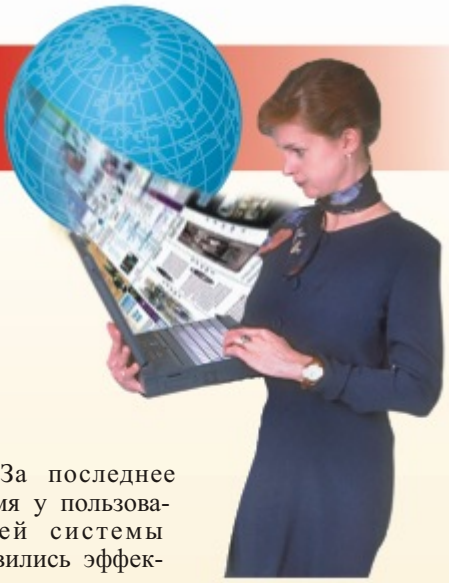
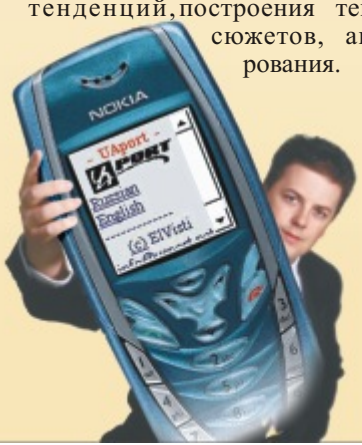
InfoStream



Новые возможности анализа медиа-пространства Интернет

За время своего существования на украинском рынке система InfoStream обрела широкую популярность и надежную клиентскую базу. 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 25000 документов с более чем 700 Web-сайтов в сутки, так и с необходимостью не только получать документы в режиме поиска, но и проводить эффективный анализ его результатов.

Для решения этих задач в системе InfoStream, наряду с развитием информационной базы и поисковых возможностей, были созданы средства содержательного анализа текстовой информации, выявления основных тенденций, построения тематических сюжетов, автореферирования.



За последнее время у пользователей системы появились эффективные возможности выявления смыслового дублирования и поиска подобных документов, уточнения запросов с помощью информационных портретов, управляемого подключения средств морфемной обработки слов запросов и т.д.

Большое внимание в системе InfoStream уделено развитию сервисов для владельцев мобильных устройств, что обусловлено сегодня широким внедрением телекоммуникационной технологии GPRS. На основе системы InfoStream был создан новостной сайт для пользователей режима InfoStream Online - владельцев карманных компьютеров. На этом сайте через лаконичный интерфейс PDA обеспечивается оперативный доступ к ресурсам системы.

Пользователи мобильных телефонов имеют доступ к оперативной информации системы InfoStream через WAP-сайт.

http://wap.uaport.net

Новые возможности системы InfoStream

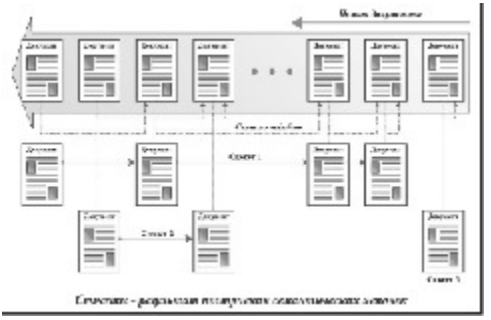
Сюжеты

Функция «Сюжеты» обеспечивает семантическое ранжирование результатов поиска и позволяет ответить на вопросы:

- что нового?

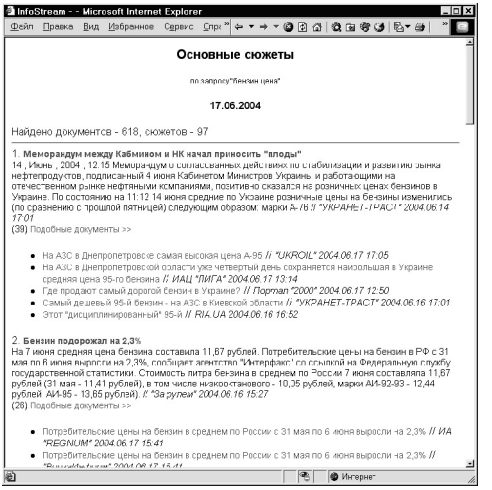
- о чем больше всего сообщений в Интернет?

При построении сюжетов методами “глубинного анализа” текста (Text Mining) выявляется смысл отобранных в результате поиска документов и происходит автоматическое определение наиболее значимых сюжетов, которые освещаются в сообщениях. Все сообщения группируются по принадлежности автоматически определяемым сюжетам. В качестве названия сюжета используется заголовок сообщения, наиболее точно отражающего суть данного сюжета. Порядок отображения сюжетов определяется числом сообщений в сюжете (длиной сюжетной цепочки), что отражает общий интерес к данной теме, и временем публикации сообщений, входящих в сюжет.



Для получения сюжетов строится семантическая сеть

Для того чтобы воспользоваться новой возможностью, любому пользователю режима InfoStream Online достаточно ввести поисковый запрос и нажать на клавишу «Сюжеты». При этом составление запроса максимально упрощается - для получения точных результатов вполне достаточно указать одно-два слова, относящихся к необходимой тематике, например, "банки", "Microsoft" или "бензин цена". Например, при поиске по последнему запросу было выдано свыше 600 сообщений, многие из которых, хотя и имеют в своем тексте указанные слова, относятся к теме лишь косвенно. С другой стороны, этот же запрос привел к построению 97 сюжетов.



Основные сюжеты по запросу “бензин цена”

При этом на первом месте находится сюжет о меморандуме между Кабмином Украины и нефтяными компаниями. Второй по значимости сюжет повествует о повышении цен на бензин в России.

Режим «Морфология»

Режим «Морфология» обеспечивает предварительную обработку слов, входящих в поисковый запрос. В каждом слове отбрасывается изменяемое окончание, что обеспечивает нахождение не только слов из запроса, но и их словоформ.

Например, если в запросе использовано слово “незалежність”, то при включенном режиме морфемной обработки в поиске участвуют все словоформы от

основы “незалежн”. При отключенном режиме «Морфология» по этому же запросу будут найдены документы, содержащие слово “незалежність” целиком.

Морфологической обработке подвергаются все слова запроса, соединенные любыми допустимыми операторами.

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

Режим «Убрать дубли»

Подключение этой возможности позволяет исключить из результатов поиска сообщения, дублирующиеся не только в полном объеме, но и по смыслу. Выявление дублей на основе частотно-весового алгоритма происходит на этапе формирования базы данных системы. Из множества дублирующихся документов оригиналом считается первый по времени, поступивший в систему. Каждому выявленному дублирующемуся документу при помещении в базу данных присваивается соответствующий признак.

Поиск подобных документов

При выводе результатов поиска каждое сообщение дополнено ссылкой «Подобные документы», которая обеспечивает переход к списку близких по смыслу документов. Эти документы, как и смысловые дубли, определяются на основе частотного анализа лексики документа, но отличаются более мягкими критериями подобия.

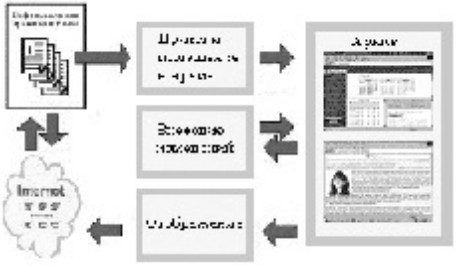
База табличных данных

Система InfoStream расширена средствами мониторинга, обработки и загрузки в базы данных табличной информации с Web-сайтов сети Интернет.

В настоящее время база табличных данных эксплуатируется в тестовом режиме и дает возможность работы с табличной информацией, представленной на целевых веб-сайтах.

Информационно-поисковая система обеспечивает полнотекстовый контекстный поиск, а также поиск с использованием значений рубрик. При обращении к тематической рубрике или по запросу пользователю выдается список наименований таблиц, являющихся ссылками на соответствующие таблицы.

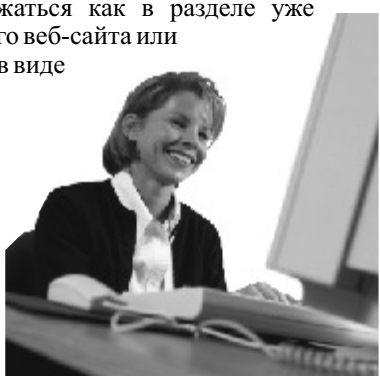
InfoStream Light: персональный архив новостей



Сообщения, отобранные пользователем системы InfoStream, в соответствии с его интересами, могут потребовать организации персонального архива. Для дальнейшего использования информации - ее сохранения, упорядочения, внесения изменений и обеспечения Web-публикации разработана система управления архивом новостей InfoStream Light.

Работа с персональным архивом производится через веб-интерфейс, который обеспечивает комплексное управление содержанием и визуализацией информации.

Помещаемые в архив сообщения могут отображаться как в разделе уже существующего веб-сайта или портала, так и в виде отдельного новостного веб-сайта.



http://pda.uaport.net/

Доступ к новостям владельцев карманных компьютеров

В рамках технологии InfoStream создан специальный веб-сайт для владельцев PDA (PDA - Personal Digital Assistent), который учитывает ограниченность размера экрана этих устройств. На этом сайте обеспечивается доступ к новостям, сгруппированным по тематикам и источникам. С помощью таких устройств, как Palm или Pocket PC, пользователи режима InfoStream Online получают простой, надежный и удобный доступ к новостям.

С развитием информационных ресурсов Интернет вечная проблема поиска информации сегодня получила новое звучание: «поиск информации в неограниченной, неоднородной, динамической информационной среде». Или, другими словами, «поиск иголки в стоге сена».

Традиционные поисковые системы предлагают лишь частичное решение этой проблемы. Им присущи такие недостатки, как низкая оперативность, зависимость от спектра источников, слабые возможности ранжирования результатов поиска.

Система InfoStream решает проблему поиска необходимой новостной информации с учетом таких аспектов, как обобщение данных и их анализ.

Одним из самых перспективных направлений обработки информационных потоков в настоящее время является контент-мониторинг - непрерывный процесс анализа текстовых массивов. Именно непрерывная обработка информационных потоков является самой характерной чертой технологии InfoStream.

ElVisti
ІНФОРМАЦІЙНИЙ ЦЕНТР

Информационный центр "ЭЛВИСТИ"
Адрес: г. Киев, ул. Максима Кривоноса, 2-А, офис 20
Телефон/факс: (380 44) 2399091, 2473940, 2473941
E-mail: stream@visti.net