УДК 025.4.03:81

В. Ф. Пархоменко

Работа с русскоязычными полнотекстовыми базами данных в ИПС АРТЕФАКТ

Рассматривается история разработки семейства лингвистических ИПС и последний представитель этого семейства — ИПС "Артефакт", предназначенная для обработки больших массивов текстовой информации произвольной тематики; организация этих массивов в базы данных; обеспечение быстрого и качественного поиска информации в локальных и глобальных сетях. Высказывается предположение, что ИПС, работающей с текстами, написанными на русском языке, необходим морфологический анализ.

ИСТОРИЯ РАЗРАБОТКИ

Первыми предшественниками ИПС "Артефакт" можно считать ИПС семейства "Пусто—Непусто" [1], которые разрабатывались в 1963–70 гг. в "Информэлектро" на ЭВМ "Урал-4" и "Минск-22". В 1973–75 гг. появилась первая версия ИПС "Скобки" на ЭВМ "Минск-32" [2, 3], переведенная в 1980–1983 гг. на ЕС ЭВМ [4, 5]. Руководили этими работами В. С. Чернявский и Д. Г. Лахути, не только создавшие один из передовых коллективов разработчиков ИПС, но и основавшие оригинальную научную школу в информатике.

Следует отметить, что ИПС "Пусто—Непусто" была не только одной из первых реально действующих документальных ИПС в мире, но и первой в мире системой с автоматическим индексированием — этот примечательный факт, к сожалению, в последующие годы упорно замалчивался, и отнюдь не по научным соображениям*. Что касается ИПС "Скобки", то ее язык запросов превосходил по мощности большинство языков современных ИПС.

Будущие разработчики ИПС "Артефакт" с 1967 г. принимали активное участие в этих и других разработках, которые проводились в "Информэлектро", таких, как система синтаксического анализа русских текстов, квазиреферирование, системы анализа словосочетаний, псевдоперевод, многоязычное индексирование и т. д.

Все упомянутые ИПС использовались в реальных информационных технологиях, обрабатывая солидные по тем временам объемы информации.

Если попробовать выделить основную идею, являющуюся стержнем всех разработок, то это, в первую очередь, понимание системой естественного языка, на котором написаны тексты документов, хотя бы на лексическом уровне. Для реализации этой идеи в ИПС "Пусто—Непусто" использовался морфологический словарь объемом около 5 тыс. основ, в ИПС "Скобки" — до 20 тыс. основ, снабженных грамматической информацией. Но для создания коммерческой поисковой системы общего назначения, индифферентной к тематике обрабатываемых документов, были нужны другие словари, другого качества и объема.

Создание таких словарей и действующих алгоритмов морфологического анализа стало возможным после того, как руководитель машинного фонда русского языка В. М. Андрющенко в начале 1991 г. любезно предоставил нам электронную версию "Грамматического словаря русского языка" А. А. Зализняка. Переработка этого словаря (электронной копии книги) в словарно-алгоритмическую систему, которая велась под руководством и при непосредственном участии Д. Г. Лахути, потребовала значительных усилий, в результате чего появилась электронная система морфологических словарей для анализа и лемматизации слов русского языка. В основу этой системы были положены наработки группы лингвистов: Г. А. Лесскиса, Н. А. Еськовой, Т. Ю. Кобзаревой и др. На базе этой системы в 1991–1995 гг. и были созданы такие коммерческие продукты, как программа проверки правописания "Пропись" (версия 1.X–3.X) и лингвистические ИПС "Агама", "Мирс" и "Артефакт".

Общее в этих ИПС — их полная независимость (в том смысле, что они не используют какую-либо стандартную СУБД в качестве основы), примерно одинаковые лингвистические возможности и высокая скорость работы.

"Агама" — это наш первый опыт, "Мирс" более современная программная архитектура и структуры баз данных, позволяющие быстро работать с таким специфическим носителем, как лазерный диск.

ИПС "Агама" мы начали разрабатывать в феврале 1992 г., намереваясь выпустить через год первую коммерческую версию. И действительно, уже в декабре на конференции в ВИНИТИ мы сумели продемонстрировать нечто, вызвавшее интерес у немногочисленной публики, а в марте 1993 г. были проданы первые экземпляры. В конце 1993 г. появилась специализированная CD-ROM ИПС "Мирс" (мы тогда работали в CD-ROM издательстве "Media Mechanics"), в 1995 г. была выпущена первая версия ИПС "Артефакт", а в 1997 г.

*После эмиграции в 1973 г. на имя В. С. Чернявского в советской научной литературе было наложено табу. Профессор Чернявский скончался в г. Брауншвейге в январе 1996 г.

начато коммерческое использование ИПС "Артефакт" — версии "Клиент-Сервер" для WWW-клиентов сети Интернет, которая успешно демонстрировалась на выставке "Cebit-97" в Ганновере.

"Мы" — это не литературный прием, а Н. С. Назарова и Е. В. Травкина, разделившие с автором нелегкий труд по созданию системы. Большой вклад в разработку, особенно на ее первых этапах, внес Д. Г. Лахути — консультант проекта. Разработкой архитектуры и интерфейсов Интернет-версии руководит С. А. Романенко.

"Артефакт" — это не только дальнейшее развитие идей и методов, реализованных в ИПС "Агама" и "Мирс", но и воплощение нового понимания информационных технологий. Такое понимание возникло в результате слияния двух коллективов — разработчиков информационных инфраструктур для аналитической деятельности (руководитель Ю. П. Поляков, ныне генеральный директор "Интегрум-Техно") и разработчиков лингвистического программного обеспечения. ИПС "Артефакт" является программно-технологической базой крупных информационных центров, устойчиво работающим коммерческим продуктом, поддерживающим в общей сложности у всех корпоративных пользователей до 100 Гбайт различных баз данных (только в "Интегрум-Техно" более 250 баз данных общим объемом свыше 18 Гбайт)[7].

Полнотекстовых ИПС в наше время существует довольно много. Это и универсальные в смысле носителя баз данных системы, и специализированные системы, предназначенные для поддержки баз данных на CD-ROM. Широко известны, например, такие системы, как CDS-ISIS, специально разработанная для слаборазвитых стран и бесплатно распространяемая ЮНЕСКО, системы ZyIndex, SPIRS фирмы Silver Platter, Lexis-Nexis, Reuter и многие другие. В последнее время появилась такая новинка, как Excalibur, усиленно внедряемая в российские государственные и коммерческие структуры. Есть и отечественные разработки, имеющие свою историю и опыт эксплуатации, такие, например, как ИРБИС, ШЕРШЕ, PCBIRS. Кроме того, практически все базы данных на отечественном информационном рынке "одеты" в специализированные программные оболочки, каждая в свою.

Все упомянутые ИПС (за исключением Excalibur, у которой "нечеткий" поиск и о которой разговор особый) отличаются друг от друга степенью дружественности интерфейса, сложностью загрузки баз данных и т. п. Объединяют же все ИПС, заслуживающие упоминания, следующие обстоятельства:

во всех ИПС используется тот или иной алгоритм инверсного поиска (поисковый индекс), что позволяет минимизировать время поиска (впрочем, время все-таки иногда отличается на порядок, но это уже зависит от искусства выбора оптимальных структур данных и программирования);

все ИПС содержат "джентльменский набор" поисковых операторов — И, ИЛИ, реже И НЕ, операторы вхождения в заданный контекст (обычно это поле, абзац, предложение), совсем редко встречается чистое отрицание НЕ;

при отождествлении слов запроса и слов документов используется правое усечение (truncation) слова запроса (это необходимо при работе с русскоязычными документами). Впрочем, в последнее время одна за другой появляются ИПС, использующие морфологический анализ при загрузке базы данных и поиске, — "Апорт", "Яndex", "Следопыт" и другие.

НУЖЕН ЛИ МОРФОАНАЛИЗ?

Рассмотрим подробнее проблему отождествления слов запроса со словами документов при поиске в русскоязычных базах данных.

Интересующие нас слова встречаются в документах в разных грамматических формах (*автомобиль, автомобилями, автомобилях* и т. п.). Это означает, что ИПС должна уметь отождествлять разные словоформы одного и того же слова и при этом не ошибаться или ошибаться как можно реже. Существует два способа решения этой проблемы.

При правом усечении решение проблемы перекладывается на пользователя. В запросе цитируются не полные слова, а только их начало. С помощью какого-нибудь значка (или по умолчанию) ИПС понимает, что словам запроса в документах соответствуют все слова, имеющие такое же начало. В этом случае запрос, например, выглядит так:

таможенн и пошлин* и (автомобил* или трактор*).*

Главная, не всегда сразу очевидная, проблема при таком подходе — правильно выбрать точку усечения. Если мы зададим в запросе слишком длинное начало, то рискуем потерять нужные словоформы, слишком короткое — получим слова, не относящиеся к делу. Ситуация осложняется и тем, что слова русского языка имеют очень развитую систему окончаний: русский язык — язык флективный.

И если в английском языке с его слабой флективностью наличие аппарата усечения решает проблему практически полностью, то для русского языка не все так просто, даже если не принимать во внимание совсем уж неестественный вид запроса. Кроме того, и это самое главное, даже в нашем примере "хорошими" словами будут признаны *таможенник, автомобильный, тракторный* и прочие, может быть не очень относящиеся к делу. Ограничение длины окончания, применяемое в некоторых системах, также не решает проблему полностью.

Намного более эффективный способ решения этой проблемы — это автоматический морфологический анализ слов документов и запросов, когда решение проблемы перекладывается на ИПС. В этом случае система должна уметь для любой пары "слово запроса — слово документа" безошибочно решать, одно и то же это слово, или нет. Желательно, чтобы точность автоматического отождествления могла регулироваться, например, с точностью до окончания (до части речи) — *автомобиль, автомобилями*, или с точностью до корня (до большой парадигмы) — *автомобили, автомобильными*. Кроме того, желательно иметь возможность употреблять в запросах разные словоформы. Так будет удобно пользователю.

Хорошее решение этой проблемы (с точностью до омонимии) возможно только при наличии в системе средств полного морфологического анализа, использующих развитую систему словарей и

точные лингвистические алгоритмы. Заметим, что только при наличии таких средств можно безошибочно отождествить такие пары "слово запроса — слово документа", как *ребенок —дети, шел — идущий* и т. п.

Разумеется, такой способ отождествления предпочтителен: пользователю дается возможность сосредоточиться на содержательных аспектах запроса и не думать, что будет, если он поставит усечение в другом месте.

МОРФОАНАЛИЗ В ИПС "АРТЕФАКТ"

Полнота понимания "Артефактом" лексики русского языка основана на словаре более чем в 100 тыс. основ, а точность понимания — на морфологической системе, в которой около 2000 различных классов слов (1000 — глаголы, 1000 — существительные, прилагательные и все остальное). Реально слов распознается не 100 тыс., а намного больше — распознаются и нормализуются приставки, имеющие самостоятельное смысловое значение, а таких около 500 (например, *авто-, авиа-, радио-, сельско-* и т. п.). Сколько распознается словоформ, никто не знает: может быть, 3–4 млн, а может быть и больше (нельзя учесть все возможные присоединения приставок). Тем не менее, сколько бы слов не распознавалось, этого всегда будет мало — невозможно учесть всю научную и техническую терминологию, кроме того, каждый день появляются и прочно входят в употребление *регионы, брифинги* и т. п. Поэтому документы, вводимые в "Артефакт", могут содержать и незнакомые слова, отсутствующие в словаре. Они поступают в базу данных во всех встреченных словоформах, а их морфологическое отождествление со словами запросов производится на этапе поиска. И это отождествление практически безошибочно, так как и научная терминология, и слова типа *регионы* склоняются достаточно регулярно (впрочем, и *регион* и *брифинг* в словаре ИПС "Артефакт" все-таки есть). Точно так же обрабатываются и слова английского языка: в систему введена и английская морфология — в скромном объеме, но достаточном, тем не менее, для грамотной обработки англоязычных текстов.

ПОЛНОТЕКСТОВЫЕ ДОКУМЕНТЫ

Следующая проблема, уже не связанная с языком, — обработка полнотекстовых документов. Не будем вдаваться в чисто технические аспекты: очевидно, что с большими документами управляться труднее, чем с маленькими. Главная сложность в том, что в этом случае не обойтись без структурирования документов и развитых средств контекстного поиска, для реализации которых в индексе базы данных для каждого слова нужно иметь не просто соответствие "слово — номера документов", но и сведения о всех вхождениях слова во все документы, с точностью до всех элементов структуры. Появляются новые проблемы: размер индекса, сложность структурирования (вручную?, автоматически?), сложность алгоритмов, а, в конечном итоге, все лимитирует допустимое время разработки. Приходится идти на компромиссы.

Итог этих компромиссов в ИПС "Артефакт" выглядит так. Объем обрабатываемого документа — не более 1,5 Мгб. В большинстве случаев этого достаточно, ну а если документ все-таки больше, то его можно поделить на части. Возможности структурирования позволяют учитывать вхождение слов в поле документа, предложение, а также учитывать порядок слов в предложении и расстояние между ними. Эти возможности мы постарались максимально использовать: в языке запросов "Артефакта" большой выбор контекстных операторов.

МНОГОБАЗОВЫЙ ПОИСК

В отличие от ИПС с линейным поиском, когда последовательно просматриваются документы базы данных (даже в наше время такое иногда встречается), в системах с индексом скорость поиска не зависит напрямую от размера базы данных. Гораздо важнее частотность терминов запроса: вычислить термин, входящий в десять документов, проще, чем термин, входящий в каждый (или почти каждый) документ базы. Но в любом случае, при превышении какого-то определенного объема база становится плохо управляемой — снижается скорость поиска (ведь на практике в формулу запроса входят термины разной частотности), замедляется процесс дозагрузки документов в базу, начинаются технологические сложности с резервным копированием и т. д. По нашему мнению, оптимальный размер базы данных — это 400–500 Мгб исходных текстов; такую базу можно поместить на CD-ROM или на магнитооптический диск. А поисковая система должна уметь искать не только в одной базе данных, но и в нескольких базах, и чем больше, тем лучше (для пользователя это выглядит как параллельный поиск). Поэтому важной особенностью "Артефакта" является многобазовый поиск — до 45 БД в ДОС-версии и до 900 в Интернет-версии, причем возможна параллельная обработка при наличии нескольких процессоров.

ЯЗЫК ЗАПРОСОВ

Эффективность поисковой системы — полнота* и точность поиска — зависит от мощности языка, на котором формулируются запросы к системе, от его выразительных средств. Возможность реализовать эти средства напрямую зависит от полноты описания в индексе вхождений слов документов и от структуры индекса, которая должна позволять быстро обрабатывать даже такой экзотический поисковый оператор, как двухстороннее усечение (поиск всех слов, содержащих заданный фрагмент). Структура индекса БД ИПС "Артефакт" обеспечивает и быстрый контекстный поиск, и двустороннее усечение.

Поисковыми признаками в тексте документа могут быть не только слова, но и словосочетания, даты, числа в различной форме и т. п. Чем

*Исследования Д. Г. Лахути показали, что полнота поиска в первую очередь зависит от выбранного критерия смыслового соответствия и наличия развитого тезауруса [6]. Заметим, что в ИПС "Артефакт" есть все средства ведения и использования тезауруса.

богаче язык, тем больше таких элементов текста может учитываться при поиске в языке запросов и тем точнее можно формулировать запросы.

Синтаксис языка запросов, с нашей точки зрения, должен быть "двуслойным", т. е., с одной стороны, позволять начинающему пользователю сразу вводить простые запросы и получать на них осмысленные ответы, а, с другой — по мере накопления опыта, пользователь должен иметь возможность естественным образом переходить ко все более сложным запросам, осваивая все новые элементы синтаксиса. С этой точки зрения, нельзя признать удачным существование в некоторых системах двух языков — простого и профессионального. Язык должен быть один и, так же как и естественный, позволять выражать смысл запроса просто или более утонченно, и, соответственно, более точно.

Кирпичиками, из которых строится запрос в ИПС "Артефакт", являются слова естественного языка, их фрагменты — слова, усеченные слева, справа или с двух сторон, даты и интервалы дат. Будем называть их элементарными термами — элементарными потому, что, с точки зрения языка, их нельзя разложить на более мелкие части. Элементарный терм — это уже полноценный запрос, например, в ответ на запрос *судьи* будут выданы документы, в которых встречается слово *судья* в различных падежах, на запрос **ализаци** — документы, в которых содержатся такие слова, как *индустриализация, коллективизация* и тому подобные **ализации* (* является знаком усечения), на запрос 22.06.1941–08.05.1945 — все документы, в которых есть даты, относящиеся к периоду Великой Отечественной войны. В некоторых случаях уже и этого достаточно. Но если в запросе есть хотя бы два слова, то нужны средства для описания условий их совместного вхождения в документы. В первую очередь это логические операторы И, ИЛИ и НЕ, которыми связываются элементарные термы, образуя фразы запроса. Оператор И требует совместного вхождения связываемых им термов в документ, оператор ИЛИ — вхождения хотя бы одного из связываемых термов, НЕ запрещает вхождение в документ следующего за ним терма. Запрос структурируется скобками — в языке ИПС "Артефакт" не определено старшинство операторов. У нас уже достаточно средств, чтобы сформулировать запросы типа

народные судьи (((городские или муниципальные) власти) или муниципалитет).

Мы получили довольно-таки сложную фразу, напоминающую арифметическое выражение, в которой операторы связывают не только элементарные термы, но и составные — фразы, заключенные в скобки. При отсутствии оператора между словами обычно подразумевается оператор И. Заметим, что для многих ИПС такой запрос является пределом сложности, да еще каждое слово должно быть усечено справа. Для небольших документов этого, может быть, и достаточно, но с увеличением их размера быстро растет вероятность получить текст, в начале которого говорится об отношении народной власти к судьям, а в конце — о городской канализации. Т. е. необходимо ограничить размер контекстов, на которых выполняются фразы *народные судьи (городские или муниципальные) власти*, например, потребовать, чтобы они выполнялись на одном предложении или на группе слов в пределах предложения. В "Артефакте" это делается так:

(народные судьи /n) — контекст, на котором должна выполняться фраза запроса, ограничен одним предложением — в документе должно быть хотя бы одно предложение, в которое входят слова *народный и судья;*

((городские или муниципальные) власти /c4) — контекст, на котором должна выполняться фраза запроса, ограничен четырьмя смежными словами (*n*-кой слов): в документе должно быть хотя бы одно предложение, в котором есть слово *власть* и слово *городской* или слово *муниципальный*, они могут стоять в любом порядке и между ними может быть не более двух других слов.

Операторы */n* и */c* действуют в пределах скобок или в пределах запроса.

Поиск в пределах поля определяется его именем, за которым следует операнд или фраза в скобках. Возможен поиск по нескольким полям, в этом случае операнд предваряется несколькими именами полей. Например:

/ДАТА 09.1997 /ЗАГЛ/ТЕКСТ ((визит или посещение) Ельцина в Нижний Новгород /n).

Оператор следования, необходимый для поиска словосочетаний, выглядит так:

влияние :0 краситель :2 бактерия,

где ":2" оператор следования, соединяющий термы; ":2" указывает, что между словами *краситель* и *бактерия* может быть до двух других слов. Обратим внимание, что оператор следования, так же, как и оператор И, соединяет именно термы, а не слова. Например:

влияние : (краситель или краска) : бактерии.

Жесткий контекст можно определить, заключив словосочетание запроса в кавычки, например,

"Нижний Новгород" будет эквивалентно Нижний :0 Новгород.

РАНЖИРОВАНИЕ И ПРОСМОТР НАЙДЕННЫХ ДОКУМЕНТОВ

Найденные документы могут быть упорядочены в порядке их ввода в БД (обычно это прямой хронологический порядок), в обратном порядке или ранжированы по смыслу. В последнем случае самые релевантные, с точки зрения системы, документы будут выведены первыми. В ИПС "Артефакт" есть все три возможности упорядочения документов, найденных по запросу.

Необходимо подчеркнуть важность того, как в ИПС реализован просмотр найденных документов. Конечно, можно просто показать документ, как это делается в любом редакторе, а дальше пусть пользователь сам разбирается, нужен ему этот документ или нет, сам отыскивает релевантные фрагменты и т. д. Но если документ достаточно велик, то в этом случае не обойтись без дополнительных средств навигации.

При показе найденных документов как минимум необходимо выделить в тексте найденные слова, чтобы не заставлять пользователя читать документ в поисках нужных фрагментов текста. Еще

лучше иметь возможность позиционировать просматриваемый документ, переходя от одного релевантного фрагмента к другому, минуя текст, лежащий между фрагментами. Желательно выводить не только найденные документы, но и их фрагменты, выделенные пользователем, выводить только документы, признанные пользователем релевантными и т. д.

В ИПС "Артефакт" эти проблемы решены следующим образом.

В ДОС-версии сначала показывается общая статистика поиска по всем базам данных, по которым проводился поиск. Далее можно просматривать документы, найденные в конкретной базе. Просмотр может быть двухступенчатым — просмотр заголовков/просмотр документов или непосредственно просмотр документов, минуя заголовки. При просмотре документов показываются релевантные фрагменты документов. В тексте они выделены цветом, слова (даты), по которым произошло отождествление, — другим цветом. Возможен просмотр с переходом от одного релевантного фрагмента к другому. Найденные документы (или их фрагменты) можно записывать в журнал поиска или отдельными файлами. Можно исключить ненужные документы (шум), пометив их.

В Интернет-версии приблизительно такая же схема показа документов, но при заголовках еще показываются и релевантные фрагменты — кусочки текста, с точки зрения системы отвечающие на запрос. Как показывает практика, этого в большинстве случаев достаточно, чтобы решить вопрос о релевантности или нерелевантности документа. Интересно отметить, что похожая схема показа была реализована в первой версии — в ИПС "Агама" показывались дайджесты документов — указанные поля и релевантные фрагменты, но в дальнейшем развитии эта схема была утрачена (а жаль!) и восстановлена в новом качестве только в Интернет-версии.

ЧТО РЕАЛЬНО СДЕЛАНО

В настоящее время реализованы следующие версии ИПС "Артефакт":

- Версия, работающая в операционной среде MS DOS в локальных сетях.
- Специализированный поисковый модуль, работающий с базами данных на CD-ROM (лазерные диски "Промышленность России и ближнего зарубежья", "Продукция Российских предприятий", "Мониторинг телерадиоэфира 1991–1997", "Антологии газет "Московские новости 1994–1997" и "Moscow News 1992–1997" и др.).
- Поисковый сервер, обслуживающий WWW-клиентов сети Internet, работающий в операционной среде Windows NT (www.integrum.ru и www.integrum.com)

СПИСОК ЛИТЕРАТУРЫ

1. Бернштейн Э. С., Лахути Д. Г., Чернявский В. С. Вопросы теории поисковых систем. М.: ОВНИИЭМ, 1966.

2. Лахути Д. Г. АИПС с грамматикой и автоматическим индексированием в перспективе развития документальных ИПС // Тез. докл. Всесоюз. науч.-техн. конф. "Проблемы автоматизированной обработки научно-технической информации": Секция II. Ч. 2.— М., 1976 .— С. 44–48.

3. Пархоменко В. Ф., Лахути Д. Г., Мясковский И. Ф. Программное обеспечение документальной АИПС с грамматикой и автоматическим индексированием // Проблемы автоматизированной обработки научно-технической информации.— М., 1976 .— С. 40–42.

4. Лахути Д. Г., Пархоменко В. Ф. Пакет прикладных программ документально-фактографического поиска СКОБКИ-II // Проблемы автоматизированной обработки научно-технической информации: Тез. докл. IV Всесоюз. конф.— М.: ВИМИ, 1983 .— С. 95–97.

5. Пархоменко В. Ф. Система автоматического индексирования документов СКОБКИ ОС ЕС.— М.: МЦНТИ, 1983 .— 79 с.

6. Лахути Д. Г. Автоматизированные документально-фактографические информационно-поисковые системы // Итоги науки и техники. Сер. Информатика.— М., 1988 .— С. 6–79.

7. Пархоменко В. Ф., Поляков Ю. П. Автоматизация технологии процессов обработки и хранения слабоструктурированной разнородной информации и организация эффективных механизмов поиска на базе лингвистической ИПС // Библиотеки и ассоциации в меняющемся мире: новые технологии и новые формы сотрудничества: Материалы II Междунар. конф. "Крым-95".— М., 1995 .— С. 237–238.

Материал поступил в редакцию 18.12.97.