Методы классификации и технология Галактика-Zoom

А. В. АНТОНОВ

Корпорация "Галактика", Москва, Россия

Рассматриваются различные методы классификации текстовой информации. Проводится их сравнительный анализ с технологией Zoom, применяемой в промышленной системе Галактика-Zoom. Приводятся примеры классификации текстовой информации, выполненной этой системой.

ВВЕДЕНИЕ

Перед любой поисковой системой, работающей с большими объемами текстов (сегодня доступны десятки и сотни гигабайт) стоят задачи полноты и точности выдачи материалов по запросу пользователя. При этом актуальность решения задачи точного ответа на запрос пользователя сейчас выше проблемы полноты в связи с огромными объемами доступных данных.

Один из подходов, позволяющих решить проблему точности в часто употребляемых пользователями многозначных запросах, — разбиение информации на тематические группы-рубрики. Пользуясь достаточно коротким списком рубрик, под которые попадают все возвращенные документы, пользователь существенно сужает границы поиска.

Все методы классификации используют один и тот же обобщенный алгоритм, который состоит из следующих этапов:

- задание/построение описаний для всех тематических групп-рубрик;
- построение описания рассматриваемого документа;
- вычисление оценок близости между описаниями тематических групп и описанием документа и выбор наиболее близких тематических групп.

Различия же между методами определяются реализацией этих этапов. В данной статье описываются некоторые из основных методов отнесения документов к рубрикам. Многие методы классификации основываются на работах Большакова И. А., Белоногова Г. Г., Гиляревского Р. С., Лахути Д. Г., Чёрного А. И., Нариньяни А. С.

ТЕЗАУРУСЫ

Для классификации текстов с помощью тезаурусов используются синонимические ряды понятий, другие понятийные отношения, различные лексические связи между понятиями в тексте.

Примерами систем, созданных на основе тезаурусов, могут служить: универсальная информационная система "Россия", Excalibur, Auto-Categorizer, Кросслексика. Тезаурусы, применяемые в этих системах, существенно отличаются от тезаурусов, предназначенных для использования при ручном индексировании текстовых документов [1]. Человек-индексатор сначала должен прочитать текст, понять его и затем изложить содержание текста, пользуясь дескрипторами, указанными в информационно-поисковом тезаурусе. Индексатор должен хорошо понимать всю терминологию, использованную в тексте, тогда для описания основной темы текста ему понадобится значительно меньшее количество терминов.

При автоматической обработке текстов человек-посредник между текстом и описанием его содержания в виде дескрипторов отсутствует. Есть только автоматический процесс и тезаурус, который должен содержать как знания, имеющиеся в традиционных информационно-поисковых тезаурусах, так и знания (насколько это возможно), используемые индексатором для определения основной темы текста. Именно поэтому традиционные тезаурусы, разработанные для ручного индексирования, невозможно использовать при автоматическом индексировании.

Машинный тезаурус должен включать в виде отдельных единиц семантически близкие понятия в отличие от тезаурусов для ручного индексирования, где совокупности близких понятий сводятся к одному, наиболее представительному понятию для уменьшения субъективности индексирования. Таким образом, единицы машинного тезауруса должны быть значительно ближе к понятийному аппарату предметной области, чем дескрипторы традиционного тезауруса. Синонимические ряды понятий должны быть богаче, чем совокупности вариантов дескриптора в тезаурусе для ручного индексирования, поскольку синонимы предназначены для автоматического описания различных способов выражения данного понятия в тексте. Ряды синонимов включают не только существительные и именные группы, но и прилагательные, глаголы, глагольные группы. Для снижения лексической многозначности требуется поиск многословных словосочетаний.

Для автоматического разбора необходимо существенно увеличить покрытие реального текста терминами, описанными в тезаурусе. По этой причине объем тезауруса может насчитывать сотни тысяч терминов и несколько сотен тысяч связей. Способ, помогающий пополнять тезаурус словосочетаниями, предложен в системе Кросслексика [2].

ПРОСТЫЕ СИСТЕМЫ

Существуют задачи классификации текстов, где очень важно решение проблемы быстродействия системы. К таким задачам относятся, в частности, спам-фильтры (фильтрация

реклам) [3]. Причем рекламы отличаются исключительным разнообразием — от предложений купить те или иные товары/услуги до обещаний золотых гор за рассылку "писем счастья" или откровенных просьб. Естественно, алгоритмы, которые используются в таких системах, должны быть максимально простыми.

В этих системах используются, в частности, сигнатуры — частотные словарные характеристики образцов писем-спамов для определения схожих документов и отнесения их к категории спамерских. При этом используется морфология для сведения лексем. Такие сигнатуры не чувствительны к некоторым методикам изменения спамерских писем — перестановке отдельных кусков текста, вставке случайных последовательностей букв. Естественно, такая методика годится только как апостериорный фильтр, реагирующий на данное письмо-спам. Актуальность таких действий сильно зависит от того, насколько рано удается перехватить новое письмо.

Кроме этого, используется также взвешенный список терминов, присущих спамерским документам. Такие веса позволяют отнести документ к этой категории по совокупности нескольких терминов из списка. Пример такого списка: "Хотите заработать", "способ зарабатывания денег в Интернете", "способ заработка", "WebMoney".

Естественно, имеется и "черный список" словосочетаний и отдельных слов, употребление которых дает однозначный ответ на причисление данного письма к категории спамерских. Например, такие часто встречающиеся выражения, как "Ваш адрес был взят из открытых источников", "Рассылка была проведена в соответствии с Конституцией РФ".

АВТОМАТИЧЕСКИЕ СИСТЕМЫ

Автоматические системы не содержат тезаурусов и проводят разбор текстов и выделение из них главных тем на самих массивах текстов. Именно к этому классу относится и технология Zoom, на которой мы остановимся подробнее. Но прежде рассмотрим существующие методы классификации.

Метод SOM — Self-Organizing Maps

Этот метод предполагает классификацию документов с использованием самонастраивающейся нейронной сети. В методе используется интегральная значимость — *tf•idf* (term frequency, inverse document frequency). По сути, SOM — это нейронная сеть Кохонена, выполняющая задачу классификации входных данных и обучающаяся без учителя (unsupervised learning), т. е. самонастраивающаяся [4]. К таким системам относятся: Текст-Аналист, NeurOk Semantic Server, SmartLogic

В системе (SmartLogic) упор делается на технике персонификации информации. Одни и те же термины по-разному значимы для разных людей: "Bill & Gates" для одних связан с "Sun", "Java", "Windows", для других — с "Монополия", "Миллиардер". Семантическая сеть здесь называется CMAP (Concept Map). Разница с SOM состоит в том, что для преодоления недостатков подхода (статистический анализ — неизменяемый результат) применяется наложение "общего" CMAP на персональный. Таких персональных CMAP может быть несколько. Составляются они по истории пользователя, в частности по терминам его запросов.

Байесовские алгоритмы разбора

Метод использует байесовские выражения для вычисления значимости терминов, учитывая условную вероятность появления терминов в данном контексте [5], на основании значимости строится "смысловая" сеть.

Ссылочное ранжирование

Еще один метод следует упомянуть, поскольку применяющая его поисковая система Google произвела настоящую революцию в поиске в сети Интернет.

Основными параметрами данного метода для оценки значимости документов являются гипертекстовые ссылки на данный документ. При этом учитывается текст ссылок и значимость документа, в котором сделана эта ссылка. Также могут учитываться и обратные ссылки из данного документа. Этот метод подобен методу определения ценности статьи в научном обществе — в зависимости от количества ссылок на нее. На основании ссылок вычисляется рейтинг популярности документа (сетевой страницы) PageRank, предложенный разработчиками системы Google. Рейтинг учитывает не только количество ссылок на данный документ, но и ценность каждой ссылки. Конечно, данный показатель не имеет решающего значения при расчете рейтинга ресурса, а используется только для корректировки рейтинга, рассчитанного более традиционным способом на основе ключевых слов.
PageRank вычисляется по формуле:

$$PR(A) = (1-d) + d\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)}, \qquad (1)$$

где $PR(X)$ — рейтинг (PageRank) документа X; A — оцениваемый документ; d — коэффициент затухания; n — количество документов, ссылающихся на A; T_i — документы, ссылающиеся на A; $C(X)$ — общее количество ссылок со страницы X.

По сути, $PR(A)$ — вероятность попадания хаотически путешествующего по сети пользователя на страницу A. Величина $d = 0,85$, характеризует вероятность того, что, находясь на странице, участвующей "в оценочной формуле", посетитель перейдет на произвольную страницу в сети Интернет путем набора URL-адреса прямо в соответствующем поле своего браузера.

Как видно из формулы, "рекомендация" со страницы, имеющей высокую "репутацию", обладает большим весом, что позволяет правильно оценивать значимость непопулярных, но качественных сайтов. Этот метод внетекстовой и подходит в основном для Интернета, в котором распространены такие ссылки, по этой причине сравнивать его с Zoom мы не будем. К системам, использующим подобные методы, относятся Google и Yandex.

Привлечение информации из других запросов тех же пользователей

Такую технологию предлагает известная поисковая машина Рамблер [6]. Для формирования связей запросов, схожих с данным запросом, используются протоколы работы сетевого сервера поисковой машины. В протоколах есть информация о времени обращения, адресе пользователя, его уникальном идентификаторе и собственно самом запросе к поисковой машине.

Программа группировки получает на вход тройки вида $(x,\ y,\ f_{xy})$, где x, y — запросы, а f_{xy} — количество пользователей, подавших данную пару запросов (частота совместной встречаемости). Результат работы программы — список

связей (ссылок) для данного запроса. Например, исходный запрос — "отдых на Кипре", список ассоциаций — "отели на Кипре", "погода на Кипре", "апартаменты", "детский отдых на Кипре", "карта Кипра", "лимассол"... Этот метод также внетекстовой и подходит для поиска в Интернете (иначе просто не собрать соответствующую статистику), по этой причине сравнивать его с Zoom мы не будем.

СИСТЕМА ГАЛАКТИКА-ZOOM

Система Галактика-Zoom создана в корпорации "Галактика". В этой системе применяется технология Zoom, дающая пользователю возможность приближения-удаления от поискового образа для более полного обзора: от отдельных документов до информационного портрета. В системе применяется ряд вышеизложенных технологий, в частности классическая поисковая технология, использующая словарь с автоматическим пополнением и инвертированные списки. В системе представлены как полные, так и документные списки для повышения скорости обработки простых одно-двусловных запросов и контекстных запросов. Имеются разные типы запросов: простой, игнорирующий всё, кроме дескрипторов, и строгий с развитым ИПЯ. Применяется также метод сличения сигнатур — частотных словарных характеристик документов для определения дублей, а для различения омонимов используется метод аналогий [7]. Кроме того, для текстового анализа в системе используется технология Zoom.

ОТЛИЧИЯ ТЕХНОЛОГИИ ZOOM

Отсутствие тезауруса или других словарей

В технологии Zoom не применяются тезаурусы или другие словари, например синонимический словарь. Таким образом, Zoom относится к категории автоматических систем. При этом Zoom осуществляет морфоразбор, используя морфологический словарь.

Относительный анализ

Взгляд "смысловой" и других сетей на документ абсолютен и инвариантен и не зависит от направленности выборки, следовательно, от интересов данного пользователя в данный момент. В Zoom принцип относительности работает всегда (при каждом запросе), а не только по отношению к конкретному пользователю. Пользователю не нужно задавать свой профиль или ждать образования истории для получения разумного результата.

Рассмотрим пример выделения сути предложения: "Дантисты из Урюпинска решили поехать в отпуск в Крым на поезде". Что здесь важно: дантисты из Урюпинска, наличие у них отпуска, желание провести его в Крыму или выбор данного типа транспорта? Технологии инвариантного выбора такого ответа не дадут. Не слишком поможет и персонификация информации, поскольку потребности пользователя не зафиксированы раз и навсегда. Технология Zoom даст ответ в зависимости от интересов данного пользователя в данный момент, точнее в зависимости от контекста интересующей его выборки документов.

Анализ больших выборок

Процесс обработки больших выборок документов требует обучения сети, следовательно выполняется очень медленно. Скорость работы технологии Zoom существенно быстрее, поскольку в ней нет обучаемой сети.

Отсутствие пошагового анализа

В отличие от байесовского метода для повышения производительности в технологии Zoom мы отказались от пошагового вычисления функций правдоподобия. Обработка полученной выборки текстов документов производится за один шаг. Сначала вычисляются сравнительно простые характеристики (простые суммы, квадратичные суммы) для каждого документа, а уже затем вычисляются собственно оценки по функциям правдоподобия. Такая методика существенно экономит вычислительные ресурсы и позволяет проводить статистический анализ тысяч документов в секунду.

ОТНОСИТЕЛЬНЫЙ КОНТЕНТ-АНАЛИЗ ТЕХНОЛОГИИ ZOOM

Такой метод заключается в сопоставлении частотного спектра выборки документов со спектром-образцом. Соображение здесь следующее: значимость данного слова (словосочетания) для предметной области тем больше, чем больше относительная плотность вероятности встречаемости этого слова в данной предметной области по сравнению с некоторой существенно большей областью (надобластью), подмножеством которой наша предметная область является. Например, область "Нейрохирургия" относится к надобласти "Хирургия", область "Хирургия" — к надобласти "Медицина" и т. д. Частным случаем надобласти может быть весь язык.

При этом подходе сравниваются две величины: плотность встречаемости данного слова в области и плотность в надобласти.

$$p = v_w/v, \tag{1}$$

где p — плотность встречаемости в области; v_w — число встреч данного слова в области (надобласти); v — число встреч всех слов (объем области в словоместах). Этот подход позволяет выявить все часто встречаемые слова, которые выделяют данную область. Простейший критерий, позволяющий это сделать, представляет собой отношение плотностей встречаемости в области и надобласти.

$$k_{\text{зн}} = p_0/p_{\text{но}}, \tag{2}$$

где $k_{\text{зн}}$ — коэффициент значимости (вес); p_0 — плотность встречаемости в области; $p_{\text{но}}$ — плотность встречаемости в надобласти.

Однако здесь есть несколько трудностей, главная из которых — работа с редко встречаемыми словами. Очевидно, что, согласно формулам (1) и (2), $k_{\text{зн}}$ достигает максимума для слов, которые встречаются только в области, и равен отношению объемов области и надобласти, независимо от числа встреч данного слова. Так, слово, встречающееся один раз только в области, будет иметь вес больший, нежели слово, встречающееся в области 999 раз из 1000 раз в надобласти. Следовательно, такое решение не обладает устойчивостью и инвариантностью по отношению к разным выборкам одной предметной области. Для получения необходимой устойчивости нужен немного другой критерий значимости.

КРИТЕРИЙ НЕСЛУЧАЙНОСТИ СЛОВА В ПРЕДМЕТНОЙ ОБЛАСТИ

Для получения такого критерия нам поможет понятие нулевой гипотезы. Если слово случайно оказалось в нашей выборке, то нулевая гипотеза выполнена. Критерием значимости нам может послужить вероятность отбрасывания нулевой гипотезы. При этом можно ограничить множество слов

в спектре, скажем теми, у которых эта вероятность больше 0,99 или любой другой достаточно близкой к единице величиной. Итак, допустим, что выполняется нулевая гипотеза для данного слова. Для этого будем считать, что слово распределено по надобласти абсолютно равномерно. Объем надобласти — $v_{но}$, объем области — v_0. Введём в надобласть $v_{но}$ N раз данное слово, и посчитаем вероятность попадания k раз в нашу область v_0 : $P(\xi = k|N)$. Рассматривается серия из N испытаний. С каждым испытанием связано случайное событие A — попадание выбранного слова w в область, и событие $\bar{A}$ — непопадание выбранного слова в область. Вероятность p события A равна n/N при условии равномерного распределения. Тогда количество ξ попадания слова в область в серии из N испытаний будет случайной величиной с биноминальным распределением вероятностей:

$$P(\xi = k) = C_N^k \cdot p^k \cdot (1-p)^{N-k},\ k = \overline{0, N}. \qquad (3)$$

Здесь C_N^k есть число сочетаний из N элементов по k. Вероятность $P(i_l)$ случайного попадания слова w в область определяется по формуле:

$$P(w) \equiv P(\xi > n) = \sum_{k=n+1}^{N} C_N^k \cdot p^k \cdot (1-p)^{N-k}. \qquad (4)$$

На этом можно бы закончить, однако вычисления по приведенной формуле довольно трудны. Поэтому применяются небольшие модификации для отбрасывания тех слов, которые явно не удовлетворяют заданному нами критерию значимости. В частности, распределение ξ аппроксимируется нормальным. В качестве критерия значимости берется величина

$$k_{зн} = -\ln(1 - p_w). \qquad (5)$$

ПРИМЕРЫ ИСПОЛЬЗОВАНИЯ ТЕХНОЛОГИИ ZOOM

Классификация

Пример 1

Центральная российская пресса. Документы (статьи) за 19.09.02. Выборка — около 1200 документов.

Здесь приведены темы, которые освещали СМИ в этот день, а также статьи, в которых данные темы даны наиболее приближенно к инфопортрету. Это — ежедневные обзоры (документы 1 и 2), а также статьи по одной из наиболее представленных тем

- Документ 3 — Шеварднадзе-Грузия-Панкисси-Грузинский
- Документ 4 — Кукуры-Кукура-Олигарх-Лукойл-нефть-Похищение-Похититель.

На приведенном примере видно, что система позволяет выявить ключевые термины и соответствующие документы, характеризующие контекст событий одного дня, описанного центральной российской прессой.

Пример 2

Общая российская пресса. Документы (статьи) за 2002 г., содержащие фразу "Атомная энергия". Выборка — около 3 тыс. документов.

1. Завтра (Москва). № 038 ... ТАБЛО. Выступление Буша 11 сентября, по оценкам **экспертов** СБД, подтвердило **прогнозы** относительно формального объявления войны Ираку. Структура речи американского президента свидетельствовало о том, что ООН ставится перед фактом решения США нанести удар, и позиция главной международной организации не будет приниматься во внимание официальным **Вашингтоном** в том случае, если она не пойдет на поводу у республиканских «ястребов». В ближайшие дни будет вынесен новый вариант ультимативных требований к Ираку с включением туда абсолютно неприемлемых... 2. Парламентская газета (Москва). № 177 ... Ирак сделал правильный шаг. Председатель Совета **Федерации** Сергей Миронов считает, что именно последовательная позиция России побудила Ирак вернуть международных инспекторов в страну. – Это правильный шаг, – подчеркнул **спикер** верхней палаты российского парламента. – Теперь задача инспекторов – посмотреть, действительно ли иракцы прекратили разработку оружия массового поражения. И если это так, то нужно делать следующий шаг – снимать санкции с Ирака. Александр ШАРОВ. *** Письмо из Грузии признано «неудовлетворительным" Президент ... 3. Завтра (Москва). № 038 ... «КУДА ТЕБЯ НЕСЕТ?..». Против кого собирается **воевать** в Грузии Путин? Против боевиков в Панкисси? Но их там давно нет. Лидеры боевиков своевременно получили информацию о подготовке Россией ударов по панкисским базам, и уже месяца два разведка докладывает о передислокации чеченских отрядов в глубь Грузии, где им выделены объекты на территориях бывших советских воинских частей. Главное же командование боевиков, координационный центр, который управляет потоками денег, наемников и оружия, надежно укрыт в центре Тбилиси на одной из баз МГБ Грузии. Конечно, все это было невозможно без протекции и помощи грузинских ... 4. Московская правда (Москва). № 172 ... КУКУРЫ НА ВСЕ НЕ ХВАТИТ. Телефон «ЛУКОЙЛа» обрывают любители легких денег.	Отметьте слова, которые хотите добавить в запрос, и нажмите кнопку «Уточнить» Вкл/Выкл / Слово ☐ КУКУРЫ ☐ ИРАК ☐ КУКУРА ☐ РЕФЕРЕНДУМ ☐ ШЕВАРДНАДЗЕ ☐ ГОСПОДИН ☐ КУКУРУ ☐ ГРУЗИЯ ☐ ОЛИГАРХ ☐ ЛУКОЙЛ ☐ НЕФТЬ ☐ ПАНКИССИ ☐ МЛРД ☐ ПОХИЩЕНИЕ ☐ ЕВРО ☐ ООН ☐ БАГДАД ☐ Ъ ☐ САДДАМ ☐ ПОХИТИТЕЛЬ ☐ ФИЛЬМ ☐ БУШ ☐ ЦЕНТРИСТ ☐ ГРУЗИНСКИЙ

1. Наше время (Ростов-на-Дону). № 140-141 ... *Прежде всего потому, что там со ссылкой* на Министерство **обороны** США утверждалось (правда, довольно туманно), что некоторые из названных материалов были похищены с **Волгодонской** АЭС. Британская газета повторяет пентагоновскую информацию о том, что **данные** о краже якобы поступили в **Международное агентство** по • **атомной** • **энергии** (МАГАТЭ) из официальных российских источников. МАГАТЭ немедленно поставило в известность об инциденте... 2. Труд (Москва). № 173 ... В австрийской столице Вене прошла 46-я сессия Генеральной **конференции** Международного **агентства** по • **атомной** • **энергии** (МАГАТЭ), посвященная актуальным вопросам в области **мирного** использования • **атомной** • **энергии**. Подводя на пресс-конференции итоги сессии, министр по • **атомной** • **энергии** РФ Александр **Румянцев**... 3. Наше время (Ростов-на-Дону). № 084-085 ... – прозвучала вполне оптимистично: сегодня **выплачено** около половины всей суммы, а вторая половина, по заявлению министра по • **атомной** • **энергии** А. **Румянцева**, также должна быть вскоре перечислена. *Надо сказать, теперь о страстях*, что бушевали вокруг обещанных Минатомом денег еще не так давно, уже мало кто вспоминает. Другое дело – вопрос **финансирования** и **выполнения** строительно-монтажных работ по объектам соцкультбыта в 30-километровой зоне. В очередной раз комиссия... 4. Известия (Москва). № 036 ... Кризис миновал, желающих посидеть на поплавке резко прибавилось. Вчера министру по • **атомной** • **энергии** Александру **Румянцеву** пришлось срочно опровергать **доклад ЦРУ** американскому конгрессу о том, что, по **сведениям** "российского официального лица", в 1998 году в России украли "значительное количество расщепляющего материала, достаточное **для изготовления** атомной **бомбы**". Накануне по иску челябинских граждан, ...	Отметьте слова, которые хотите добавить в запрос, и нажмите кнопку Уточнить. Вкл/Выкл / Слово ☐ АЭС ☐ ФЕДЕРАЦИЯ ☐ МИНАТОМ ☐ РЕАКТОР ☐ РАДИОАКТИВНЫЙ ☐ ОЯТ ☐ ЭНЕРГОБЛОК ☐ МАГАТЭ ☐ РОСЭНЕРГОАТОМ ☐ УРАН ☐ ОТХОД ☐ РАДИАЦИОННЫЙ ☐ ИСПОЛЬЗОВАНИЕ ☐ ОБЪЕКТ ☐ АТОМЩИК ☐ ТОПЛИВО ☐ СООТВЕТСТВИЕ ☐ ЗАКОНОДАТЕЛЬСТВО ☐ РУМЯНЦЕВ ☐ ЭНЕРГЕТИКА ☐ ОБЕСПЕЧЕНИЕ ☐ ЭКСПЛУАТАЦИЯ ☐ ЭНЕРГЕТИК ☐ ЭЛЕКТРОСТАНЦИЯ

Здесь приведены темы, идентифицирующие объект "Атомная энергия" и характеризующие контекст данного объекта. В документе 1 приводятся сведения о краже плутония, а в документе 2 описывается сессия МАГАТЭ. Из приведенного примера видно, что релевантность вышеуказанных документов довольно велика. Система хорошо определяет подходящие к запрошенной теме "Атомная энергия" документы.

Составление обзоров, выявление тенденций

Тема войны

Попробуем решить задачу определения основной направленности информационных сообщений мировых агентств, связанных с темой войны. Решив эту задачу, мы сможем выяснить, какие объекты (понятия, регионы, страны, личности) ассоциируются в массовом сознании с войной. Иначе говоря, сможем предсказать, какой ответ мы получим, разбудив среднего обывателя ночью и спросив о войне. Зная это, службы по связям с общественностью соответствующих стран смогут решать конкретную задачу, как бороться против такого образа в общественном сознании.

Запрос: Война (WAR) по недельным периодам (начало-середина-конец 1999 г.).

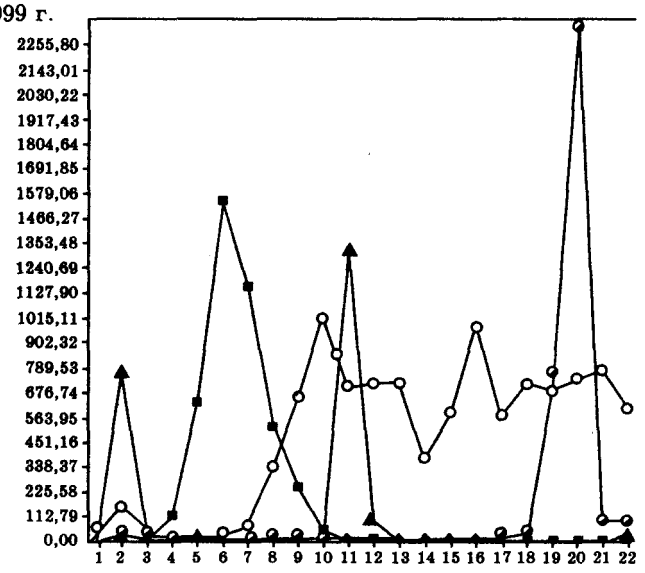
Поисковая база: сообщения мировых агентств.

Проанализировано около 24 тыс. документов. Время анализа 7 сек. на каждую выборку (средний объем — около 1 тыс. документов). Приведем три выборки за указанный период.

Главные темы выборок

Query_10.01.2001 15:11:40 War&$MainDate: 1.1.1999..7.1.1999		Query_10.01.2001 15:13:07 War&$MainDate: 25.6.1999..1.7.1999		Query_10.01.2001 15:14:06 War&$MainDate: 24.12.1999..30.12.1999	
Слова	Коэфф. значимости	Слова	Коэфф. значимости	Слова	Коэфф. значимости
CASTRO	1903,07	JUNE	2459,08	GROZNY	4101,44
JAN	1545,95	WAR	1910,42	WAR	2363,91
IRAQ	1485,55	KASHMIR	1511,41	RUSSIAN	2137,29
ANGOLA	1480,36	LEBANON	796,15	CHECHEN	1931,58
WAR	1373,06	PAKISTAN	722,35	DEC	1888,16
IRAQI	1369,46	HORTA	673,43	PORTILLO	1808,15
UNITA	1353,76	KOSOVO	612,64	KALEJS	1414,08
JANUARY	1012,24	KOREAN	601,24	HIJACKER	1350,40
HUAMBO	996,26	TALK	577,60	DECEMBER	1305,28
REBEL	933,78	INDIA	545,13	CHECHNYA	1299,46
CUBA	859,23	PEACE	522,44	MILLEN-NIUM	1220,05
NO-FLY	738,16	BEIRUT	508,68	HIJACK	961,19
KHMER	724,31	GUER-RILLA	478,14	REBEL	893,63
ANGOLAN	705,72	SERB	462,11	KANDAHAR	819,08
ROUGE	702,29	SHARIF	461,64	CHRISTMAS	799,40
SADDAM	692,12	HERSHER	447,70	VUKOVIC	789,81
ZONE	673,89	NKOMO	430,21	APPLE	717,45
SHIV	671,44	KOREA	408,04	GUATE-MALA	710,91
SENA	638,49	NORTH	402,57	GALIC	698,29
LUANDA	604,63	LUSAKA	399,66	PRO-MOSCOW	678,74
HIROHITO	591,12	INDIAN	393,05	INDIAN	655,59
LIPKIN-SHAHAK	575,96	SKIRBALL	390,54	CITY	609,23
SEVAN	526,09	JAMHOUR	375,76	PLANE	562,49
FARC	499,44	INFIL-TRATOR	369,59	IVORY	555,58
BAGHDAD	477,22	PRISTINA	361,67	TROOP	553,67
FIDEL	472,12	SCHEFFER	356,52	CHATHAM	524,94
KHIEU	366,87	OCALAN	342,00	GUATE-MALAN	501,30

Итак, мы видим, что вектор таких сообщений, следовательно, и вектор общественного сознания резко изменился в течение 1999 г. В начале года это были (Castro, Angola, Cuba, UNITA, Iraq, Saddam) две основные темы. Для середины года характерно (Kashmir, Pakistan, India, Lebanon, Beirut, Kosovo, Korea) несколько тем. Но в конце года практически все заняла чеченская тема (Grozny, Chechen, Chechnya, Russia) и намного менее значимы были Kandahar и Guatemala. Ниже приведен график изменения значимости некоторых наиболее важных тем по неделям во второй половине 1999 г.



На этом графике также видна постоянная, суммарная значимость чеченской темы (представлена одним словом Chechen). Это произошло после подъема в конце сентября 1999 г. Взлет же остальных тем (Golan, Timor, Pakistan) был кратковременным.

Тема наркотиков

Еще один пример анализа направленности сообщений прессы, связанных с проблемой наркотиков. Можно начать с запроса "Наркотики" (Narcotics), но чаще применяется термин Drugs, и здесь возникает другая проблема. В получаемую выборку попадают документы, относящиеся к медикаментам, появление их вызвано многозначностью слова Drugs. Эта проблема хорошо известна и называется проблемой разделения объектов. С помощью технологии БИО-Zoom она решается в большинстве случаев довольно просто. Получается вот что:

Запрос: DRUGS

Поисковая база: сообщения мировых агентств.

Главные темы выборок

VIAGRA	MYANMAR
HEROIN	TRAFFIC
COLOMBIA	DALLAGLIO
PATIENT	SUBSTANCE
COCAINE	NARCOTICS
TRAFFICKERS	DOCTOR
DOPING	ANTI-DRUG
COLOMBIAN	MYANMAR
DISEASE	BOGOTA
CANCER	PROTEIN
IMPLANT	ATHLETE

Видно, что в данном информационном портрете содержатся синонимы (Narcotics), а также темы, не касающиеся темы наркотиков (Athlete). Применив отсечение, получаем следующее:

Запрос: "(NARCOTIC | NARCOTICS | Drugs) & ^VIAGRA & ^PATIENT & ^DISEASE & ^VACCINE & ^DOPING & ^ATHLETE"

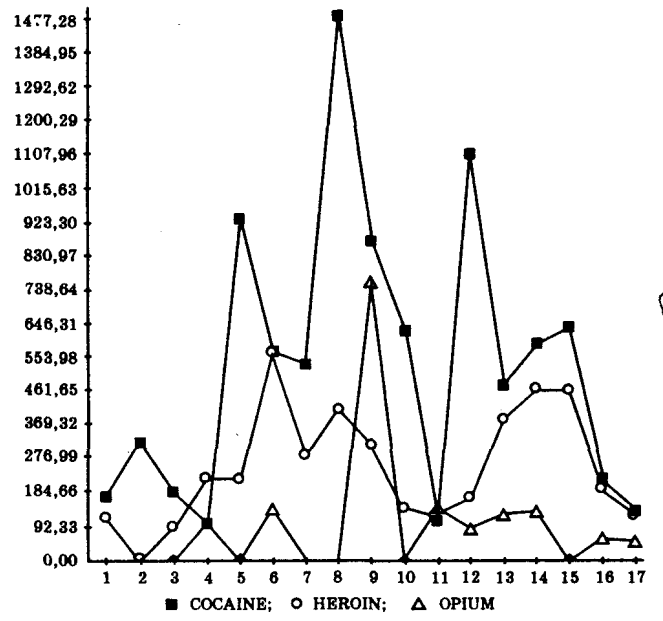
Получилась довольно чистая выборка.

Главные темы выборок

COLOMBIA	JUAREZ
COCAINE	NARCOTICS
MEXICO	DALLAGLIO
COLOMBIAN	OPIUM
HEROIN	MCCAFFREY
TRAFFICKERS	ANTI-DRUG
MEXICAN	SALINAS
CARTEL	PASTRANA
MYANMAR	CRIME
TRAFFIC	ADDICT

Как видим, ушли не только темы, напрямую отсеченные в запросе, но и связанные с ними (Cancer, Protein). Заметьте, мы не просматривали сами документы для отделения ненужных.

Теперь проведем анализ временных трендов. Проанализировано около 11 тыс. документов. Время анализа — 7 сек. на каждую выборку (средний объем — около 600 документов). Ниже приведен график изменения значимости некоторых тем, относящихся к типам наркотических веществ по двухнедельным интервалам во второй половине 1999 г.



01: 10.5.1999 – 23.5.1999; 02: 24.5.1999 – 6.6.1999; 03: 7.6.1999 – 20.6.1999; 04: 21.6.1999 – 4.7.1999; 05: 5.7.1999 – 18.7.1999; 06: 19.7.1999 – 1.8.1999; 07: 2.8.1999 – 15.8.1999; 08: 16.8.1999 – 29.8.1999; 09: 30.8.1999 – 12.9.1999; 10: 13.9.1999 – 26.9.1999; 11: 27.9.1999 – 10.10.1999; 12: 11.10.1999 – 24.10.1999; 13: 25.10.1999 – 7.11.1999; 14: 8.11.1999 – 21.11.1999; 15: 22.11.1999 – 5.12.1999; 16: 6.12.1999 – 19.12.1999; 17: 20.12.1999 – 31.12.1999

На этом графике видна большая постоянная значимость темы Cocaine. Из остальных тем редко превышают эту планку Opium, Heroin.

ЗАКЛЮЧЕНИЕ

Применение технологии Zoom в рамках промышленной системы Галактика-Zoom показало её эффективность для многих пользователей. Технология позволяет решать задачи текстового анализа, в том числе классификации на больших объемах текстов с приемлемой скоростью, что очень важно, поскольку часто решение требуется принимать "вчера", следовательно всю информацию необходимо собрать как можно скорее.

ЛИТЕРАТУРА

1. *Лукашевич Н. В., Салий А. Д.* Представление знаний в системе автоматической обработки текстов // Науч.-техн. информ. Сер. 2 .— 1997 .— № 3.

2. *Большаков И. А., Гельбух А. Ф.* Рубрикация словосочетаний в базах данных по элементам толкования сочетаемых слов // Науч.-техн. информ. Сер. 2 .— 2000 .— № 6.

3. *Ашманов И. С., Власова А. Е., Зоркий К. П.*, и др. Технология фильтрации содержания для Интернет // Труды Международного семинара Диалог '2002.

4. *Kohonen T.* Self-organization of very large document collections: State of the art.— Helsinki University of Technology, Neural Networks Research Center, PO Box 2200, FIN-02015 HUT, Finland.

5. *Alexandrov M., Gelbukh A., Makagonov P.* Some keyword-based characteristics for evaluation of thematic structure of multidisciplinary documents // Proc. of CICLing-200, International Conf. on Intelligent text processing and Computational Linguistics, CIC-IPN, Mexico City.

6. *Шабанов В. И., Власова А. Е.* Алгоритм формирования ассоциативных связей и его применение в поисковых системах // Труды Международного семинара Диалог '2003.

7. *Белоногов Г. Г., Зеленков Ю. Г., Новоселов А. П., Хорошилов А. А., Хорошилов А. А.* Метод аналогии в компьютерной лингвистике // Науч.-техн. информ. Сер. 2 .— 2000 .— № 1.

АВТОРСКИЙ УКАЗАТЕЛЬ К Т. 28, 2003