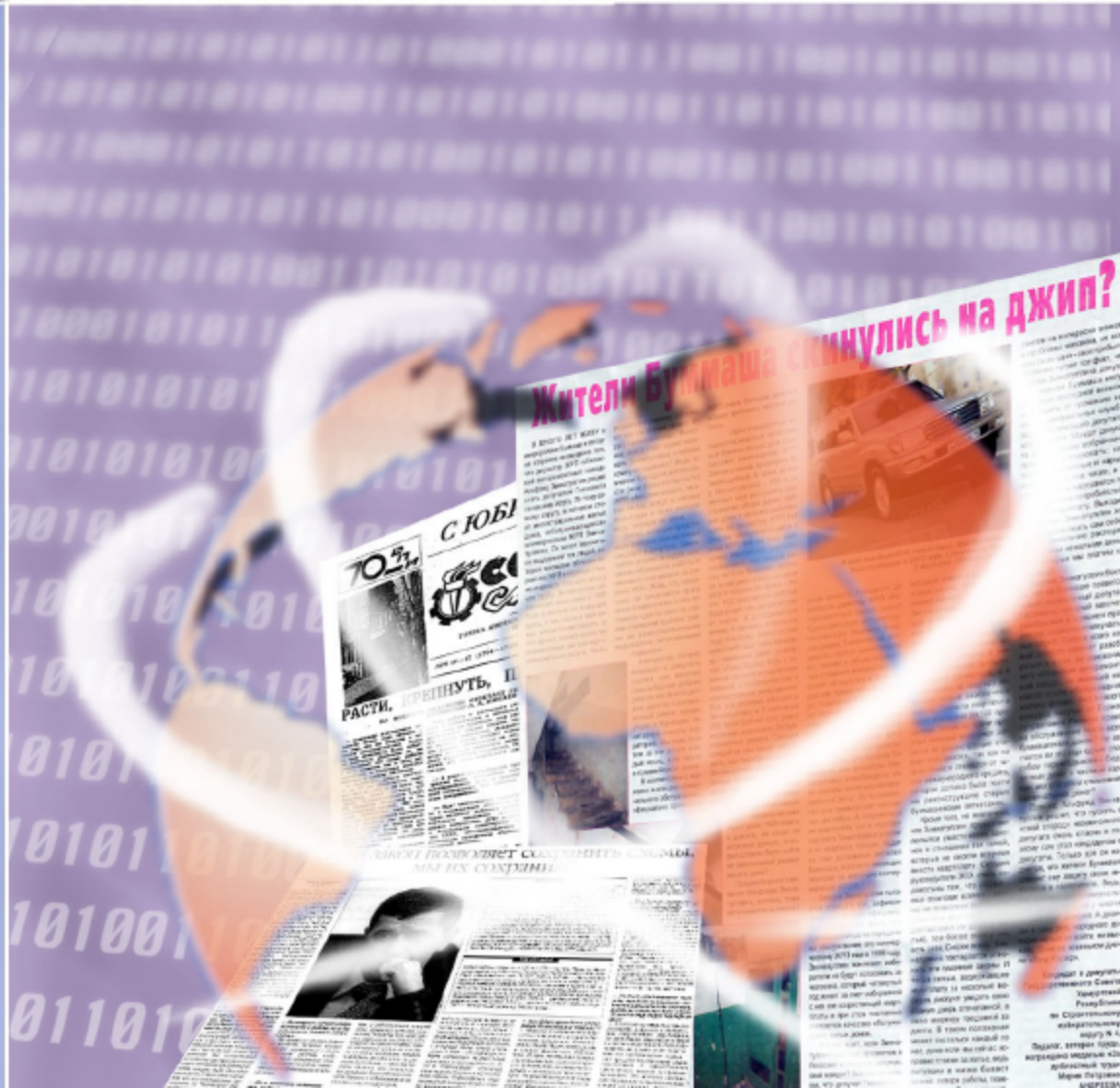
ElVisti
ИНФОРМАЦИОННЫЙ ЦЕНТР

InfoStream

Система мониторинга новостей из Интернет

Методическое пособие



Киев-2005

Система мониторинга новостей из Интернет
/Методическое пособие/, издание 2-е

Редакторы: С.А. Бороденков, Т.Г. Дроботюк, В.Н. Пацьора, О.А. Шевчук

© Информационный центр "ЭЛВИСТИ"
г. Киев, ул. Максима Кривоноса, 2-А. “Internet-офис ElVisti”
Телефон/факс: (380 44) 239-90-91, 247-39-40, 247-39-41
http://infostream.ua
e-mail: stream@visti.net

Назначение данного пособия - ознакомить пользователей (как реальных, так и потенциальных) с системой интеграции новостей InfoStream®*, научить их правилам составления запросов, использованию широких возможностей поиска информации и ее дальнейшей аналитической обработки.

Система InfoStream предназначена для нахождения в сети Интернет новостной информации по интересующим пользователя темам, оперативной доставки результатов поиска, предоставления единого интерфейса доступа к информации с сотен Web-сайтов, и, таким образом, минимизации усилий пользователя на отсеивание дублирующейся информации, шума.

Пособие состоит из четырех основных разделов и трех приложений. Первый раздел включает общее описание системы InfoStream, технологические аспекты решаемых ею задач.

Второй раздел посвящен описанию возможностей интерфейса системы InfoStream, среди которых поисковые возможности системы, средства уточнения запросов, автоматического формирования сюжетных цепочек, дайджестов, построение гистограмм динамики понятий, таблиц взаимосвязей и т.д.

В третьем разделе описаны сервисы и решения, посредством которых реализуются возможности системы для пользователей.

В четвертом разделе описана работа пользователя с системой на примере одного из ее сервисов (он-лайн доступ к оперативным и ретроспективным базам), а также персонализации поискового интерфейса.

В приложениях к данному пособию содержится поясняющая информация - это правила использования языка запросов ИПС InfoReS®* при работе с системой InfoStream, примеры поисковых запросов к новостной базе данных, а также описание возможностей технологии RSS, доступных пользователям системы InfoStream.

За время своего существования система InfoStream обрела широкую популярность и надежную клиентскую базу на украинском рынке. Вместе с тем требования, которые предъявляют пользователи к системе мониторинга новостных ресурсов Интернет, продолжают расти. Это связано как с увеличением информационных потоков (в настоящее время система InfoStream охватывает свыше 25 000 документов в сутки с более чем 1 000 Web-сайтов), так и с необходимостью не только находить документы, но и проводить эффективный анализ результатов поиска.

Для решения этих задач, наряду с развитием информационной базы и поисковых возможностей, был создан новый интерфейс системы InfoStream для предоставления он-лайн доступа к оперативным и ретроспективным базам данных, включающий средства персонализации и содержательного анализа результатов поиска – InfoStream Online.

---

1. Система InfoStream

1.1. Спектр задач

На протяжении последних лет во всем мире и в Украине интенсивно развивается информационное наполнение сети Интернет, что обусловливает появление ряда серьезных проблем, в частности, проблему нахождения в Интернет актуальной новостной информации по необходимой пользователю тематике.

Для решения задач автоматизированного сбора новостной информации из Интернет, ее обработки, систематизации, обобщения и обеспечения доступа к ней в Информационном центре "ЭЛВИСТИ" была разработана система InfoStream (Рис.1).

Рис.1.
Общая схема функционирования системы InfoStream

Система InfoStream предоставляет пользователям широкий спектр сервисных возможностей по обеспечению доступа к новостным ресурсам Интернет. Ее применение позволяет:

- оперативно получать необходимую информацию по мере ее появления в Интернет, анализировать события, своевременно на них реагировать;
- формировать персональные информационные каналы, определяемые запросами на информационно-поисковом языке, формировать архивы для последующей обработки и ретроспективного анализа;
- анализировать поток поступающей информации в режиме реального времени;
- своевременно выявлять тенденции развития и состояния рынков товаров или услуг;
- отслеживать в Интернет информацию о деятельности конкурентов и партнеров, их PR-активности;
- оценивать возможные сферы влияния конфликтных или кризисных ситуаций, осуществлять информационный контроль вероятных источников рисков;
- находить потенциальных клиентов и партнеров.

1.2. Технологические аспекты

Система InfoStream обеспечивает интеграцию сетевых информационных ресурсов на базе эффективных средств сбора, обработки, хранения данных и ор-

ганизации эффективного доступа к ним. С помощью InfoStream выполняется автоматизированный сбор информации с Web-сайтов в режиме реального времени, ее структурирование, группировка по семантическим признакам, а также эффективное тематическое избирательное распределение и предоставление доступа к информационным базам данных в поисковых режимах.

Технологическим ядром InfoStream является полнотекстовая информационно-поисковая система InfoReS. Система InfoStream обеспечивает обработку информации в трех основных режимах, и, соответственно, состоит из трех комплексов:

- избирательного распространения информации (ИРИ);
- интерактивного доступа к полнотекстовым базам данных;
- контент-мониторинга.

Комплекс избирательного распространения информации (Рис.2) позволяет:

- выполнять автоматическое сканирование доступных информационных ресурсов сети Интернет;
- нормализовать информацию, приводить ее к единому текстовому формату;
- автоматически классифицировать информацию, выполнять ее избирательное распространение.

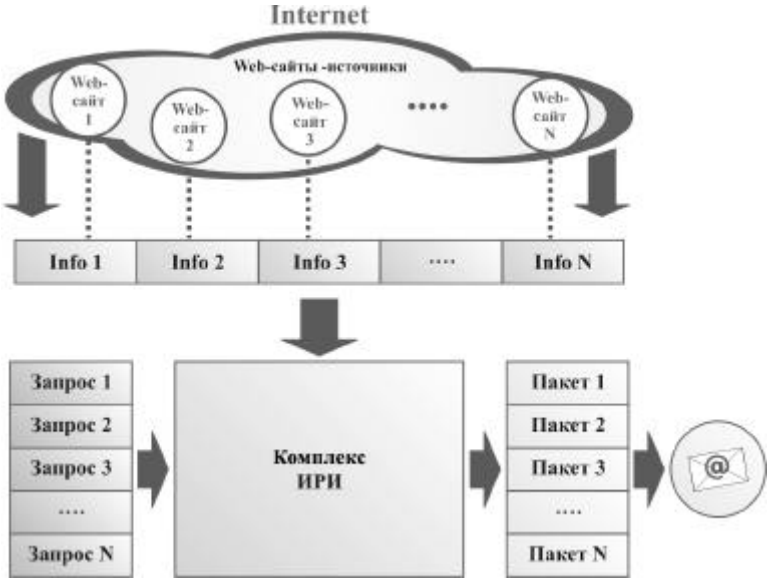


Рис.2. Избирательное распространение информации в системе InfoStream

Комплекс организации интерактивного доступа к базам данных (Рис.3) обеспечивает:

- автоматическое создание оперативных и ретроспективных баз данных;
- интерактивный доступ пользователей к базам данных;
- санкционированный доступ пользователей к базам данных.

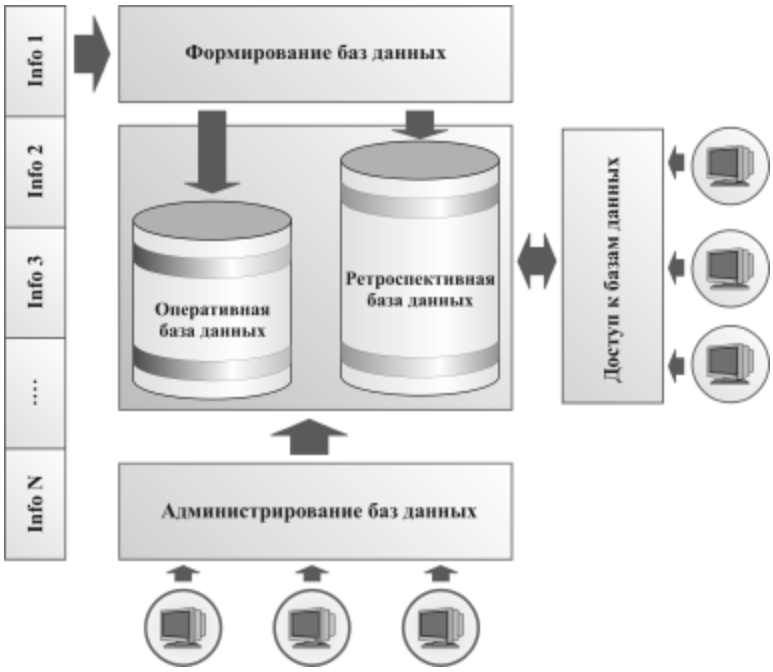


Рис.3. Организация интерактивного доступа к базам данных системы InfoStream

Комплекс контент-мониторинга (Рис.4), базирующийся на технологии Text Mining, обеспечивает формирование:

- информационных портретов;
- дайджестов;
- сюжетных цепочек;
- диаграмм распределения и динамики понятий;
- таблиц взаимосвязей понятий.

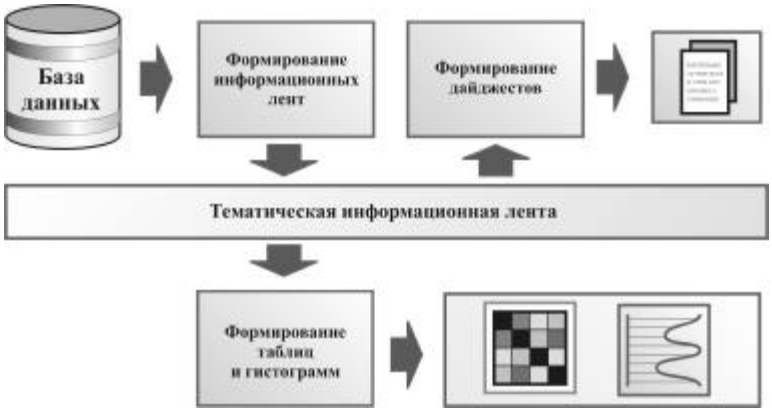


Рис.4.
Комплекс контент-мониторинга системы InfoStream

В настоящее время система InfoStream охватывает мощнейший поток информации, превышающий 25 000 документов в сутки более чем с 1 000 Web-сайтов. Сервер системы InfoStream установлен на площадке ISP ElVisti, одного из ведущих провайдеров в Украине.

2. Возможности системы InfoStream

2.1. Поиск и отображение информации

Поиск и отображение найденной новостной информации – это основные задачи системы InfoStream. Технологическим ядром системы InfoStream является полнотекстовая информационно-поисковая система InfoReS, обеспечивающая поиск информации с использованием логических и контекстных операторов, а также отображение результатов поиска в соответствии с заданными шаблонами.

Алгоритмы обработки документов, поступивших в базу данных системы обеспечивают удобное и максимально информативное отображение найденной по запросу пользователя информации.

Правила и возможности использования языка поисковых запросов при работе с системой описаны в Приложении 1 "Использование языка запросов ИПС InfoReS при работе с системой InfoStream".

2.2. Информационные портреты

Портрет в широком понимании можно рассматривать как модель реального объекта (или субъекта), выраженную его наиболее узнаваемыми чертами.

Построение информационных портретов выполняется на основе эмпирических и частотно-статистических методов, основу которых составляет определение весов отдельных терминов в информационном канале (результате отработки поискового запроса к базе документальных данных).

С помощью информационного портрета визуально можно детализировать и уточнять поисковый запрос, либо конфигурировать персональный информационный канал. В частности, информационный портрет существенно облегчает выбор источников информации, релевантных заданному запросу.

В информационном портрете отображаются такие характеристики массива документов, соответствующих критериям запроса (информационного канала), как:

- рубрики базы данных;
- языки;

- размер сообщений (малый, средний, большой);
- цифровая насыщенность (малая, средняя, большая);
- страны источников;
- названия источников;
- наиболее характерные для данной выборки документов термины (слова).

Все приведенные характеристики ранжируются с учетом их "веса" в информационном канале.

Подробные сведения об использовании информационного портрета приведены в п. 4.5.

2.3. Дайджесты

Дайджест строится на основе алгоритмов автоматического реферирования массивов документов - результатов поиска по запросу. Автоматическое реферирование, как и построение информационных портретов, выполняется на основе частотно-статистического метода. При этом основу его составляет определение весов как отдельных терминов, отдельных предложений и абзацев, так и целых документов.

В программе автоматического формирования дайджестов определяется заданное количество наиболее весомых по статистическим критериям документов, которые берутся в качестве его основы. При формировании дайджеста всегда используются заголовки выбранных документов. Кроме того, в дайджест включаются абзацы, которые имеют наивысшие весовые показатели. В дайджест не включаются дублирующиеся фрагменты. Для каждого фрагмента дайджеста указываются дата его публикации и гиперссылка на первоисточник.

Дайджест представляет собой самостоятельный документ, который можно при необходимости распечатать или сохранить в файле. Вместе с тем электронный дайджест можно также рассматривать как аннотированный источник гиперссылок на документы, лежащие в его основе.

Интерфейс формирования дайджестов при работе пользователя с системой описан в п. 4.7.

2.4. Сюжеты

Функция «Сюжеты» позволяет ответить на вопросы:

- что нового?
- о чем больше всего пишут?

путем семантического ранжирования результатов поиска.

При построении сюжетных цепочек система определяет лингво-статистические характеристики отобранных в результате поиска документов и автоматически выявляет наиболее значимые темы, освещаемые в информационных потоках. Все весомые сообщения группируются по принадлежности автоматически определяемым сюжетам. В качестве названий сюжетных цепочек используются заголовки сообщений, наиболее точно отражающих их суть. Порядок отображения сюжетов определяется количеством сообщений в сюжетной цепочке, что отражает общий интерес к данной теме, и временем публикации сообщений.

При этом составление запроса максимально упрощается - для получения точных результатов вполне достаточно указать одно-два слова, относящихся к необходимой тематике.

Интерфейс формирования сюжетов при работе пользователя с системой описан в п. 4.8.

2.5. Динамика понятий

Форма представления динамики встречаемости понятий - это гистограмма, которая строится как результат информационного поиска по множеству запросов. Эти запросы представляют собой комбинацию ключевых слов, соответствующих понятию, и дат, которые определяют необходимый период времени. Каждая дата в гистограмме является гиперссылкой, ведущей к результату поиска по указанному критерию.

2.6. Взаимосвязь рубрик

Таблица взаимосвязей рубрик строится как статистический отчет, отражающий близость (совместную встречаемость в новостных сообщениях) отдельных понятий реального мира. Это симметричная матрица, элементы которой – коэффициенты взаимосвязей тематических рубрик, соответствующих ее строкам

и столбцам. Эти коэффициенты пропорциональны количеству документов входного информационного потока, которые одновременно соответствуют обеим рубрикам.

С целью выявления блоков - множеств наиболее взаимосвязанных рубрик - применяется алгоритм кластерного анализа.

Как выглядит и используется таблица взаимосвязи рубрик описано в п. 4.10.

3. Сервисы и решения на основе системы InfoStream, предоставляемые пользователям

Существует целый ряд вариантов использования системы InfoStream – сервисов, доступных пользователям. В рамках простейшего сервисного пакета пользователь может подписаться на получение по e-mail ленты новостей по своей тематике, выраженной запросом, имеющим, например, такой вид: **банк&(защит~/1/информаци)**. Данный запрос относится к защите информации в банковской сфере. Получение документов, соответствующих такому запросу, относится, скорее, к общему анализу отрасли, его субъектов и событий. Для анализа деятельности партнеров и конкурентов с помощью системы InfoStream можно подписаться на определяемые запросами информационные ленты, сообщения которых включают названия соответствующих фирм, имена и фамилии, бренды и т.п. Для работы пользователей, нуждающихся в постоянном варьировании запросов, предусмотрен режим он-лайн, который идеально подходит для проведения постоянного контроля, например, источников рисков и конкурентов, оценки состояния рынков и т.д. Для того чтобы избежать информационной "перегрузки", получить наиболее существенные документы необходимой широкой тематики, можно воспользоваться средствами обобщения и уточнения запросов (информационными портретами) или технологиями автоматического построения сюжетных цепочек и дайджестов, реализующих современный подход "глубинного анализа текстов" (Text Mining). Еще один очень важный аспект - это персонализация поискового интерфейса в режиме он-лайн - возможность сохранения запросов пользователями, организации подписки на них.

3.1. Он-лайн доступ к оперативным и ретроспективным базам данных

InfoStream Online - это сервисный пакет, обеспечивающий доступ к базам данных системы в режиме он-лайн. Системой формируются следующие базы данных, доступные пользователям:

- оперативная - содержит документы, поступившие за последние 7 дней. Эта база данных обновляется в режиме реального времени;
- ретроспективная - содержит документы, по-

ступившие за последний квартал;
- оперативная англоязычная – содержит документы, полученные из англоязычных источников за последние 7 дней.
- ретроспективная англоязычная – содержит документы, поступившие за последний квартал;
- табличная – содержит документы со структурированной информацией, например, прайс-листы или котировки.

В рамках этого сервисного пакета предоставляются возможности:
- поиска документов и построения их информационных портретов;
- построения дайджестов;
- построения сюжетов;
- построения диаграмм распределения и динамики встречаемости понятий;
- построения таблиц взаимосвязей рубрик.

3.2. Рассылка новостной информации по e-mail

В настоящее время существуют семь вариантов подписки на получение по электронной почте результатов поиска по запросам:

MiniStream - рассылка информации по одному запросу один или два раза в сутки. Пакет услуг MiniStream предусматривает доставку информации только по одному электронному адресу.

MidStream - рассылка информации по одному запросу один, два или шесть раз в сутки. Пакет услуг MidStream предусматривает доставку информации на один или два электронных адреса.

BizStream - получение информации пользователем по одному запросу один, два или шесть раз в сутки. Пакет услуг BizStream предусматривает рассылку информации на один или два электронных адреса. Кроме этого, пакет BizStream предполагает дополнительно возможность получения информации по второму запросу, состоящему из ключевых слов, определяющих название компании, бренда, персоналии и т.д. в рамках пакета MiniStream, предоставляемого как бонус.

MainStream - рассылка информации по одному запросу в режиме реального времени, круглосуточно, по мере появления информации в Интернет. Рассылка может отправляться по трем электронным адресам и, по желанию, в виде заголовков и анонсов статей размещаться непосредственно на сайте пользователя с помощью встраиваемого JavaScript-приложения в виде ленты новостей.

InfoStream Topics - получение рассылки логически структурированной информации по запросу в виде последовательных, автоматически формируемых сюжетов, каждый из которых представлен наиболее "характерной" для сюжета новостью (приводится полностью) и пятью заголовками подобных публикаций со ссылками на источники. Количество сюжетов определяется объемом документов, относящихся к запросу пользователя (до 20). Пакет представляет идеальный инструмент для оперативного ознакомления с новостями рынков, сфер деятельности, событий, явлений, происшествий и т.д.

InfoStream Ukrainian Day - ежедневная рассылка по одному электронному адресу (утром или вечером) наиболее важных событий дня, произошедших в Украине или касающихся Украины. Рассылка осуществляется в виде сюжетов (см. InfoStream Topics).

InfoStream Rating - получение результатов анализа публикаций по заданным компаниям, брендам, персоналиям и др. Общее предлагаемое количество запросов - 20, результаты предоставляются по первым 10 позициям. Возможно как уменьшение, так и увеличение количества запросов.

3.3. Поток новостей на Web-сайт

WebStream - это специальный сервисный пакет, предназначенный для информационной поддержки Web-ресурсов, который обеспечивает экспорт данных, определяемых запросами пользователей, на страницы их Web-сайтов и порталов. В рамках этого режима информация, соответствующая запросам, в виде заголовков и анонсов статей помещается на страницы Web-ресурсов. Информация экспортируется с помощью встраиваемого JavaScript-приложения. Модери-

руемый администратором Web-ресурсов WebStream представляет собой идеальный инструмент для организации на Web-сайтах пользователей колонок новостей профильной тематики, публикаций об отрасли, компаниях и т.п.

3.4. Сервер InfoStream Port

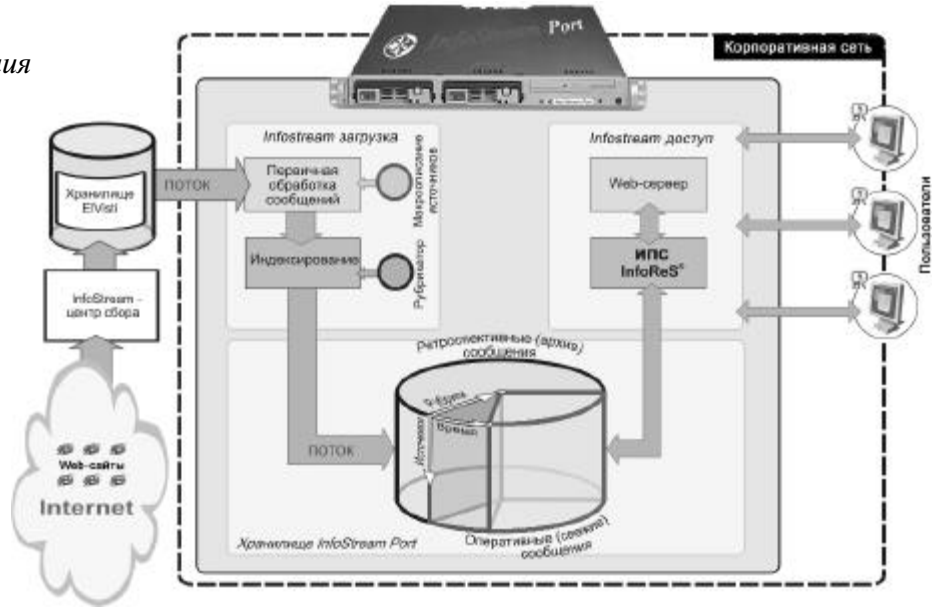
InfoStream Port - это впервые созданное в Украине аппаратно-программное решение - реализация корпоративного новостного поискового сервера, предназначенного для информационного обеспечения компаний разного уровня (Рис.5).

Рис.5.
Сервер InfoStream Port

Информационное обеспечение InfoStream Port строится на основе использования информационного хранилища, формируемого на технической площадке ISP ElVisti в результате последовательности технологических операций:

- сбор информации из сети Интернет;
- нормализация информации, приведение ее к единому формату;
- автоматическая классификация информации;
- помещение данных в информационное хранилище;
- предоставление санкционированного доступа к информационному хранилищу (Рис.6) .

Рис.6.
Принцип функционирования сервера InfoStream Port



Использование InfoStream Port обеспечивает:

- существенную экономию Интернет-трафика;
- формирование и хранение ретроспективных баз данных практически неограниченных объемов;
- интерактивный доступ корпоративных пользователей к базам данных;
- комфортную работу пользователей с неограниченного количества рабочих мест;
- высокий уровень защиты данных;
- экономию затрат на администрирование.

4. Работа пользователя InfoStream Online

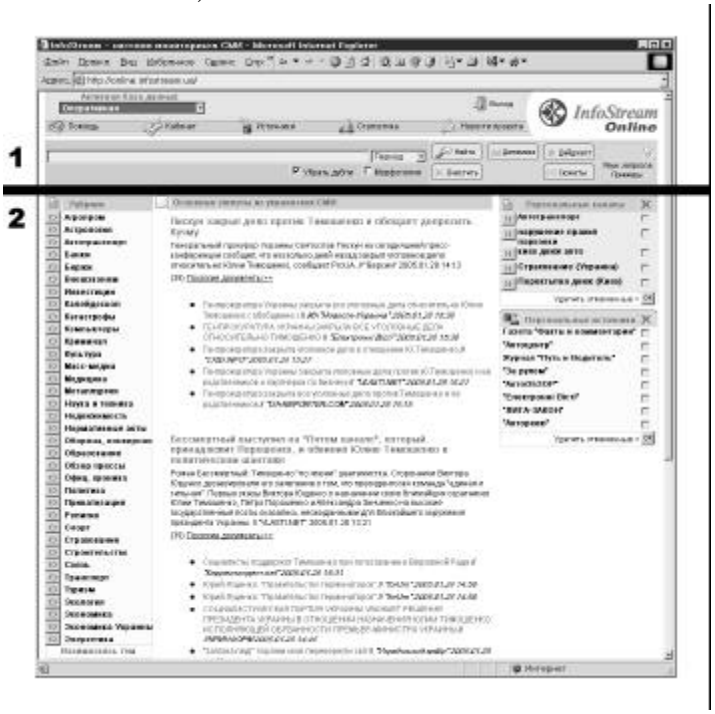
Для работы с системой InfoStream в режиме он-лайн пользователь должен обратиться к серверу системы в сети Интернет по адресу **http://online.infostream.ua** и ввести свои регистрационные данные (логин и пароль)**.** После авторизации пользователь получает в распоряжение удобный, интуитивный интерфейс с возможностью персонализации.

Ниже приведено описание основных экранов интерфейса и предоставляемых пользователю инструментов.

4.1. Основной экран

Первое, что видит пользователь после входа в систему - это "Основной экран". На всех экранах интерфейса системы можно выделить две области:

1) верхняя (основное "Меню системы");
2) рабочая (отображение результатов работы системы).



4.1.1. Меню он-лайн доступа к системе InfoStream

Меню системы (1) постоянно и располагается на всех экранах интерфейса. Это меню предоставляет для пользователя следующие возможности:

Выбор базы данных

Пользователь одновременно может работать (просматривать и производить поиск сообщений) только с одной из баз данных, подключенных к системе.

К системе подключены следующие базы данных:

- оперативная;
- ретроспективная;
- англоязычная;
- англоязычная/ретроспективная;
- табличная.

Выбор активной базы данных производится при помощи селектора, расположенного в левом верхнем углу экранной формы. Каждая база данных, подключенная к системе, предполагает свой вид "Основного экрана" работы с ней.

Так, на основном экране **"Оперативной" базы данных** располагаются:

1) меню системы;
2) форма ввода поискового запроса;
3) перечень рубрик;
4) перечень основных сюжетов из украинских СМИ;
5) перечень персональных информационных каналов пользователя системы;
6) перечень источников, выделенных пользователем как персональные.

На основном экране **"Ретроспективной" базы данных** располагаются:

1) меню системы;
2) форма ввода поискового запроса;
3) календарь выбора интервала дат для поиска документов.

На основном экране **"Англоязычной" базы данных** располагаются:

1) меню системы;
2) форма ввода поискового запроса;
3) перечень рубрик;
4) перечень основных сюжетов из зарубежных СМИ про Украину.

На основном экране **"Ретроспективной/Англоязычной" базы данных** располагаются:

1) меню системы;
2) форма ввода поискового запроса;
3) календарь выбора интервала дат для поиска документов.

На основном экране **"Табличной" базы данных** располагаются:

1) меню системы;
2) форма ввода поискового запроса;
3) рубрики базы данных.

Помощь

Здесь содержится описание системы, интерфейса пользователя, методическое пособие по использованию технологии InfoStream, описание языка запросов, примеры сложных запросов.

Кабинет пользователя

Здесь пользователь имеет возможность:

- изменить настройки системы,
- настроить пользовательский интерфейс,
- просмотреть персональную статистику работы в системе.

Источники

Каталог источников, подключенных в систему. Пользователь имеет возможность объединить отдельные источники в группу "Персональные". Каталог позволяет также отслеживать активность источников.

Статистика поступлений

Представляет статистику поступлений сообщений в базу данных системы.

Выход из системы

Команда выхода из системы.

4.1.2. Рабочая область экрана (отображение результатов работы системы)

На рабочей области экрана располагаются:

Форма поиска (2) позволяет вводить, редактировать, сохранять поисковые запросы, таким образом конфигурируя информационный канал, организуемый посредством системы. Специальными возможностями системы являются автоматическое формирование "Дайджестов электронной прессы" и "Сюжетов".

В качестве информационно-поисковой системы (ИПС) в InfoStream Online используется ИПС InfoReS. Морфология языка запросов описана в

Приложении 2 “Использование языка запросов ИПС InfoReS при работе с системой InfoStream”.

Поисковая форма предоставляет возможность указания критериев запроса (учитывать / не учитывать).

Критерии запроса:

- поисковые термины;
- логические операторы;
- морфология естественного языка (учитывать / не учитывать);
- дублирующиеся сообщения (показывать / не показывать);
- диапазон дат сообщения.

Кроме этого, к критериям запроса относятся термины и операнды из "Информационного портрета" (см. п.4.5.).

В момент, когда результаты обработки поискового запроса системой удовлетворят требованиям пользователя, запрос может быть сохранен для дальнейшего неоднократного обращения к нему. Таким образом осуществляется добавление в систему нового (или сохраняются новые настройки уже существующего) персонального информационного канала.

Перечень рубрик (3)

В качестве тематических рубрик в системе используются предустановленные поисковые запросы. Примеры запросов, соответствующих рубрикам, приведены в Приложении 2. Кликнув на названии рубрики, можно просмотреть полный перечень соответствующих ей документов. "Обзор основных сюжетов" по рубрике - кнопка "(·)".

Перечень основных сюжетов из украинских СМИ (4)

На основном экране системы представлен перечень наиболее популярных сюжетов из украинских СМИ на момент подключения пользователя к системе. Кликнув по ссылке "Похожие документы", можно просмотреть все сообщения, хранящиеся в базе данных и относящиеся к данному сюжету.

Просмотреть сообщение можно, кликнув по его заголовку.

Перечень персональных информационных каналов пользователя системы (5)

Персональные информационные каналы, организованные пользователем, приводятся в "Перечне каналов", в котором представлены:

- "Обзор основных сюжетов" для персонального канала - кнопка "(·)";
- название канала.

Управление организованными каналами:

- просмотр персонального канала;
- просмотр "Обзора основных сюжетов";
- удаление персонального канала производится непосредственно из "Перечня каналов".

Перечень источников, выделенных пользователем как персональные (6)

В "Перечне персональных источников" приводятся:

- название источника;
- кнопка удаления источника из перечня.

Управление "персональными источниками":

- просмотр сообщений из "персонального" источника;
- удаление источника из перечня "персональных" производится непосредственно из "Перечня персональных источников";
- добавление источников в "персональные" происходит из каталога-перечня источников системы (см. п.4.3.).

Календарь выбора интервала дат для поиска документов (7)

Интервал дат для поиска документов в ретроспективной базе данных можно указать, отметив поля выбора соответствующих месяцев года.

Интервал в за весь год можно выбрать, кликнув по номеру года. Также можно использовать максимальный интервал для поиска (выбрав "Весь диапазон дат").

4.2. "Кабинет пользователя"

Здесь, помимо "Меню системы" (см. п.4.1.), пользователь имеет доступ к:

- настройкам персональных информационных

каналов;
- настройкам персональных источников;
- статистике работы пользователя в системе.

Настройка персональных информационных каналов (1)

В момент, когда результаты обработки системой поискового запроса удовлетворят требованиям пользователя, запрос может быть сохранен для дальнейшего многократного обращения к нему.

Таким образом осуществляется добавление в систему нового (или сохраняются новые настройки уже существующего) персонального информационного канала. Сохраняющиеся параметры каналаможно видеть в строке ввода в "форме ввода поискового запроса", а именно, это критерии запроса:

- поисковые термины;
- логические операторы;
- морфология языка (учитывать / не учитывать);
- дублирующиеся сообщения (показывать / не показывать).

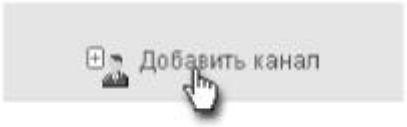
Кроме этого, к критериям запроса относятся термины и операнды, выбранные из "Информационного портрета запроса" (см. п.4.5.):

- рубрики;
- языки;
- размер сообщений;
- цифровая насыщенность сообщений;
- страны источников;
- источники;
- слова.

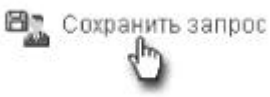
Персональные информационные каналы хранятся и отображаются в виде перечня.

Создание нового канала происходит следующим образом:

- анализируется результат отработки поискового запроса;
- если необходимо, корректируется поисковый запрос (либо на языке запросов, либо при помощи "Информационного портрета") до получения приемлемых результатов;
- параметры запроса сохраняются в системе под собственным именем ("Добавить канал").



Конфигурирование существующего канала производится при его просмотре. Новые настройки можно сохранить для этого же канала ("Сохранить



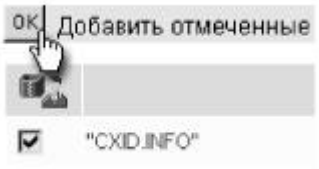
настройки") или добавить новый канал ("Добавить канал") с этими параметрами. В этом случае параметры исходного канала останутся без изменений.

Удаление персонального информационного канала происходит путем выбора чекбокса соответствующего канала в перечне и нажатия кнопки "Удалить отмеченные".

Настройка персональных источников (2)

Любой источник, подключенный к системе, может быть выделен в перечень "Персональных источников". Персональные источники хранятся и отображаются в виде перечня.

Добавление Источника в перечень персональных происходит в "Каталоге-перечне источников". В строке с названием источника необходимо отметить чекбокс и нажать кнопку "Добавить отмеченные в "Персональные источники".



Удаление источника из перечня персональных происходит путем выбора чекбокса соответствующего источника в перечне и нажатия кнопки "Удалить отмеченные".

Статистика работы пользователя в системе (3)

Статистика отображает:

- количество страниц, просмотренных пользователем;
- количество документов, просмотренных пользователем;
- количество уникальных документов, просмотренных пользователем;
- дата первого события;
- дата последнего события за отчетный период.

4.3. Каталог источников

Открытые источники сети Интернет, сканируемые системой, называются "Каталог источников". Все источники для удобства отображения сгруппированы по категориям:

- информационные агентства;
- теле-радио каналы;
- газеты;
- еженедельники и журналы;
- официальные источники;
- ассоциации, компании;

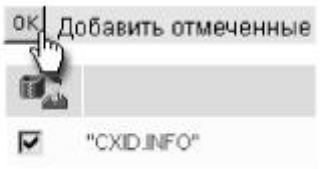
- интернет-издания;
- другие;
- архивы.

Источник описывается следующими полями:

- категория (рубрика в каталоге);
- название источника;
- URL Web-сайта источника;
- страна принадлежности источника;
- язык источника;
- периодичность обновления источником своего содержания;
- дата поступления последнего сообщения данного источника в систему;
- индикатор активности источника.

Каталог является функциональным, т.е.:

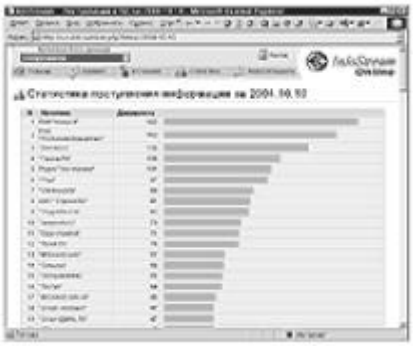
- имеется возможность поиска источника по названию и/или адресу Web-сайта;
- нажав на название источника, можно просмотреть все его документы, хранящиеся в базе данных системы;
- нажав на надпись "www", можно перейти непосредственно на Web-сайт источника в сети Интернет;
- по индикатору активности источника можно судить о частоте обновления источником своего содержания и, соответственно, поступлении документов из него в базу данных системы;
- подсветка строк перечня говорит о принадлежности источника к Украине (желтый тон) и России (синий тон).



Любой источник, подключенный к системе, может быть добавлен в перечень персональных источников пользователя системы. О работе с перечнем персональных источников см. п.4.2.

Источники к системе подключаются администраторами. Пользователь может обратиться к администратору системы с соответствующей заявкой, если не найдет в каталоге необходимого для себя источника.

4.4. Статистика поступления информации

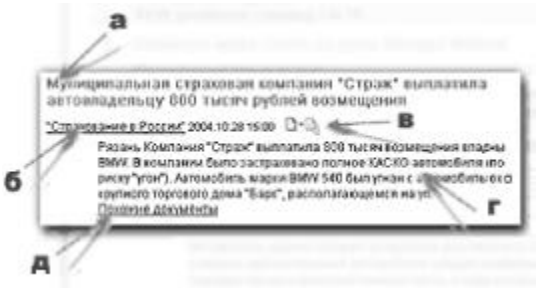


База данных (БД) системы наполняется документами из открытых источников сети Интернет, подключенных к системе. Функция анализа поступлений в БД представляет гистограмму распределения количества документов, которые поступили с указанием количественных параметров распределения по дням.

Нажав на дату, можно просмотреть детальную статистику по поступлениям документов в БД за выбранный день из подключенных к системе источников.

Следующий уровень анализа - детальный просмотр документов из источника. Представление документов аналогично представлению результатов поиска или просмотра персонального информационного канала (см. п.4.5.).

4.5. Результаты поиска (просмотр канала)



Результаты поиска отображаются в виде перечня документов, которые отвечают критериям поискового запроса и включают:

а) заголовок перечня, содержащий:
- название просматриваемой рубрики (если был выбран просмотр рубрики) или поисковый запрос;
- выбранные базы данных и ссылку на ее смену;
- количество найденных в БД документов и страниц выдачи;

б) собственно перечень документов;

в) "Информационный портрет" запроса/рубрики/канала.

Документ в перечне представляется так:

а) название сообщения (является переходом к полному тексту документа);

б) источник сообщения (название источника является переходом на просмотр всех сообщений из данного источника в активной базе данных), дата поступления документа в БД;

в) признак того, что это сообщение дублирует

другое сообщение, поступившее в БД ранее;

г) аннотация сообщения;

д) переход к перечню похожих сообщений в активной базе данных ("Похожие документы").

"Информационный портрет" запроса/рубрики/канала - это модель реального объекта (или субъекта), выраженная его наиболее узнаваемыми чертами. "Информационный портрет" представляет собой множество ключевых слов, наиболее точно отображающих информацию, полученную в результате поиска.

В информационном портрете отображаются такие характеристики массива документов, соответствующих критериям запроса:

- рубрики;
- языки;
- размер сообщений (малый, средний, большой);
- цифровая насыщенность (малая, средняя, большая);
- страны источников;
- источники;
- слова.

Все характеристики ранжируются с учетом их "веса" в портрете. Наиболее весомые (характеризующие большинство документов в результате обработки поискового запроса) имеют обозначение "*" (звездочка).

При помощи информационного портрета можно легко детализировать и уточнять поисковый запрос либо конфигурировать персональный информационный канал.

Для уточнения запроса по информационному портрету достаточно кликнуть на одной из характеристик. При этом уточнение будет производиться с учетом морфологии русского и украинского языков (от слова будет автоматически отделено окончание). Для уточнения запроса сразу по нескольким характеристикам, из информационного портрета, можно отметить несколько чекбоксов "AND" - логическое "И" (слева от слова) или "NOT" - логическое "НЕТ" (справа от слова), а затем кликнуть на кнопке "Уточнить запрос". При уточнении в этом режиме также будет учитываться морфология.

4.6. Просмотр сообщения/документа

Документ открывается в отдельном окне программы просмотра Web-страниц. Его представление включает:

а) название рубрики или запроса, или канала, которому соответствует данный документ;

б) название и полный текст сообщения (в случае непосредственного ввода запроса в "форму ввода поискового запроса" найденные ключевые слова в тексте документа выделяются цветом);

в) команды "Распечатать" (печать на принтере) и "Сохранить" (сохранение текста документа в отдельный файл на компьютере пользователя);

г) название источника, дата поступления документа в БД системы, признак того, что это сообщение дублирует другое, которое поступило в БД раньше;

д) перечень названий документов в активной базе данных, похожих на данный по содержанию;

е) переход к полному перечню похожих документов в активной БД системы.

4.7. Дайджест электронной прессы

Дайджест формируется на основе алгоритмов реферирования массивов документов - результатов поиска по запросу (а). При формировании дайджеста используются заголовки (б). Кроме того, в дайджест включаются абзацы, которые имеют наивысшие весовые показатели (в). Для каждого фрагмента дайджеста указывается дата его публикации и гиперссылка на первоисточник.

Дайджест электронной прессы представляет документ, состоящий из заблаговременно заданного количества фрагментов, который, при необходимости, можно распечатать (г).

Вместе с тем дайджест электронной прессы можно также рассматривать как аннотированный источник гиперссылок на документы, которые лежат в его основе (в).

4.8. Обзор основных сюжетов

Функция "Обзор основных сюжетов" обеспечивает семантическое ранжирование результатов поиска. Все сообщения, соответствующие запросу/рубрике/каналу (а), группируются по принадлежности автоматически построенным сюжетам (б). В качестве названия сюжета используется заголовок сообщения, которое наиболее точно отражает суть данного сюжета (в). Порядок отображения сюжетов определяется количеством сообщений в сюжете (длинной сюжетной цепочки), которое отражает общий интерес к данной теме, и временем публикаций входящих в сюжет сообщений.

Представление сообщений в сюжетах практически аналогично представлению сообщений, выдаваемых в результате поиска, за исключением аннотаций.

Обзор основных сюжетов представляет собой документ, который, при необходимости, можно распечатать (г).

4.9. Динамика понятий

Для динамического анализа встречаемости понятий строится гистограмма распределения сообщений в базе данных системы по датам поступления.

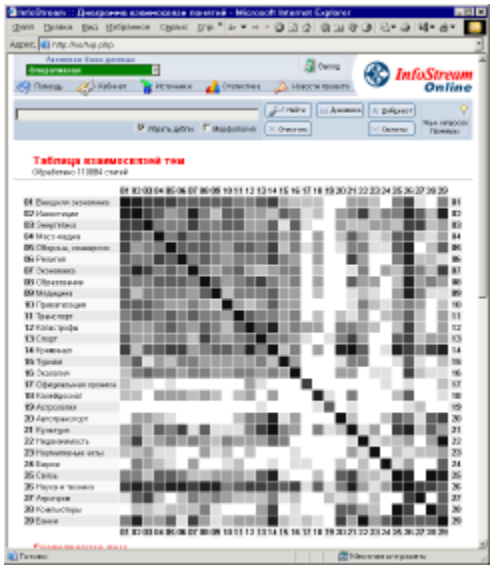
Гистограмма динамики понятий позволяет проанализировать, как изменяются упоминания того или иного понятия, изучить его развитие во времени.

Непосредственно из гистограммы можно перейти к перечню документов, в которых данные понятия встретились в течение выбранного периода.

4.10. Взаимосвязь рубрик

Для наглядности отображения таблицы взаимосвязей рубрик отдельные ее элементы окрашиваются в различные оттенки серого цвета (в зависимости от значений коэффициентов взаимосвязи).

При наведении курсора на ячейку таблицы взаимосвязей рубрик отображается нормированный



вес, а также количество документов, одновременно соответствующих рубрикам столбца и строки данной ячейки.

Каждая ячейка таблицы взаимосвязей рубрик представляет собой гиперссылку, которая указывает на список документов, одновременно соответствующих рубрикам столбца и строки данной ячейки.

Заключение

InfoStream представляет собой систему интеграции новостных ресурсов сети Интернет, охватывающую в настоящее время практически все основные информационные Web-сайты Украины и России. Профессиональное использование возможностей системы InfoStream, доступное пользователям ее сервисов, обеспечивает качественно новые возможности для информационно-аналитической работы в самых различных областях деятельности - от политики, макроэкономики, банковской деятельности - до управления персоналом или индустрии развлечений.

Сегодня системой InfoStream охватывается ежедневно свыше 25 000 документов из более чем 1 000 информационных источников, перечень которых постоянно изменяется, количество постоянно растет. Сведения о новых информационных источниках поступают как непосредственно от разработчика, так и от пользователей сервисов InfoStream, в результате чего реализуется эффективный механизм обратной связи между службой сопровождения системы и пользователями.

Для тех, кто еще не является пользователем системы, но желает ознакомиться с ее возможностями на практике, предоставляется бесплатный тестовый доступ.

Навыки, получаемые пользователями в процессе работы с InfoStream, могут быть использованы при поиске и обобщении информации многих типов (не только новостной) с помощью самых разнообразных поисковых систем. Формализация поисковых предписаний, поиск значимых ключевых слов, учет формальных логических и лингвистических особенностей, использование логических операторов, поэтапное уточнение критериев поиска и многие другие подходы и приемы будут способствовать повышению эффективности информационно-аналитической деятельности в любой области.

ПРИЛОЖЕНИЕ 1

Использование языка запросов ИПС InfoReS при работе с системой InfoStream

Формирование запросов - это искусство, но искусство, доступное каждому. Запросы составляются с использованием определенных правил, называемых в совокупности "языком запросов".

Запрос вводится в область ввода текста поисковой формы и передается поисковой системе при нажатии на кнопку "Поиск". В режимах подписки запросы сохраняются администратором системы в базе данных.

Запросы состоят из термов (слов или их правых усечений) и операторов.

Особенности составления запросов

Òåðì û

Термы - это слова естественного языка или их правые усечения, состоящие как минимум из двух букв. По умолчанию каждое введенное слово воспринимается как основа для поиска, т.е. введя, например, запрос **завод**, можно найти документы, содержащие словоформы: "**завода**", "**заводить**", "**заводы**" и др. При необходимости нахождения точного вхождения слова, при вводе запроса следует добавить к слову символ "]", например: **завод]**.

Система не различает прописных и строчных букв, поэтому для поисковой процедуры запросы **завод]** и **Завод]** равнозначны.

Ñëîâîñî÷åòàíèÿ

Словосочетания - это термы, состоящие из нескольких слов. Для реализации возможности поиска по словосочетаниям используется специальный оператор контекстной близости ADJ (возможно равнозначное написание - "~").

Оператор контекстной близости обеспечивает отбор документов, в которые входят слова, связанные этим оператором. Эти слова должны находиться в документах в указанной последовательности рядом друг с другом. По умолчанию предполагается, что это соседние слова в документе (между ними

отсутствуют какие-либо другие слова). Существует возможность задания в запросе расстояния между словами: /0/ - соседние слова (по умолчанию), /1/ - не более 1 слова в тексте документа между словами; /2/ - не более 2-х слов и т. д. Например, запрос **транспорт~/1/нефти** обеспечивает нахождение документов, в состав которых входят словосочетания "**транспортировка нефти**" и "**транспорт иранской нефти**", в то время, как запрос **транспорт~нефти** позволит выбрать только документ с первым словосочетанием.

Ëîãè÷åñêèå îïåðàòîðû è ñêîáêè

В системе используется следующий набор логических операторов:

· NOT - логическое НЕТ, понимаемое как И-НЕТ;

· AND - логическое И;

· OR - логическое ИЛИ.

При употреблении операторов допускается также их сокращенное написание:

· NOT равнозначно "!" или "^";

· AND равнозначно пробелу или "&" или "+";

· OR равнозначно "|" или "," или ";".

Например, запрос **банк&кредит&украин** равнозначен запросам **банк кредит украин**, **банк+кредит+украин** и обеспечивает отбор документов, в которые входят все три терма – "**банк**", "**кредит**", "**украин**".

Запрос может быть многоуровневым. Различные уровни определяются с помощью круглых скобок. С помощью скобок также рекомендуется выделять термы-словосочетания.

Îïöèîíàëüíûå âîçìîæíîñòè

Язык запросов позволяет использовать в качестве термов определенные сочетания символов, которые могут трактоваться как рубрики, коды источников, даты и т. п.

«Источники»

При поиске по источникам в качестве термов можно использовать правые части соответствующих доменных имен, например, **www.elvisti.com**, **www.lenta** или **4vlada.net**.

«Даты»

Для поиска по датам в базах данных, доступных

в режиме он-лайн, как термы для поиска можно задавать даты в формате ГГГГ.ММ.ДД, например, **2003.06.12**. Допускаются также правые усечения дат, например, если указать в запросе **2003.06.0**, то будут выданы документы с 1 по 9 июня 2003 года.

«Страны»

В системе применяется двубуквенное кодирование стран, к которым относятся сайты - источники информации. Например, для поиска по сайтам, относящимся к Украине, достаточно уточнить запрос термом country.ua, соединив его с остальной частью запроса оператором "&".

«Рубрики»

В запросе, так же как обычные термы, можно использовать коды рубрик. Например, запрос **rubr02&(нбу | (нацбанк~укра)|(нац~банк~укра))** обеспечивает отбор документов по банковской тематике, в которых есть информация о Национальном банке Украины. В качестве тематических рубрик в системе используются предустановленные запросы.

«Морфология»

Режим «Морфология» обеспечивает предварительную обработку слов, входящих в поисковый запрос. В каждом слове отбрасывается изменяемое окончание, что приводит к охвату системой не только слов, но и их словоформ.

Важно, что пользователь всегда имеет возможность как активизировать этот режим, так и отменить его.

«Убрать дубли»

Эта возможность позволяет исключить из результатов поиска сообщения, дублирующиеся не только целиком, но и по смыслу. Выявление дублей на основе лингво-статистических алгоритмов происходит на этапе формирования базы данных системы.

«Подобные документы»

При выводе результатов поиска каждое сообщение дополнено ссылкой «Подобные документы», которая обеспечивает переход к списку содержательно близких сообщений. Содержательная близость, как и смысловое дублирование, выявляется

на основе лингво-статистического анализа сообщений, но отличается более мягкими критериями.

«Насыщенность цифровой информацией»

Пользователю сервиса InfoStream Online доступна возможность указания уровня насыщенности документов цифровой информацией. Эта возможность полезна, например, при поиске аналитических документов, ценовых таблиц, результатов рейтингов и т.п. В системе выделено три уровня насыщенности документов цифровой информацией: высокая (numb.large) - свыше 10%, средняя (numb.medium) - свыше 3% и низкая (numb.small) - до 3%. Обращение к этой возможности, в частности, предусмотрено в информационном портрете.

«Размер сообщений»

Пользователю сервиса InfoStream Online также доступна возможность задания характеристик размеров искомых документов. Эта возможность может быть использована, например, как при поиске объемных аналитических материалов, обзоров, законодательных актов, так и при поиске кратких, насыщенных цифрами таблиц котировок, курсов валют или сводок погоды. В системе предусмотрено три уровня размера сообщений: высокий (leng.large) - свыше 10000 символов, средний (leng.medium) - свыше 1000 символов и низкий (leng.small) - до 1000 символов. Как и в предыдущем случае, эта возможность также отражена в информационном портрете.

ПРИЛОЖЕНИЕ 2

Примеры запросов

Ниже приведены примеры поисковых запросов к системе InfoStream, составленных для получения информации по заданным темам.

Тематика	Запрос
ВР Украины	(верховн~рад) | (парламент~украин) | (парламент~україн)
Экономика Украины	(украи | украї) & (економі | экономи | макроэкон | макроекон)& (бюджет | ввп] | інфляц | инфляц | торгов | бизнес | бізнес | финанс | фінанс)
Инвестиции	(инвестици | інвестиці) & (капіталовкл | капиталовл | інвестор | инвестор | инвестпроект | інвестпроект)
Приватизация	(приватизац | приватизир | реприватиз) & ((гос~имущ) | госимущ | фгиу] | (пакет~акц) | (продаж~/1/акц))
Зерновой рынок	(рын~зерна) | (рин~зерна) | зернов
Недвижимость	(рын~недвижимост) | (продаж~недвижимост) | (торг~недвижимост) | риелт | (агентс~недвижимост) | (ипотечн~кредит)
Маркетинг	(маркетинг~исследован) | (маркетинг~политик) | (маркетинг~кампан)
Политика	(политич | політичн) & (парти | парті | выборы | вибори] | выборов | виборів | оппозиц | опозиці)
Оборона	(миноборон | (минист~оборон) | (вооружен~сил) | всу]) & (военн | (боев~действ) | вооружени | перевооруж | миротвор)
Информационные войны	(информ~войн) & (компромат | дискредит | манипул | сб] | спецслуж | психо)
Экология	(еколог | эколог | довкілля | (окружающ~/1/сред)) & (защита | захист | гринп | greenp | охрана | отход | загрязнен | забруднен)
Туризм	(туризм | туроператор | туристическ | туриндустри | туррын | турбизнес) & (турист | курорт | автотуризм | велотуризм | экотуризм)
Здоровье	(медицин | здравоохран | минздрав) & (врач | лечебн | пандем | болезн | заболев | эпидем)
Наука	(учены | археолог | астроном | космонавт | нкау] | наса] | nasa]) & ((научн~открыти) | космос | спутник | галактик | планет | телескоп | раскопк)
Безопасность в сфере ИТ	(компь | кибер | microsoft | unix | linux) & (сеть | сете | сети | интернет | internet) & (взлом | червь | вирус | антивирус | хакер)
Терроризм	террор | теракт
Нефтепровод «Одесса-Броды»	(одесса~броды) | (одеса~броди)
Анонсы событий	(анонс~/2/событий) | (анонс~/2/подій)
Владимир Высоцкий	(владимир~/1/высоцк) | (володимир~/1/висоцьк) | в.высоцк | в.висоцьк

ПРИЛОЖЕНИЕ 3

Технология RSS и система InfoStream

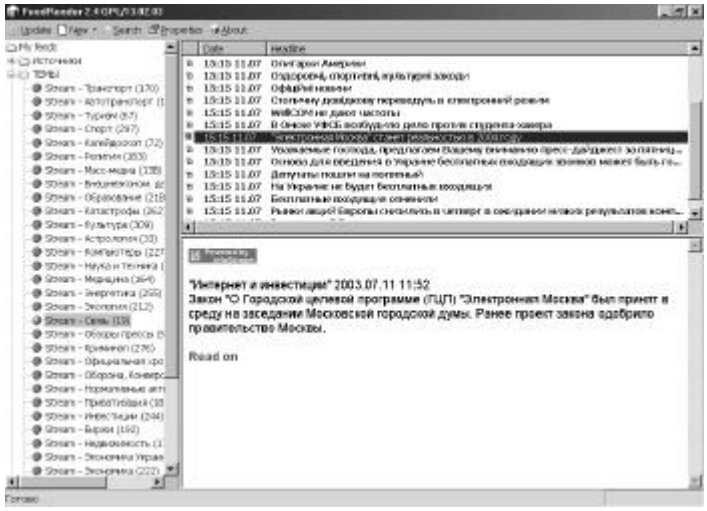
Персонализация на основе технологии RSS

Персонализация интерфейса пользователей, работающих в режиме он-лайн, т.е. сохранение их постоянных запросов и организация подписки на них, может быть реализована на основе современной технологии RSS (Realy Simple Syndication), формата данных и технического стандарта, который обеспечивает интегрированный доступ к новостной информации на Web-сайтах.

Îðãàíèçàöèÿ äîñòóïà êèíôîðìàöèè

Пользователи могут получить доступ к данным в формате RSS с помощью специальных программ, называемых RSS-агрегаторами.

Интерфейс RSS-агрегатора



В качестве RSS-агрегатора рекомендуется использовать FeedReader версии 2.4. (дистрибутив приведен, например, по адресу: ***http://infostream.ua/prg/feedreader24.exe)***

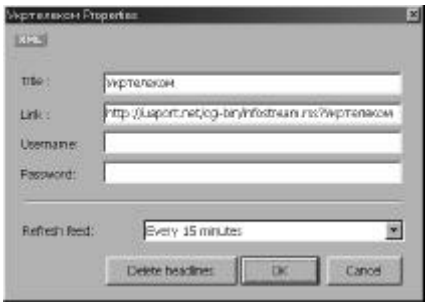
Для получения тематической ленты (RSS-фида) от системы InfoStream в соответствующее поле RSS-агрегатора следует ввести адрес в формате:

http://online.infostream.ua/rss.php[?<ЗАПРОС>]

где в качестве ЗАПРОСа можно ввести слово или словосочетание на языке запросов InfoReS.

Íàñòðîéêà ïîäïèñêè

Для настройки подписки на тематическую ленту (RSS-фид) следует в основном окне системы InfoStream отладить запрос, после чего запустить программу FeedReader, активизировать опцию New и ввести следующую информацию:



Окно подписки

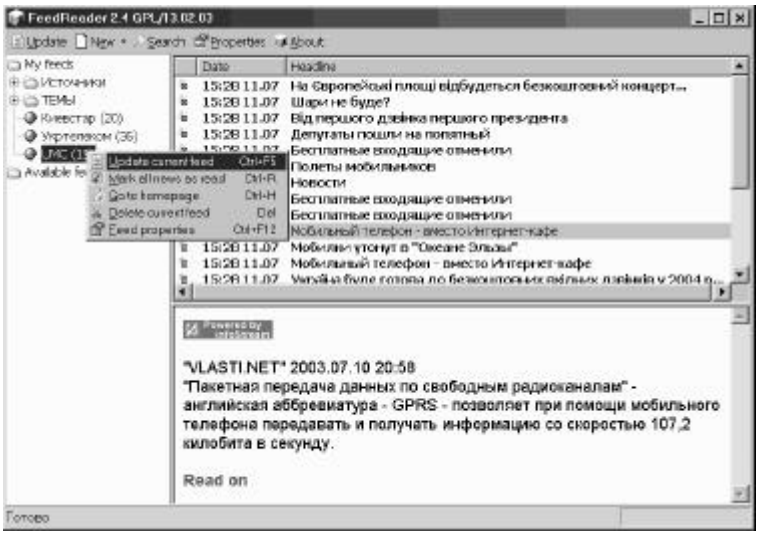
- адрес RSS-фида, включая запрос на информационно-поисковом языке системы InfoReS в формате, приведенном выше;
- название фида, которое может быть определено пользователем;
- периодичность обновления.

Имеется возможность изменения кодировки, размеров шрифтов, помещения фида в отдельную папку, группировки фидов и т.д.

Для управления подпиской в этом режиме существуют дополнительные опции, активируемые для каждого RSS-фида отдельно:

- обновление фида (списка активных сообщений);
- отметка всех сообщений как уже прочитанных;
- удаление списка сообщений;
- изменение свойств подписки, включая тему, периодичность и др.

Опции FeedReader



Для получения полного текста сообщения, заголовок и аннотация которого вызвали интерес, следует:

- произвести двойное нажатие левой клавиши мыши на заголовке,
- нажать на ссылку "Read on" в поле аннотации, или
- нажать на соответствующую кнопку, стоящую перед заглавием, или
- нажать правую клавишу мыши, находясь курсором на заглавии, при этом можно открыть текст сообщения в новом окне Интернет-браузера.