Российские Internet базы данных по объектам промышленной собственности

Беляев Виктор Олегович
Федеральный институт промышленной собственности.
adminm@fips.ru

.

Аннотация

Статья посвящена проблемам разработки Российских Internet БД по объектам промышленной собственности (изобретения, полезные модели, промышленные образцы, товарные знаки) в федеральном институте промышленной собственности Роспатента.

Дана информация о существующих БД, их наполнении, форматах, возможностях поиска и используемом программном и техническом обеспечении. Рассматриваются вопросы лингвистического обеспечения, использования классификационных систем, ссылочного аппарата, безопасности и отказоустойчивости.

1. ЭТАПЫ РАЗРАБОТКИ ИНФОРМАЦИОННО-ПОИСКОВОЙ СИСТЕМЫ

Разработка патентной информационно-поисковой системы в Федеральном институте промышленной собственности России началась в 1997 году.

Основной целью разработки являлось обеспечение поиска информации об основных объектах промышленной собственности России (изобретения, полезные модели, товарные знаки, промышленные образцы).

В начале 1998 года подготовлен необходимый комплекс технических средств и осуществлен выбор и приобретение программного обеспечения.

В конце 1998 года система начала функционировать в экспериментальном режиме. Открыт доступ к БД ограниченному кругу пользователей (в основном с целью тестирования основных функций системы).

В марте 1999 года открыт доступ к системе через INTERNET широкому кругу пользователей. Система сдана в опытную эксплуатацию.

С августа 1999 года доступ к полнотекстовым БД и БД товарных знаков осуществляется на договорной основе. Реферативные БД (на русском и английском языках) предоставляются в доступ бесплатно.

2. ДОСТУП В СИСТЕМУ

Доступ к Российским патентным БД возможен через WEB-сайт WWW.FIPS.RU.

Данный сайт содержит английскую и русскую части. Для осуществления поддержки пользователей системы в каждую из частей сайта введены разделы «Поддержка», позволяющие осуществить взаимодействие с техническим персоналом Федерального института промышленной собственности по вопросам использования системы, а также получить заранее предопределенную информацию технического характера.

Рис.1 WEB сайт www.fips.ru - Home page.

3. ПРОГРАММНОЕ И ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ

В качестве основной технической платформы для реализации системы было выбрано оборудование фирмы Compaq. Сервер баз данных – Proliant 6000 4CPU Ppro200, RAID 5 36Gb & RAID 5 72 Gb, 1340 MB RAM.

В качестве основной программной платформы для реализации системы было выбрано программное обеспечение фирмы Microsoft. Операционная система – Windows NT 4.0.

В качестве поисковой машины программное обеспечение фирмы Excalibur – Excalibur Retrieval Ware 6.6 (с осени 2000 года 6.7). **[4] [5] [6]**

Доступ в Internet возможен через канал ISDN (256 Kb). Провайдер – COMSTAR. Коммуникационное оборудование фирмы Ascend. WEB сервер на основе программного обеспечения Microsoft Internet Information server 4.0.

4. ИНФОРМАЦИОННЫЕ МАССИВЫ

К середине 2000 года подготовлено шесть основных баз данных:

- **Полнотекстовые** базы данных по **изобретениям** (**RUPAT01, RUPAT02, RUPAT03, RUPAT04, RUPAT_NEW**) содержат информацию о более чем 185 тысячах патентов на изобретения с 1994 по 2000 год. Пополнение – ежемесячно. В состав полнотекстового документа входят следующие структурные части: библиографическое описание, реферат, формула изобретения, описание изобретения, чертежи и таблицы.
- **Реферативная** база данных по **изобретениям** (**RUABRU**) содержит информацию о более чем 295 тысячах патентов и заявок на изобретения с 1994 по 2000 год. Пополнение – ежеквартально. В состав реферативного документа входят следующие структурные части: библиографическое описание, реферат, основной чертеж.
- **Реферативная** база данных по **изобретениям на английском языке** (**RUABEN**) содержит информацию о более чем 185 тысячах патентов на изобретения с 1994 по 2000 год. Пополнение – ежеквартально. В состав реферативного документа на английском языке входят следующие структурные части: библиографическое описание, реферат, основной чертеж.
- **Реферативная** база данных **полезных моделей** (**RUABU1**) содержит информацию о более чем 12 тысячах полезных моделей с 1994 по 2000 год. Пополнение – ежеквартально. В состав реферативного документа входят следующие структурные части: библиографическое описание, реферат, основной чертеж.
- База данных **товарных знаков** (**RUTM**) содержит информацию о более чем 100 тысячах Российских товарных знаках с 1991 по 2000 год. Пополнение – 1 раз в два месяца. В состав документа входят следующие структурные части: библиографическое описание (включая описание классов и словесное воспроизведение знака), графическое воспроизведение знака.
- База данных **международных товарных знаков с указанием России (W_RUTM, W_RUTM_NEW)** содержит информацию о более чем 100 тысячах международных товарных знаках с 1978 по 2000 год. Пополнение – 1 раз в месяц. В состав документа входят следующие структурные части: библиографическое описание (включая описание классов и словесное воспроизведение знака), графическое воспроизведение знака. Язык БД – французский (частично английский).

Общее количество документов в БД приближается к 1 миллиону. Что соответствует более 30 Gb дискового пространства (текст+индекс+графика).

Планируется разработка следующих БД :

- ретро БД по Российским изобретениям (с 1924 г. по 1993 год) – полная факсимильная графика, библиографическое описание и реферат.
- БД по опубликованным заявкам (полные тексты),
- БД действующих патентов,
- БД промышленных образцов,
- БД классификаторов (МПК, МКТУ, МКПО) на русском языке **[7] [8]** ,
- специализированных (регистрационных) БД в области химии и генной инженерии.

Если необходимость и реальность разработки патентных БД не вызывает сомнения, то специальные БД в области химии и биотехнологии (даже только в части Российских патентных документов) могут быть разработаны только при условии тесной интеграции со специализированными организациями в этих областях.

Необходимость создания подобных БД очевидна. Они являются наиболее эффективными для поиска и в тоже время наиболее трудоемкими при разработке. При этом общеизвестно, что поиск по текстовой информации в этих областях не дает желаемых результатов.

Существует обширный международный опыт в части разработки и эксплуатации подобных БД (CAS Registry, DARC Questel и т.д.). Однако использование этих БД в России сопряжено с рядом трудностей и финансовая не самая главная из них (стоимость проведения одного поиска по заявке в несколько раз превышает стоимость ее полного рассмотрения). Не говоря о том, что при всей полноте этих БД они могут не содержать части информации из Российских документов и попол-

няться теми же Российскими документами не стой периодичностью и не в тот срок, в который хотелось бы.

5. ПРЕДСТАВЛЕНИЕ ДОКУМЕНТОВ БД

Каждый документ БД хранится в файловой системе в формате SGML **[1].** Растровая графика в формате TIFF gr.4 (для товарных знаков частично в формате GIF и JPEG).

Текстовая часть патентного документа при отображении средствами Internet браузера разбивается (на стороне сервера БД) на составные части: библиографическое описание, реферат, формула изобретения, описание изобретения. При необходимости описание изобретения может быть дополнительно разбито на несколько составных частей. Причем каждая часть документа не может привышать 70 Kb.

Таким образом, достигается приемлемое время реакции при работе удаленных пользователей с системой (что при существующих в России каналах связи немаловажно).

«Малая графика» (химические, математические формулы и т.д.) хранится в формате GIF либо в кодированном виде (до 1998 года) и воспроизводится (если в кодированном виде) при помощи специально разработанного JAVA апплета. Для устранения проблем с печатью больших документов и с целью унификации доступа к документам БД с различных версий Internet браузеров с осени 2000 года вся «малая графика» будет переведена в GIF формат.

Постоянно ведутся работы по совершенствованию форматов представления документов БД.

После появления Microsoft Internet Explorer 5.0 появилась реальная возможность использования для подготовки и просмотра патентных документов метаязыка XML **[2] [3].** И хотя данная технология еще находится в процессе становления, ее нельзя не учитывать. Об этом свидетельствует то, что XML (в отличии от SGML) уже поддерживает большинство крупных фирм-производителей программного обеспечения для Internet, появились XML сервера БД (например TAMINO SoftwareAG). Намечается явная тенденция интеграции XML с технологией реляционных СУБД, хранилищами данных, EDMS, почтовыми системами и т.д..

В тоже время развиваются и специализированные (отраслевые) языки на основе XML: Mathematical Markup language (MML), Chemical Markup language (CML), Bioinformatic Sequence Markup Language (BSML), языки описания векторной графики, финансовой информации. Использование данных подмножеств языка XML позволит обеспечить поиск через Internet таких объектов, как химические и математические формулы, генные последовательности, графические материалы (в части деловой графики) и т.д..

В Федеральном институте промышленной собственности проведены предварительные исследования по использованию XML для патентных документов. Подготовлено несколько полнофункциональных тестовых примеров в различных областях (в том числе в области химии). Тестовые документы включают в себя такие элементы, как химические и математические формулы, символы различных алфавитов, подстрочные и надстрочные индексы, растровую графику.

6. ВОЗМОЖНОСТИ ПОИСКА, ПРОСМОТРА ИНФОРМАЦИИ И ПОЛЬЗОВАТЕЛЬСКИЙ ИНТЕРФЕЙС

6.1. Возможности поиска информации.

Существующая ИПС позволяет:

- выбирать базы данных для проведения поиска (одну или более);
- использовать для поиска документов семантические сети английского и русского языков (на основе английского и русского тезаурусов);
- производить поиск с использованием маскирования терминов запроса, поиск в диапазоне дат и чисел, точный поиск по слову или словосочетанию;
- производить поиск с использованием булевых операторов и операторов контекстной близости;
- производить нечеткий поиск по битовому образу;
- осуществлять контроль расширений терминов, используемых при формулировке запроса и устанавливать весовые характеристики терминов запроса;
- производить поиск по форматным полям (библиографическим данным патентного документа или товарного знака);
- устанавливать ограничение количества документов в выдаче;
- производить поиск на основе предыдущей выдачи;
- производить поиск «по образцу» (на основе найденного документа) *(функция будет доступна осенью 2000 г.)*;

6.2. Возможности просмотра информации.

Существующая ИПС позволяет:

- осуществлять просмотр суммарной информации о документе в списке найденных документов (название, номер, код вида, страна);
- выделять (подсвечивать) термины запроса в найденных документах *(функция будет доступна осенью 2000 г.)*;

- выделять ранее просмотренные документы в списке найденных документов *(функция будет доступна осенью 2000 г.)*;
- просматривать документы в оригинальном формате (например MS Word, Excel и т.д.);
- использовать настраиваемый алгоритм ранжирования документов выдачи *(функция будет доступна осенью 2000 г.)* ;
- сортировать документы выдачи по определенным форматным полям;
- группировать документы выдачи по определенным форматным полям;
- осуществлять раздельный просмотр библиографической информации, реферата, формулы, описания и чертежей патентного документа.

6.3. Возможности манипулирования результатами поиска.

Существующая ИПС позволяет:

- сохранять запросы и результаты поиска;
- создавать и использовать «постоянно действующие» запросы *(функция будет доступна осенью 2000 г.)*;
- создавать и использовать категории (т.е. выбороки из БД по определенной тематике) *(функция будет доступна осенью 2000 г.)* ;
- использовать функцию Expert, т.е присоединение и просмотр информации о персоне, ассоциированной с категорией *(функция будет доступна осенью 2000 г.)*;
- осуществлять печать и сохранение текстов найденных документов и печать списка найденных документов в сокращенном формате.

6.4. Пользовательские интерфейсы.

Для работы с системой можно использовать полный (Expert) или сокращенный (облегченный) интерфейс.

Первый позволяет производить все возможные в системе виды поиска по всем возможным форматным полям, а также изменять поисковые характеристики (количество документов, найденных по запросу, количество маскируемых терминов, правила сортировки и т.д.) и настраивать «лингвистический процессор». Второй позволяет производить поиск только по основным полям, не позволяет менять поисковые характеристики и настраивать лингвистический процессор.

Совершенстование пользовательских интерфейсов и поисковых возможностей системы ведется в двух направлениях: добавление новых функциональных возможностей и улучшение интуитивного восприятия

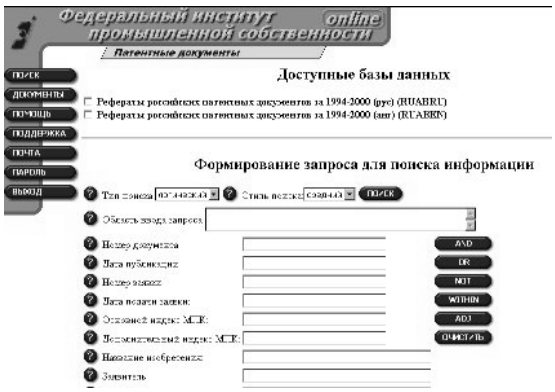

Рис. 2 Сокращенный интерфейс системы.

интерфейса пользователями (т.е. создание более интуитивно-понятного «дружественного» к пользователю интерфейса).

Планируется добавление следующей функциональности:

- обеспечение возможности экспорта найденных документов в различные форматы (HTML, TXT, DOC и т.д);
- обеспеченние возможности подготовки запроса с использованием нескольких предопределенных форм (по выбору);
- обеспечение возможности просмотра найденных документов в нескольких предопределенных форматах;
- обеспечение возможности распечатки найденных документов в нескольких предопределенных форматах;
- обеспечение возможности просмотра статистики о сеансе поиска;
- обеспечене возможности использования «истории поиска»;
- обеспечение возможности просмотра индексов по форматным полям и выбора из них терминов для запроса;
- обеспечение возможности перекрестного поиска;
- обеспечение возможности поиска графической информации по товарным знакам.

7.ЛИНГВИСТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ

В существующей конфигурации система поддерживает английский и русский языки. «Лингвистический процессор» системы состоит из следующих компонент **[5] [6]:**

7.1. Tokenization and Character Mapping

Tokenization – процесс разбиения (на основе набора правил) исходного текста на строки символов (tokens), используемые при индексировании и поиске информации. Существует три типа Tokenization (в зави-

симости от свойств определенных для поля БД): Natural Language Tokenization, Tokens only Tokenization, Custom Tokenization.

Natural Language Tokenization – набор предопределенных правил разбиения исходного текста на строки символов (tokens).

Tokens only Tokenization – все символы текста поля БД (включая пробел) интерпретируются как один token.

Custom Tokenization – пользовательские правила для разбиения текста (для конкретных полей БД). *Например, возможно определить правила обработки поля БД, содержащего номера телефонов в различных форматах, как единый формат хранения в индексе БД, что даст возможность производить поиск без учета различных вариантов написания номеров телефонов.*

Character Mapping – процесс определения символов в строке (token) на основе таблицы соответствия символов. В частности:

На основе данного механизма возможно предопределить (для каждой конкретной БД) правила интерпретации символов текста. *Например, БД товарных знаков требует возможности поиска с учетом «альфа-фонетических замен», настроив соответствующим образом таблицу символов (для этой БД), мы можем обеспечить поиск товарных знаков с различным написанием, но одинаковых (или почти одинаковых) по звучанию.*

7.2. Морфологический анализ и Stemming.

Морфологический анализ – процесс приведения слов к нормальной (словарной форме) на основе словаря основ и соответствующих правил. Дает возможность не учитывать при поиске различных вариантов написаний слов (различных окончаний, суффиксов и т.д.) и резко уменьшает объемы индексных файлов.

В существующей системе для русского языка используется модуль морфологического анализа от фирмы «Медиалингва». К сожалению данный модуль не позволяет изменять (настраивать) процесс морфологичекого анализа.

Stemming – процесс, похожий на морфологический анализ, но более простой, так-как не использует при анализе слов словаря основ.

7.3. Обработка стоп-слов и идиом.

Обработка стоп-слов – определение (на основе словаря стоп-слов) и удаление при индексировании документа (из индекса) и поиске информации по запросу (из запроса) слов и словосочитаний (идиом), не несущих смысловой нагрузки. Данный процесс может быть применен как к тексту документа, так и к определенным «форматным» полям документа, что дает возможность сократить объем индексной информации БД и улучшить время реакции системы на запрос.

Для существующей системы разработан оригинальный стоп-словарь. Планируется проведение работ по его дополнению стоп-идиомами.

Обработка идиом – определение (на основе списка идиом) устоявшихся словосочитаний (фраз). Дает возможность поиска по словосочитанию (фразе) целиком (не по отдельным словам), что улучшает точность поиска и ранжирование документов в выдаче. Предотвращает исключение из идиомы стоп-слов *(например: «Витамин А» – будет проиндексирован как «Витамин А» несмотря на то, что «А» является стоп-словом).*

В существующей системе используется список идиом от фирмы «Весть-Метатехнология». Данный список расчитан на общеупотребительную лексику и не содержит понятий из технических областей. Планируется полностью переработать данный список, пополнив его констукциями, принятыми в научно-технических и патентных документах.

7.4. Семантический анализ.

Семантический анализ (семантическая сеть) – процесс расширения терминов запроса различными вариантами на основе словарей (тезаурусов) с возможностью выбора глубины расширения. Используются следующие виды отношений: акронимы, синонимы, родо-видовые, ассоциативные и т.д.

В существующей системе используются «общие» (не специализированные) тезаурусы английского и русского языка. Данные тезаурусы мало пригодны для поиска научно-технической и патентной информации. Планируется создание ряда отраслевых тезаурусов. Для этого совместно с фирмрй «Алеста» разработано соответствующее мультипользовательское программное обеспечение.

8.ИСПОЛЬЗОВАНИЕ КЛАССИФИКАЦИОННЫХ СИСТЕМ

Другим аспектом, имеющим отношение к лингвистическому обеспечению, является использование классификационных систем. Возможность поиска по индексам МПК (Международной патентной классификации) **[7],** МКТУ (Международной классификации товаров и услуг) **[8],** МКПО (Международной классификации промышленных образцов), предоставляемая поисковой системой, может быть усовершенствована путем их органичной интеграции с поисковыми тезаурусами. Реальной становится задача создания многоуровневого тезауруса, основанного на рубриках классификационых систем с возможностью не только изменения глубины использования семантической сети (как мы говорили ранее), но и динамического изменения содержания тезауруса путем перехода по иерархическим ветвям классификаторов.

В конце 2000 года планируется предоставить пользователям системы возможность поиска по текстам МПК (6 и 7) на русском языке.

9. ССЫЛОЧНЫЙ АППАРАТ

Основной целью использования ссылочного аппарата в системе является придание базам данных дополнительной динамичности. Развитие аппарата гиперссылок (в совокупности с дополнительной функциональностью) можно сравнить с развитой в мире системой обслуживания в отелях - «все включено». Задачей системы становится обеспечить пользователя всем необходимым не выходя из базы данных.

При нажатии на ссылку (помещенную в БД) пользователь может заказать твердую копию первоисточника, просмотреть все документы БД, где упоминается номер заявки или патентного документа, просмотреть текст рубрик классификатора, провести поиск по рубрикам классификатора документа, провести перекрестный поиск в иных БД, получить информацию о фирме заявителя и родственных фирмах, занимающихся данной проблемой с указанием количества выданных им патентов, связаться по электронной почте с экспертами в данной области техники и т.д.. При развитии электронной коммерции осуществить покупку/продажу лицензии.

В 2000-2001 году планируется внедрение в систему прямых и обратных ссылок на рубрики классификационных систем и ссылок на номера патентов и заявок.

10. ЭЛЕКТРОННАЯ КОММЕРЦИЯ

Продвижение Российских патентных БД в Internet привело к необходимости создания программного обеспечения для электронной коммерции.

Для существующей на сайте www.fips.ru системы пока разработаны только отдельные компоненты, реализующие данную задачу. В частности: возможно заполнить и отослать заказы на доступ к платным БД, подготовку электронных копий первоисточников, поиск информации в пакетном режиме и подписаться на новости. Отдельные программы обеспечивают сбор и отражение информации о запросах пользователей и слежение за расходыванием сумм по договорам на доступ к платным БД с автоматической рассылкой напоминаний по E-mail и блокировкой паролей (при полном израсходовании средств).

Планируется обеспечить автоматизацию «полного цикла» электронной коммерции (представление информации – реклама – продажи (в том числе электронные платежи) – послепродажное сопровождение – аналитика). Наибольшие трудности (как и ожидалось) возникают при реализации системы платежей через Internet. В основном из-за неподготовленности Российского сегмента пользователей.

Не смотря на то, что система представлена в Internet с конца 1999 года, уже зарегистрировано более 2500 пользователей (регистрация необязательна), заключено более 200 договоров на доступ к платным БД (в том числе с зарубежными организациями). В первой половине 2000 года средний ежемесячный размер платежей за доступ к БД составил 1000 $ США. За год пользователями проведено в системе более 200000 запросов. Данные показатели могут существенно вырасти после решения проблемы «микро платежей», что даст возможность огромному количеству пользователей нуждающихся в проведении «одноразовых» поисков без труда пользоваться системой и оплачивать доступ.

11. БЕЗОПАСНОСТЬ И ОТКАЗОУСТОЙЧИВОСТЬ

Существующая система обеспечивает защиту БД от несанкционированного доступа по паролю и динамическое шифрование паролей. Права на доступ могут быть определены как для конкретного пользователя, так и для группы пользователей.

Ведутся активные работы по совершенствованию защиты WEB сайта и ИПС от внешнего деструктивного воздействия. Подготовлено оригинальное программное обеспечение для сбора и анализа статистики по использованию ИПС и WEB сайта и мониторинга пользовательской активности. Планируется внедрение программного обеспечения, позволяющего обнаруживать внешние атаки из Internet и реагировать на них в автоматическом режиме (Adaptive security).

Резервное копирование реализовано средствами программного продукта фирмы Computer Assiciates – ARCServeIT. Для ащиты по питанию установлен APC Matrix-UPS 5000. Реализована система автоматической перезагрузки сервера БД и Primary Domain Controllera в случае отключения электропитания.

Список литературы

[1] Using SGML Special edition, Martin Colby & David S.Jakson, QUE 1996

[2] XML in IE5 Progammer’s reference, Alex Homer, WROX press ltd.1999

[3] XML in Action WEB Technology, William J.Pardi, Microsoft Press 1999

[4] Excalibur Retrieval Ware. Query user’s guide.

[5] Excalibur Retrieval Ware. System administrator’s setup guide.

[6] Excalibur Retrieval Ware. System administrator’s reference manual.

[7] Международная патнтная классификация 7-я редакция (WIPO – Роспатент) 2000

[8] Международная классификация товаров и услуг 7-я редакция (WIPO – Роспатент) 2000