АГРАТИН Евгений Георгиевич *- кандидат филологических наук, начальник отдела информационного обеспечения ЗАО "Интерфакс"*

СИСТЕМА КОМПЛЕКСНОГО АНАЛИЗА НОВОСТЕЙ

Каждый хочет, чтобы его информировали
честно, беспристрастно, правдиво – и с
полном соответствии с его потребностями.
Гилберт Честертон

Стремительное развитие информатизации общества привело в последние десятилетия к невиданным темпам роста мирового информационного потока. В этих условиях возможности потребителя по восприятию и обработке информационных ресурсов остаются на ограниченном уровне, он теряется и не в состоянии найти сведения, необходимые в данный момент. Перегрузка информацией столь высока, что это ведет к существенной потере сил и времени для ее отыскания.

Противоречие между возможностями человека и необозримым потоком информации породило проблему «человек-информация», которая решается средствами сбора, обработки, хранения и последующего поиска и распространения информационных массивов. В частности, информационно-поисковой базой данных (БД), понятие которой в системе стандартов по информации трактуется как система, состоящая из одной или нескольких баз данных и системы хранения, обработки и поиска информации в них.

Несмотря на буквально лавинообразный рост полнотекстовых информационно-поисковых систем в России и мире, количество БД, специализирующихся по сбору, обработке и накоплению ресурсов средств массовой информации (СМИ) составляет небольшое число. И причина здесь не только в финансовых затратах и трудоемкости их создания, но и в том, что общество изначально уделяло больше внимания автоматизации научно-технической информации как двигателю технического прогресса, инвестируя именно в это направление как в объективную необходимость. Разница между автоматизацией научно-технической информации и СМИ составляет не один десяток лет, и лишь сегодня этот разрыв начал сокращаться с появлением интернета, который «спровоцировал» переход средств массовой информации в электронный вид и их передачу в поисковые базы данных. И, конечно же, следует отметить, что опыт, накопленный в создании автоматизированных систем в области научно-технической информации, оказал существенное влияние на построение информационно-поисковых баз данных в средствах массовой информации.

В этой связи заслуживает внимания информационно-поисковая система «СКАН» (http://scan.interfax.ru), аббревиатура которой расшифровывается как система комплексного анализа новостей. Ее основные технологические принципы в 2007 г. были реализованы сотрудниками Международной информационной группы «Интерфакс» (Interfax Information Services), которая специализируется в создании информационных продуктов и средств коммуникации, служащих для принятия решений в политике и бизнесе (Рис.1).

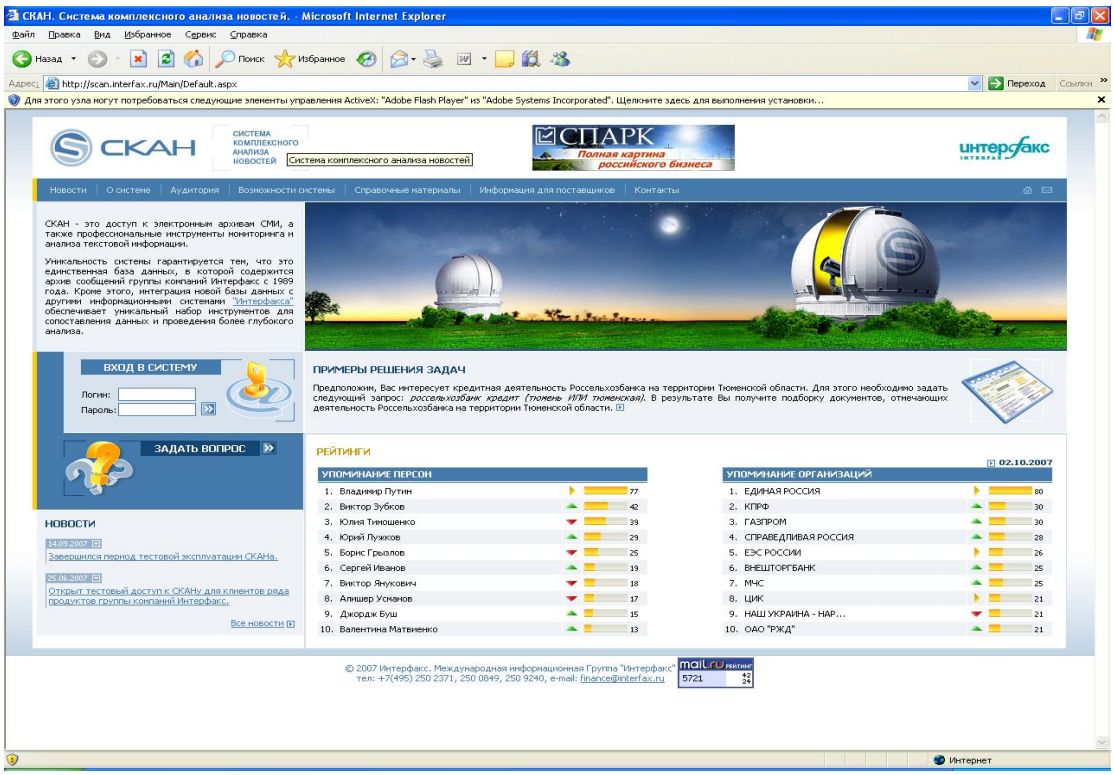

Рис. 1.

В структуру Группы «Интерфакс», объединяющую в настоящее время около трех десятков компаний, входит сеть национальных, региональных и отраслевых информационных агентств, работающих на всей территории России, стран СНГ, а также в Китае и ряде государств Центральной и Восточной Европы. Компании Группы «Интерфакс» выпускают свыше 100 специализированных информационных изданий на русском, казахском, украинском, белорусском, польском, азербайджанском, английском и немецком языках, ориентированных на различные целевые аудитории. Кроме того, разработан ряд уникальных информационных сервисов, основанных на современных IT-технологиях. Одним из таких сервисов является и поисковая база данных «СКАН», которая содержит полные тексты газетно-журнальной периодики, новости информационных агентств, материалы интернет-изданий, мониторинг прессы и телерадиоэфира России, СНГ, Ближнего и Дальнего Зарубежья, включая архив новостных сообщений группы компаний «Интерфакс» с 1989 года.

Созданию информационно-поисковой базы данных «СКАН» предшествовал ряд основных подготовительных работ, которые включали в себя:

- определение информационных потребностей (ИП) пользователей;
- установление тематических границ охватываемых источников СМИ;
- принципы комплектования БД источниками СМИ;
- выбор математического обеспечения, включая язык индексирования документов и запросов.

Изучение ИП – это этап, предшествующий всем последующим действиям информационных органов по информационному обеспечению. Он определяет стратегию поиска и логической переработки информации, возможности качественного удовлетворения ИП потенциальных пользователей. Одна из центральных проблем изучения ИП пользователей – рациональная методология. Оптимально лишь комплексное использование методов, которое только и может дать наибольший эффект. По большинству мнений и имеющемуся ответу – наиболее эффективны здесь методы конкретно-социологических исследований. При создании поисковой базы данных «СКАН» было отдано предпочтение таким методам, как беседа, интервью и наблюдение. Кроме того, была предпринята попытка решить проблему типологической характеристики потребителей информации и типы их информационных потребностей. В нашем случае для решения этой задачи использовалась совокупность признаков, образующих устойчивый тип потребителя и типы его информационных потребностей. В частности, признаками служили тип и направление работы компании и служебное положение ее сотрудника. В первую очередь исследование информационных потребностей пользователей проводилось на имеющихся подписчиках Международной информационной группы «Интерфакс» как потенциальных клиентах базы данных «СКАН».

Среди многих тысяч клиентов «Интерфакса» - ведущие российские и зарубежные средства массовой информации, государственные и правительственные структуры, крупнейшие банки, корпорации, инвестиционные компании и фонды по всему миру. Существенную пользу принес тестовый доступ к системе «СКАН», который позволил собрать большое количество откликов и пожеланий потенциальных подписчиков. В период тестового доступа к базе данных «СКАН» ею воспользовались более 300

корпоративных клиентов группы компаний Интерфакс. В результате были определены основные категории пользователей – это руководство компаний, специалисты в области массовых коммуникаций, PR и информационно-аналитической деятельности, банков и банковского дела, инвестиций, маркетинга и средств массовой информации.

Изучение информационных потребностей потенциальных пользователей позволило установить тематические границы охватываемых источников СМИ. Была представлена группа изданий по следующей тематике:

Общественно-политические и социальные издания:
- центральные;
- региональные

Экономические издания
- финансы;
- банковское дело;
- биржи;
- бизнес, предпринимательство, торговля, рынок;
- безопасность бизнеса;
- бухгалтерский учет;
- маркетинг и менеджмент;
- налоги;
- страхование.

Отраслевые издания, в том числе корпоративные:
- авиакосмическая промышленность;
- транспорт. Автомобильная промышленность;
- военная промышленность. Военные и силовые структуры. Армейская служба;
- деревообрабатывающая промышленность. Лесное хозяйство;
- добывающая промышленность;
- легкая промышленность;
- металлургическая промышленность;
- нефтегазовая промышленность;
- пищевая промышленность;
- полиграфическая промышленность;
- приборостроение. Оборудование;
- химическая промышленность;
- энергетическая промышленность;
- сельское хозяйство;
- архитектура и строительство. Недвижимость;
- связь, компьютеры, оргтехника, наука и техника;
- отдых, туризм, развлечения, здоровье, спорт;

Издания в области государственного управления:
- право;
- законодательство;
- юриспруденция;
- официальные документы.

В системе «СКАН» также предусмотрены территориальный и тематический классификатор. Первый позволяет проводить точечный поиск на основе региональных признаков источника, второй – на основе тематических признаков документа.

Комплектование источников СМИ, которых в базе данных «Скан» в настоящее время насчитывается около тысячи наименований, в обязательном порядке предусматривает соблюдение авторских и имущественных прав. С каждым владельцем изданий заключается авторский договор, по которому автор (правообладатель) обязуется передать Группе «Интерфакс» (получателю) авторское имущественное право на произведение, который, в свою очередь, обязуется использовать произведение, охраняемое авторским правом, в соответствии с условиями договора и уплачивать правообладателю установленное вознаграждение. Здесь важно подчеркнуть, что получение источников СМИ от правообладателей на юридически чистых условиях гарантирует их достоверность, в отличие от интернета, где нередки случаи либо отсутствия автора, либо его отказ от ответственности за достоверность информации, что влечет за собой искажение сведений и дезинформацию.

Среда разработки информационно-поисковой база данных «СКАН» - Microsoft.NET 2.0. Полнотекстовый индекс был разработан на базе Lucene.NET базы каталогов – SQL Server 2005 WEB, приложение – ASP.NET в среде IIS. Что касается морфологии, лингвистики, выделения объектов, индексов объектов, - все это является собственной разработкой сотрудников Международной информационной группы «Интерфакс». В частности, решались две задачи – индексирование документов и индексирование запросов. Таким образом, язык индексирования послужил мостом между языком документов и языком запросов. Индексирование документов с помощью слов, содержащихся в этих же документах, позволяет

пользователю использовать так называемый естественный язык в диалоге с базой данных «СКАН», что упрощает и облегчает задачу отыскания релевантных документов. Поисковая система дает возможность искать информацию с использованием логических выражений, последовательности слов, указывать степень удаленности слов друг от друга в предложении и многое другое. Результаты поиска можно фильтровать и сортировать, используя дополнительные возможности системы. Поисковый механизм учитывает морфологические особенности русского и английского языков, позволяет выделять объекты, людей и компании и использовать найденные объекты для связи с другими источниками информации. Для удобства работы пользователя система хранит последние десять поисковых запросов.

База данных «СКАН» обладает дополнительными сервисами, которые позволяют полнее решать информационные потребности пользователей. В частности, в системе предусмотрен рейтинг упоминаемости организаций и персон в прессе. Проиндексировано свыше тридцати тысяч крупнейших компаний и сведений о персонах. Система ежедневно рассчитывает рейтинг, учитывая все изменения в потоке приходящей информации каждые 5-10 минут. Накопленный информационный архив позволяет получить рейтинг на любую дату, начиная с 1992 года. Данный сервис тесно интегрирован с ядром поискового модуля и всегда позволяет получить подборку документов по персоне или компании, попавшей в рейтинг (Рис.2).

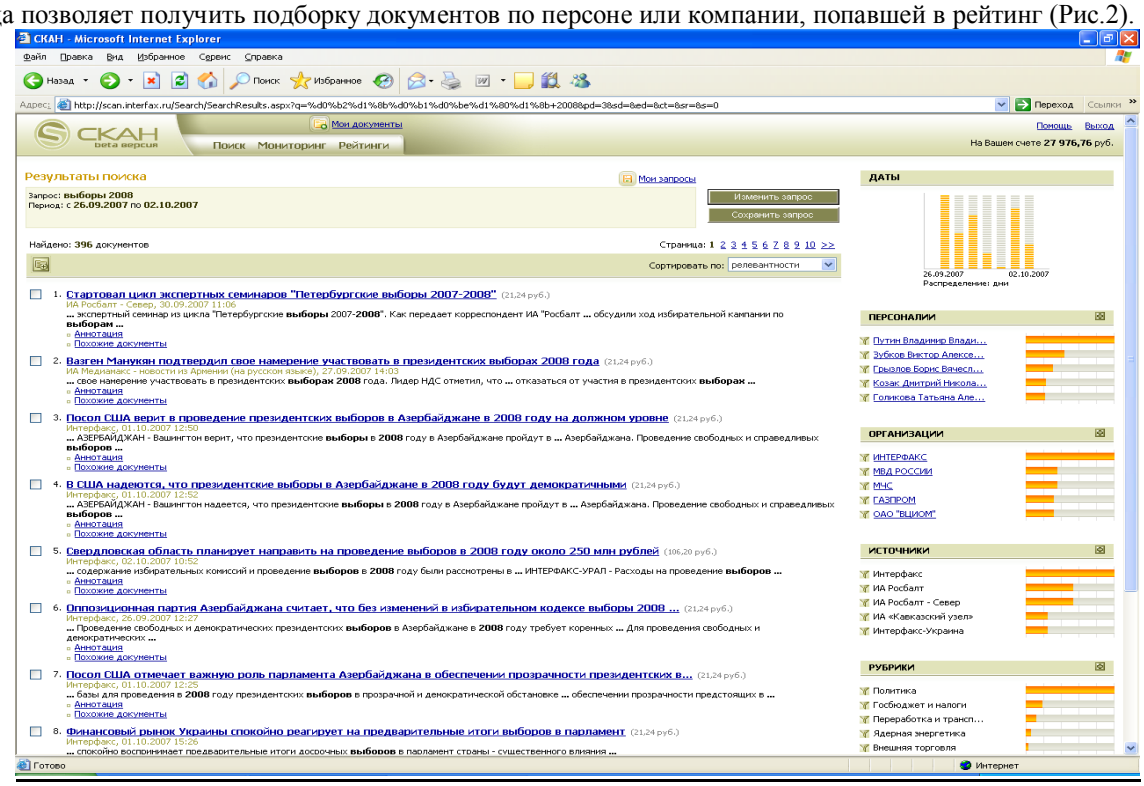

Рис.2.

Еще один дополнительный сервис системы – мониторинг. Этот механизм позволяет пользователю существенно сократить время на выполнение рутинных и ежедневных операций, требующихся при поиске информации в базе данных по заранее определенным и редко меняющимся критериям запроса. «СКАН» позволяет хранить поисковые запросы в системе, выполнять их по заранее заданному расписанию – с ежедневным или еженедельным исполнением. Результаты поискового агента пользователь получает на свой почтовый ящик, что дает возможность находиться в курсе интересующих его событий.

Интеграция базы данных «СКАН» с другими информационными системами «Интерфакса», в частности с базой данных «СПАРК», обеспечивает уникальный набор инструментов для сопоставления данных и проведения более глубокого анализа полученной информации.

База данных «СПАРК» содержит систематизированную и структурированную информацию практически по всем зарегистрированным в России юридическим лицам, а также самим компаниям, включает в себя данные, предоставляемые Федеральной службой государственной статистики, Федеральной налоговой службой, Федеральной службой по финансовым рынкам и другими ведомствами. При создании БД «СПАРК» были использованы новейшие технологии, которые обеспечивают быстрый и удобный доступ клиентов к необходимой информации. Мощный аналитический инструмент позволяет осуществлять исследования, используя весь массив имеющейся информации. В частности, это выборки на основании финансовых и производственных показателей компаний, ранжирование предприятий на основании результатов их деятельности, создание собственных отчетов на базе информации системы «СПАРК». Ранжирование компаний помогает анализировать конкурентную среду, распределение сил на рынках, получать адресную информацию о предприятиях и характеристики целевых групп потребителей и поставщиков. Пользователи могут сортировать компании по любому из показателей годового финансового

отчета. Широкие возможности поиска и сортировки данных по финансовым показателям, взятым из бухгалтерской отчетности, позволяют точно выделить сегмент интересующей части рынка.

Таким образом, обе системы дополняют друг друга, создают информационную среду, которая максимально учитывает информационные потребности пользователей, исключают погрешности анализа и принятие ошибочных решений.

В настоящее время информация, которая находится в интернете в бесплатном доступе, не настолько совершенна, чтобы в полной мере обеспечить информационные потребности пользователей. Именно специализированная информационно-поисковая система «СКАН» может предоставить необходимый пользователям сервис. Главные преимущества системы «СКАН» по сравнению с традиционными сетевыми поисковыми системами следующие:

- оперативность – обновление источников в базе данных происходит в среднем каждые пять минут, в то время как период индексации традиционных информационно-поисковых систем может измеряться неделями;
- доступность ретроспективного фонда – даже если информация удалена с Web-сайта источника, она сохранена в информационном хранилище «СКАН»;
- наличие аналитического инструментария – пользователь может в режиме реального времени не только получать результаты поиска, но и формировать дайджесты, мониторинги, аналитические обзоры, строить сюжетные цепочки, анализировать взаимосвязь событий, динамику понятий и т.д.;
- возможность исключения «дублей» - система осуществляет автоматическую маркировку идентичной по содержанию новостной информации с дальнейшим ее удалением;
- наличие инструментария уточнения запроса – интеграция базы данных «СКАН» с базой данных «СПАРК» обеспечивает уникальный набор инструментов для сопоставления данных и проведения более глубокого анализа полученной информации;
- охват источников – пользователь имеет доступ к информации по интересующей его тематике одновременно из большого числа источников, включая и те избранные, информацию которых ему необходимо просматривать ежедневно;
- расширенный доступ к информации – пользователям системы «СКАН» доступны не только заголовки найденных документов с релевантными фрагментами и аннотацией, но и их полные тексты;
- принцип централизованного интерфейса - пользователь системы, используя сервис мониторинга, имеет доступ к информации из совокупности источников с одного интуитивного интерфейса;
- поисковые возможности – поисковые машины, Web-сайты в отличие от системы «СКАН» не всегда обладают развитыми поисковыми возможностями;
- непрерывно развивающийся список источников – в систему «СКАН» ежемесячно поступает от 30 до 50 наименований источников;
- развитая служба поддержки пользователей – абоненты системы при необходимости могут обращаться с вопросами в службу поддержки и получать исчерпывающие ответы и консультации.

Б И Б Л И О Г Р А Ф И Я

1. Агратин Е.Г. Методика изучения информационных потребностей пользователей // Информационные ресурсы России. – 2002. – №7 – С. 4 – 6.

2. Агратин Е.Г. Коммерция и ГСНТИ // Информационные ресурсы России – 2003. - № 5. – с.15.

3. Агратин Е.Г. Типологическая характеристика потребителей информации // Информационные ресурсы России. – 2005. - № 5. – С. 34.

4. Андреева И.А. Состояние и тенденции развития рынка информационных продуктов и услуг // Информационные ресурсы России. – 1998. – №1. – С. 38, 44.

5. Антопольский А.Б. Мониторинг информационных ресурсов // Информационные ресурсы России. – 2002. – №5 – С. 9 – 16.

6. Дейн М., Слип Ж. Информационная служба в условиях информационного взрыва. – М.: ВИНИТИ, 1972. – 243 с.

7. Сестр Г. Стратегическое значение информации и роль базы данных в маркетинге // Проблемы теории и практики управления. – 1997. – №1. – С.104 – 109.