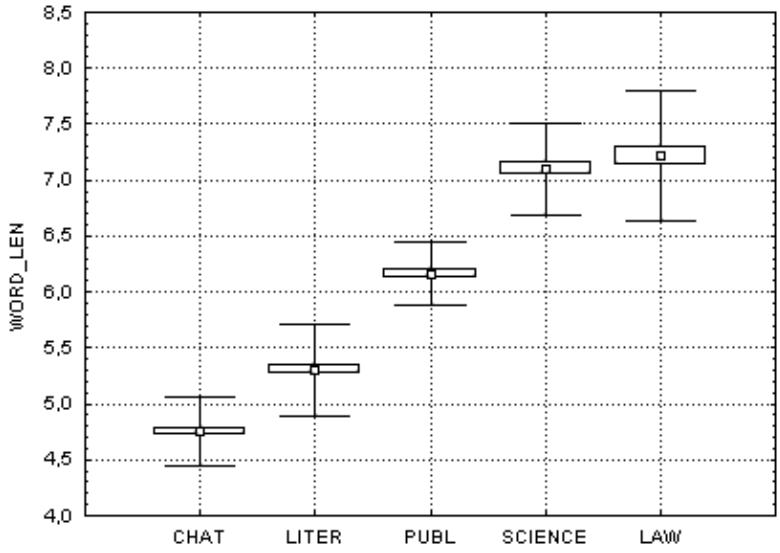

Рис. П3.28. Параметр WORLD_LEN

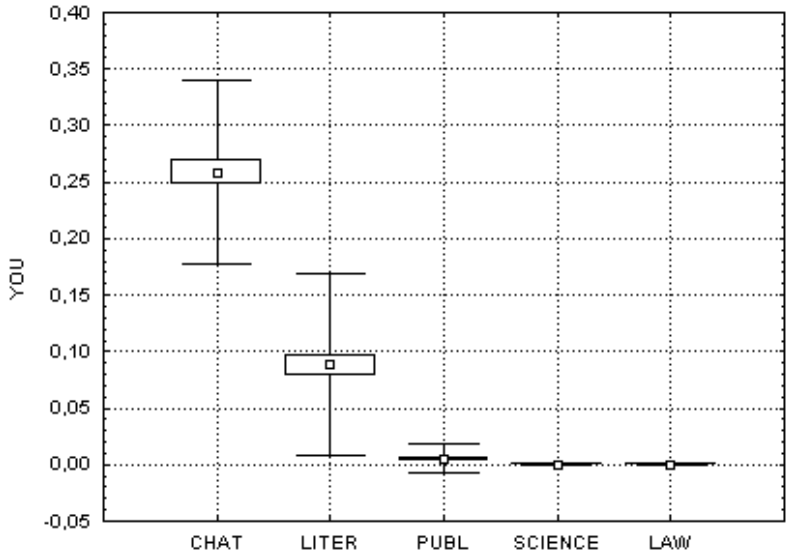

Рис. П3.29. Параметр YOU

П4. Тестовый массив документов

Документ	Источник
1. Доктрина. Архив	http://cross.ru/cgi-bin/club.pl?arc-doct-8421
2. Сан-Антонио Причесывая жирафу	http://alpha.vision.krg.kz/books/readit.htm?102
3. Люди прошлого и социализм	http://lib.cango.net.kg/ecology/ideol/RADIK003.HTM
4. Черный Л., Павлов Н. МАНИФЕСТ – 93	http://www.mi.ru/~anarch/man93.html
5. Троцкий Л. Решающий этап	http://www.1917.com/ru/Marxism/BO_No_51/xx0475.html
6. Лукьянов А. На пороге третьего тысячелетия	http://www.nasledie.ru/pressa/obozrev/N05_99/5_04.HTM
7. Павлов А., Александров Б. Деньги, пахнущие подлостью	http://www.rg.ru/Anons/arc_1999/1013/333.htm
8. Журнал "Спортивная Жизнь России", раздел «Наши консультации»	http://www.infosport.ru/press/szr/0897/topic30.htm
9. Капитонов К. Ближний восток: в ожидании омоложения и перемен	http://www.vestnik.com/issues/2000/0718/koi/kapitonov.htm
10. Вишневский Б. "Железный дровосек", не получивший сердца...	http://forum.msk.ru/files/990131091238.html
11. Московская прокуратура признала законным решение московского правительства о запрете проведения съезда РНЕ	http://www.internews.ru/~rbn/6337F2.html
12. Калашников В. Кто там шагает правой?	http://vedomosti.spb.ru/arts/spbved-2101-art-40.html
13. Вестник «Сербский крест», – 1999. – №12 (сентябрь).	http://www.rusk.ru/Press/Serb_krest/99/skrest_12.htm
14. Левашов В. Двадцать лет спустя	http://www.sccamoscow.ru/scca/exhibitions/15m2/archive/1/1_.htm
15. Верховский А. Итоги парламентских выборов 1999 года для радикальных национал-патриотов	http://www.panorama.ru:8100/works/patr/bp/4rus.html

Документ	Источник
16. ЕС показывает зубы	http://www.infonet.ee/~dd/00/Feb/04/05-1.htm
17. Шерман С. Совет Международных Отношений (CFR)	http://gazeta.ru/dossier/15-07-1999_cfr.htm
18. Кузнецов А. Предвыборный марафон под знаком "зимней войны"	http://www.presa.spb.ru/newspapers/nevrem/arts/nevrem-1927-art-16.html
19. Щекатихин Е. "Безобидный" сионизм	http://www.rus-sky.com/history/library/articles/shekatihin.htm
20. Медведев А. Не умножать зла	http://www.cfc.spb.ru/vxv/miscMedvedev.htm
21. Казеннов С., Кумачев В. Либералы: “большие маневры” в преддверии больших перемен	http://www.ji-magazine.lviv.ua/n14texts/kazenn.htm
22. Мелле Шт. Правые радикалы в немецкой армии	http://www.galaxy.orc.ru/igel/n8/germany.html
23. Антимилитарист Он-Лайн (Информационный бюллетень) – 1999. – №18 (июнь).	http://www.ara.ru/amonline/aol018r.html
24. Печкин С.М. Начальный курс квенья	http://ivbespalov.chat.ru/tolkien/pqbc.htm
25. Любой конфликт заканчивается переговорами. Беседа Максима Шевченко с председателем Совета муфтиев России шейхом Равилем Гайнутдином	http://www.ng.ru/cis/1999-09-02/konflict.html
26. Норта Д. Марксизм и профсоюзы	http://www.wsws.org/ru/1999/jul1999/tu-j08.shtml
27. Капитонов К. Ближний Восток: в ожидании омоложения и перемен	http://www.interface.co.il/politika/middle_east.html
28. Еськов К. Евангелие от Афрания	http://lib.ru/PROZA/ESKOV_K/afranij.txt
29. Коган-Ясный В.В. Политический аспект отношений федеральных органов власти Российской Федерации с Чеченской республикой в 1990–1994 гг.	http://www.memo.ru/hr/hotpoints/CHECHEN/ITOGI/kogan.htm
30. Березовский В., Червяков В. Хасбулатов и	http://www.ancentr.ru/texts/has-ruc.htm

Документ	Источник
Руцкой: портрет политического альянса	
31. Ипполитов К.Х. Время выбора решений	http://www.fact.ru/archiv/num03/2ippolitov.htm
32. Сиван Э. Радикальный ислам: причины и последствия террористического насилия	http://www.users.datacom.ru/mak/analitics9.html
33. Малашенко А. Исламский фактор в Прикаспийском регионе	http://www.transcaspian.ru/cgi-bin/webdebug.exe/rus/1091.html
34. Группы и фракции - Аграрная партия	http://notes.society.ru/bibl/partii/apr/grupp.html
35. Нодия Г. Спорные Границы на Кавказе	http://www.zakavkazye.org/georgia/policy/nodia.htm
36. Малашенко А. Исламский фактор в прикаспийском регионе	http://www.ferghana.ru/news03/074.html
37. Прейгер Д., Телушкин Дж. Глава из книги "Почему евреи?"	http://www.ir.spb.ru/potshemu14.htm
38. России после кризиса:новая расстановка сил, новые сценарии политической борьбы	http://obladm.nso.ru/oblsovet/docs/INFOTDEL/novosty/Postkriz.html
39. Шляпентох В. От любви к ненависти	http://www.lebed.com/art593.htm
40. Акишев А. Политическая элита Казахстана: история, современность, перспективы	http://eurasia.org.ru/2000/analitica/06_27_az26.htm
41. Второе полугодие 1994 г. (господствующие тенденции развития политической ситуации)	http://ns.hist.msu.ru/NIISS/Misc/2pol94.htm
42. Мухин Ю.И. Товарищ брянский волк	http://www.duel.ru/sp.jhtml?n=200007&p=1&s=07_1_2.htm&u=1
43. Фадеев Ю. Россия требует к себе важительного подхода, как к евроазиатской исламско-христианской стране (ответ на вопросы "Русского Журнала").	http://astrachan.newmail.ru/pavlovski9.htm
44. Мельников А. Егор, Вы не совсем правы	http://www.yabloko.ru/alt/Forums/Archive/146.html
45. Лейбов Р. «Бессрочная ссылка» от 20.04.98.	http://www.russ.ru/ssylka/98-04-20.htm

Документ	Источник
46. Абакумов А. В. ХОЧУ ВОЗРАЗИТЬ!? Сергею Васильцову - историку, Сергею Обухову – экономисту	http://abakum.newmail.ru/publ/text3.html
47. Экспериментальные методы изучения радикальных реакций	http://chemgu.icp.ac.ru/elect/Chap17/Chap17.htm
48. Балаболкин М.И., Клебанова Е.М., Креминская В.М. Патогенез ангиопатий при сахарном диабете	http://www.clinpharma.com/magazine/diabet/number2/art1.htm
49. АЗАКТАМ (описание лекарственного препарата)	http://www.mednet.odessa.ua/med/pharm/l/a/a99.htm
50. Тарабанько В. Е. и др. Исследование кислотности радикальных интермедиатов окисления лигнина квантово-химическими методами	http://nioch.sibstn.nsk.su/mirrors/press/chemwood/volume2/n3/stat_14.html
51. Переходцев Г.Д., Дубинский А.А., Лебедев Я.С. Ферромагнитный обмен и аномальные знаки D в радикальных парах	http://center.chph.ras.ru/~perekhod/art11r.html
52. Борщевская М.И., Васильева С.М. Развитие представлений о биохимии и фармакологии меланиновых пигментов	http://www.medi.ru/doc/8890102.htm
53. Фролков Г.Д. и др. Влияние природной механоактивации на выбросоопасность углей	http://vostnii.shakht.donpac.ru/outbursts_ru.htm
54. Воздействие радиации на ткани живого организма	http://study.npi.msu.su/zachet/handbook/gl_4/gl_4.htm
55. Романченко А. Удивительный гормон – мелатонин	http://www.oncos.ru/3-1999/21.html
56. Гудзь О.В. Итоги и перспективы клинического применения дезинфекционных средств из группы четвертичных аммониевых соединений	http://provisor.kharkov.ua/archive/N12/klin_dez.htm
57. Письменский А. В. Влияние условий окисления и структуры ароматических диаминов	http://ecolan.com.ru/alexis/autoref/index.html

Документ	Источник
на механизм и эффективность их ингибирующего действия»	
58. Автономный электронный стимулятор желудочно-кишечного тракта и слизистых оболочек "Электронный нормализатор"	http://ecomed.webzone.ru/monogr.htm
59. Строение гидроксильной группы	http://www.uic.ssu.samara.ru/~chemistry/chem4/o23.htm
60. Левицкий Е.Л. Коррекция поражений ядерного генома антиоксидантами в условиях токсического повреждения печени	http://medved.kiev.ua/arhiv_mg/stat_98/98_2_12.htm
61. Соловьев М. В. Виртуальное антистарение	http://www.kuban.su/medicine/shtm/baza/geront/baza/113rus.htm
62. Вартанян Л.С., Гуревич С.М. Влияние ионола на метаболизм супероксидных радикалов в печени мышей	http://www.ibmh.msk.su/pbmc/4504r/6r.htm
63. Словецкий Д.И. Механизмы плазмохимического травления материалов	http://www.indi.ru/private/plasma99/LECTIONS/Slov_lect.html
64. Ким Сун-чжон, Пиков Г.Г. Жан Кальвин и некоторые проблемы швейцарской Реформации (Материалы к спецкурсу)	http://lib.ru/HISTORY/CALVIN/calvin.txt
65. Чудинов А.В. Размышления англичан о Французской революции: Э.Берк, Дж.Макинтош, У.Годвин	http://liberte.newmail.ru/Tchoud_referat.html
66. Русское население республик Северного Кавказа	http://www.eawarn.ras.ru/centr/eawarn/doklad/ch10.htm
67. Никольский С.А. "Аграрная реформа" 1991 - 1995 годов и проблема модернизации российской деревни	http://fadr.msu.ru/archives/mailing-list/priv-agr/art-rus/msg00020.html
68. Калашников З.М. Политические и правовые идеи в странах Запад-ной Европы и Америки в новое и новейшее время	http://www.h-c.dp.ua/6q.htm
69. Вуколов С. Кислоты // Брокгауз Ф.А., Ефрон И.А. Энциклопедический словарь	http://www.booksite.ru/fulltext/bro/kga/brokefr/2/2123.htm

Документ	Источник
70. Мелатонин	http://melaxen.ru/Melatonin/Melatonin4.htm
71. “MIRACLE 7”- лучший антиоксидантный состав	http://www.organon.ru/~urgana/totalhealth/miracle1.htm

П5. Тезаурус по компьютерной лингвистике

П5.1. Словник

Первичными источниками для составления словника тезауруса послужили следующие издания:

1. Моделирование языковой деятельности в интеллектуальных системах /М.Б.Бергельсон, О.Ю.Богуславская, Р.И.Гафт и др.; Под ред. А.Е.Кибрика и А.С. Нариньяни. – М.: Наука, 1987. – 280 с. – (Проблемы искусственного интеллекта.)
2. Перспективы развития вычислительной техники: В 11 кн. Кн. 2. Интеллектуализация ЭВМ/Е.С.Кузин, А.И.Ройтман, И.Б.Фоминых, Г.К.Хахалин. – М .: Высш. шк., 1989. – 158 с.
3. Искусственный интеллект: В 3 кн. Кн. 1. Системы общения и экспертные системы: Справочник /Под ред. Э.В.Попова. – М.: Радио и связь, 1990. – 460 с.
4. Искусственный интеллект: В 3 кн. Кн. 2. Модели и методы: Справочник /Под ред. Д.А.Поспелова. – М.: Радио и связь, 1990. – 303 с.

Вторичными источниками для уточнения определений и пополнения словника стали словари:

1. Толковый словарь по искусственному интеллекту/А.Н. Аверкин, М.Г. Гаазе-Рапопорт, Д.А.Поспелов. – М.: Радио и связь, 1992. – 256 с.
2. Лингвистический энциклопедический словарь /Гл. ред. В.Н.Ярцева. – М.: Сов. энцикл., 1990. – 682 с.
3. Борковский А.Б. Англо-русский словарь по программированию и информатике (с толкованиями). – 2-е изд., стер. – М.: Рус. яз., 1989. – 332 с.

На первом этапе было выделено примерно 500 терминов. Окончательный объем словника тезауруса – около 250 терминов.

П5.2. Набор семантических отношений

Для тезауруса по компьютерной лингвистике за исходный был принят набор семантических отношений из книги С.Е.Никитиной[1]. В процессе работы над тезаурусом были исключены некоторые отношения (“вариант – инвариант”, “уровень языка – единица уровня”) и введены новые (“версия – исходное”). Некоторые отношения были уточнены (“операция” –> “действие/процесс/операция”, “импликация” –> “связь процессов”). Окончательный набор отношений приведен в табл. П5.1.

Таблица П5.1

Набор отношений тезауруса по компьютерной лингвистике

№	Отношение	Обратное	Описание
1	Род	2	Родовое понятие
2	Вид	1	Видовое понятие
3	Часть	4	Часть, элемент, компонент
4	Целое	3	
5	Свойство	6	
6	носитель свойства	5	
7	сфера термина	8	Дисциплина/школа/наука, к которой относится термин
8	термин сферы	7	Термины, принадлежащие данной сфере знания
9	Функция	10	
10	способ выражения функции	11	
11	действие/ процесс/ операция	12, 13, 14, 15	Процесс =<Субъект, Объект, Результат, Способ>
12	субъект действия/ операции	11	
13	объект действия/ операции	11	

[1] Никитина С.Е. Тезаурус по теоретической и прикладной лингвистике. – М.: Наука, 1978. – 375 с.

Продолжение табл. П5.1

№	Отношение	Обратное	Описание
14	Результат	11	Конечный результат действия/ процесса/ операции
15	инструмент/ способ/ метод	11	
16	языковой объект	17	Объект языка, описываемый языком науки (метаязыком)
17	способ метаязыкового представления	16	Функция отражает специфику языка науки
18	Коррелят	18	Корреляты – термины, относящиеся к одной категории, но противопоставленные по некоторому существенному признаку
19	Синонимия	19	Принудительная синонимия
20	Ассоциация	20	Семантическая близость
21	Исходное	22	Исходное для модификации
22	Версия	21	Модификация, трансформация
23	Выражение	24	Способ нотации, выражения, записи, представления объекта
24	Выражает	23	Описываемое, объект нотации, выражаемое
25	связь процессов	25	Отражает сопряженность процессов: следование во времени, причинную связь и т.д.

П5.3. Примеры установленных связей

На рис. П5.1-П5.3 приведены примеры установленных связей для термина *синтаксический анализ* и двух семантически близких ему терминов: *предложение* и *формальная грамматика*.

	<u>синтаксический анализ</u>
субъект	синтаксический анализатор
исходный объект	предложение
результат	синтаксическая структура предложения
Инструмент	формальная грамматика
связь процессов	морфологический анализ

Рис. П5.1. Термин «синтаксический анализ»

	<u>предложение</u>
род	высказывание
вид	сложное предложение
целое	сообщение
	текст
часть	слово
	словосочетание
свойство	грамматическая структура предложения

Рис. П5.2. Термин «синтаксический анализ»

	<u>формальная грамматика</u>
род	грамматика 1
вид	грамматика зависимостей
	грамматика непосредственных составляющих
	контекстно-зависимая грамматика
	контекстно-свободная грамматика
	матричная грамматика
	сетевая грамматика
действие/процесс	синтаксический анализ
версия	расширенная сеть переходов (РСП)
ассоциация	синтаксические методы

Рис. П5.3. Термин «синтаксический анализ»

П6. Акты внедрений и испытаний результатов работы