Использование Personal Avalanche для мониторинга новостей в сети Интернет

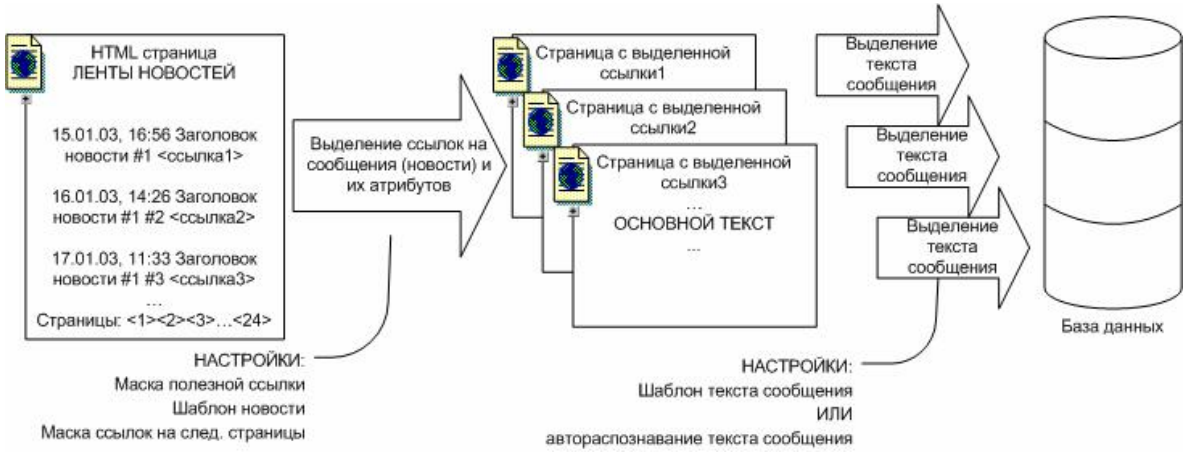
Возможность производить мониторинг Интернет-контента на заданных пользователем ресурсах – одно из основных предназначений продукта Avalanche Personal. Вам предоставляется возможность получать актуальную информацию с любых новостных лент в любое удобное для Вас время, фильтровать её, классифицировать по рубрикам, просматривать, редактировать, экспортировать и многое другое.

Основные концепции

То, что сбор информации происходит с новостных лент, может показаться на первый взгляд некоторым ограничением возможностей программы, но фактически понятие новостной ленты в Avalanche может определяться вашими настройками. Некоторые новостные ленты распознаются программой самостоятельно, практически без вашего участия на этапе настройки, но в большинстве случаев Вам придется определить формат ленты самим.

Под «лентой новостей» подразумевается страница в Интернете текстового формата (например, html), на которой присутствуют повторяющиеся участки html-текста, содержащие изменяющиеся элементы интересующей Вас информации, или ссылки на эту информацию, оформленные в едином стиле. На практике это означает, что практически все однотипные ссылки на странице, участки текста могут играть роль новостной ленты и с нее можно получать новости, или, согласно терминологии программы, «сообщения». То есть если Вы увидели на странице в Интернете список новостей сайта, список пресс-релизов, отзывов, список сообщений на форуме, в гостевой книге – все это при соответствующей настройке может играть роль новостной ленты. При этом в настройках можно также указать откуда брать атрибуты сообщений: заголовок, дата, время, подзаголовок, автор, источник.

Итак, рассмотрим схему – модель работы сбора сообщений с указанных источников:



Порядок работы сборщика сообщений (паука) при анализе одной данной ссылки следующий:

1) Из Интернета достается страница – лента новостей, URL которой Вы задали при настройке.
2) Полученная страница анализируется и в ней выявляются ссылки на документы с текстами сообщений, либо сами сообщения. Вместе со ссылками могут выделяться также и другие атрибуты: заголовок (обычно текст, являющийся ссылкой, или рядом со ссылкой), дата (если есть), время (если есть), подзаголовок. Анализ основывается либо на внутренних эвристических правилах, который в некоторых случаях позволяют интеллектуально распознавать ссылки и их атрибуты, либо на подсказах пользователя в виде маски ссылки, шаблона[1].

3) Паук проверяет каждую выделенную ссылку на предмет того, проходил ли он ее раньше и составляет список «новых» ссылок[2].

4) Паук обходит все новые ссылки, выделяя смысловой кусок текста, который становится текстом сообщения. При этом также используются интеллектуальные алгоритмы, либо пользователем задается шаблон текста сообщения. В этом случае, в отличие от страницы-ленты, алгоритм работает в 70-80% случаев. К тому же, если поиск проводится по разноформатным источникам (например, если новостная лента содержит ссылки на новости, находящиеся на различных сайтах) шаблон задавать бессмысленно. Если формат результатов поиска определен жестко (например, поиск проводится среди новостей одного сайта), то целесообразно задать шаблон текста сообщений для гарантированного результата и увеличения скорости работы системы.

5) Если задана маска ссылок на следующие страницы, то на исходной странице-ленте находится еще не обследованная ссылка, удовлетворяющая маске. В случае если количество уже обработанных лент не превышает значения поля «глубина» в настройках, по найденной ссылке получается «продолжение контента» и обрабатывается аналогично, начиная с п. 2). Если маска на следующие страницы отсутствует, или глубина установлена в «1», то паук прекращает сбор сообщений с этой ссылки (по этому запросу).

Порядок настройки одного ресурса следующий:

1. Добавьте новый сайт в список сайтов паука. При этом Вам нужно указать только название источника, удобное для Вас. Под ним дальше будут фигурировать сообщения, собранные с подчиненных ему ссылок.
2. Вы должны найти в Интернете страницу, на которой находится интересующая Вас новостная лента. Иногда сайты публикуют несколько лент новостей, пересекающихся друг с другом

[1] На практике, для качественного сбора сообщений без потерь рекомендуется задавать шаблон при настройке ссылки.

[2] Вся история пройденных ссылок храниться в файлах вида http__* в папке базы данных. Если Вы хотите, чтобы паук снова обошел уже пройденные ссылки с какого-то сайта, Вам надо удалить соответствующий файл. Удаления собранных сообщений для этого недостаточно.

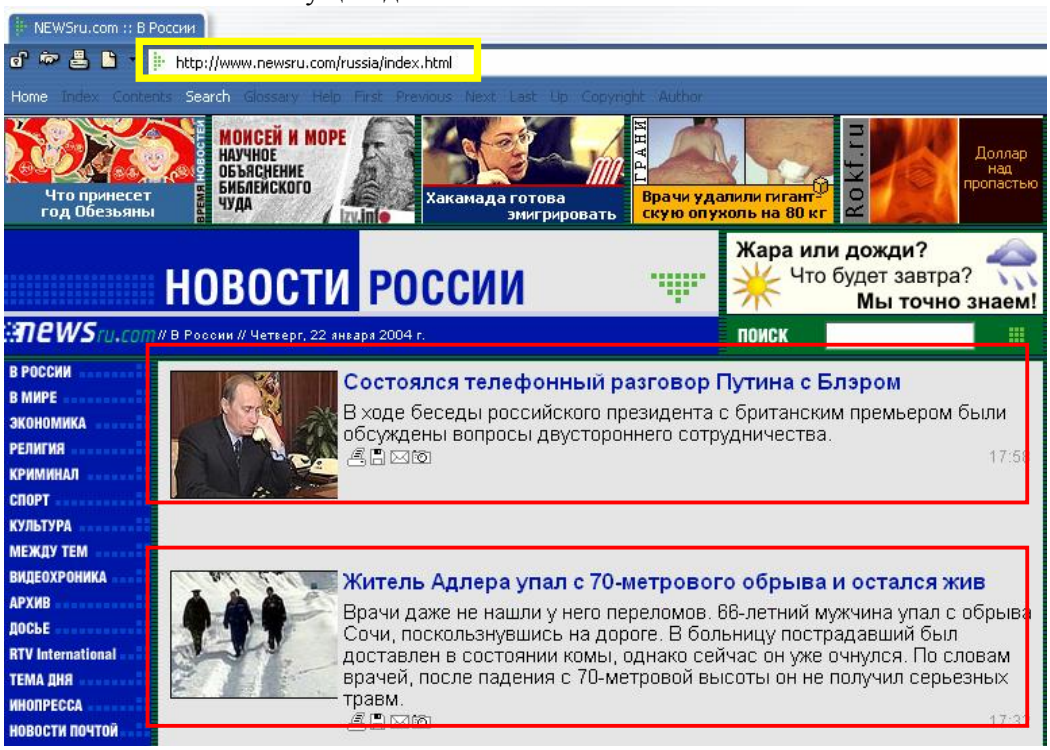
информацией. Первый критерий выбора ленты, конечно, - выбирайте ту, на которой больше всего интересующих Вас сообщений. (Часто на главной странице содержится неполная лента, но есть ссылка вроде «все новости», где содержится полная лента). Во-вторых, если есть возможность, выбирайте ленту, на которой как можно меньше всего лишнего. Так будет легче настраивать. Конечно, посмотрите, чтобы были там и необходимые атрибуты – дата, время, и т.п.

3. Настройте добавленную ссылку согласно документации, справочной информации и описанному ниже примеру настройки.

Пример настройки системы

Рассмотрим пример настройки новостной ленты портала http://www.newsru.com.

1. Добавим новый сайт в Пауке, щелкнув правой кнопкой мыши по корневому элементу списка сайтов «..». Можно ввести произвольное, но узнаваемое имя источника.
2. Открыв главную страницу, можно увидеть, что на не размещены не все новости, но справа есть ссылка на «все новости раздела», на ленте которой содержатся все новости России на текущий день.



3. Желтой рамкой отмечен URL этой страницы, которую мы добавим в качестве ссылки в только что созданный сайт. (Правой кнопкой мыши по сайту, «добавить ссылку». Можно скопировать ссылку из адресной строки в буфер, затем вставить ее в диалог добавления ссылки (Ctrl-V).

Прежде чем продолжать настройку ссылки, познакомимся с уровнями автоматизации настройки ссылок.

Можно выделить три основных этапа в процессе настройки сбора сообщений, которые также можно назвать «уровнями автоматизации сбора»:

1 уровень («автомат») – Вы снимаете отметку «собирать с новостных лент»[3], устанавливаете при необходимости «дата после заголовка», указываете формат даты. Этот уровень не требует знания HTML, достаточно только взглянуть на страницу, чтобы определить эти параметры, которые будут являться подсказками алгоритму распознавания ленты.

[3] В редких случаях на первом уровне сбор начинает работать при включенной «собирать с новостных лент»

2 уровень («полуавтомат»)– Кроме перечисленных выше параметров, вы также указываете маски обрабатываемых ссылок («обрабатывать только ссылки имеющие следующий вид») и, возможно, фрагменты HTML, обрезающие контент ленты (см. «Справочную информацию»). Этот уровень уже требует анализа HTML-кода.

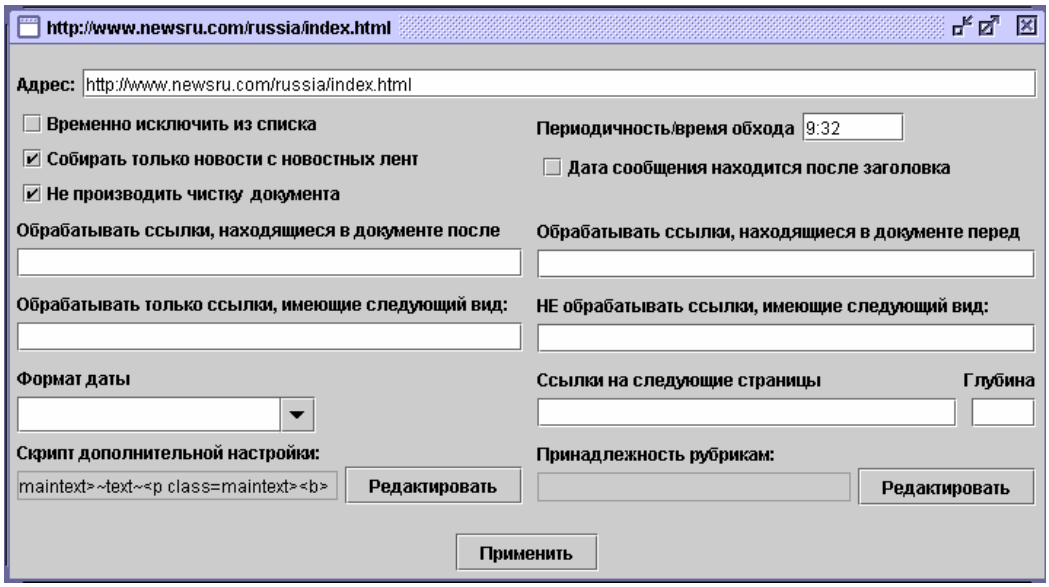
3 уровень («ручной»)– когда Вы задаете шаблон ленты в «скрипте дополнительной настройки». Этот код требует основательного анализа HTML-кода страницы, на предмет поиска подходящего шаблона, который бы охватывал все ссылки и при этом не захватывал лишние.

Чаще всего достаточно второго уровня, но третий уровень повышает качество сбора (например появляется возможность указать где находится дата и время, или заголовок сообщения).

Независимо от уровня вы можете указывать маски ссылок на следующие страницы, глубину и шаблон текста новости.

Шаблон текста имеет смысл применять в следующих случаях:

1. Тексты всех сообщений имеют единый формат (то есть расположены на одном сайте и оформлены в одном стиле)
2. В автоматическом режиме иногда теряется конец или начало текста сообщения
3. В автоматическом режиме иногда забирается не тот текст со страницы
4. Сбор работает очень медленно (более полуминуты на сообщение)



Итак, вернемся к нашей ленте newsru.com. Нас интересуют области, отмеченные красными рамками. Если Вы решили попробовать, распознается ли лента автоматически, определите, где находится дата, время (в данном время находится после заголовка). Формат даты (в данном случае можно не указывать по причине отсутствия даты – все новости сегодняшние). В данном документе мы продемонстрируем самый сложный метод настройки – ручной.

Нужно найти куски html-кода (открыв источник страницы в браузере view/source), соответствующие красным рамкам (используйте поиск в редакторе по слову «Житель»). Возьмем вторую рамку:

```
<table width=100% cellspacing=0 cellpadding=4 border=0><tr valign=top>
<td width=126><nobr><a href=/russia/22jan2004/sochi.html><font color=black><img
src="http://images.newsru.com/pict/id/621106_20040122174403.gif" width="123" height="93"
align="left" border="1" alt="Житель Адлера упал с 70-метрового обрыва и остался
жив"></font></a><img src=/img/b.gif width=1 height=130></nobr></td>
<td width=100%><a href=/russia/22jan2004/sochi.html class=headcolumn>
Житель Адлера упал с 70-метрового обрыва и остался жив </font></a><br>
<img src=/img/b.gif width=11 height=6><br>
<a href=/russia/22jan2004/sochi.html class=explaincolumn>Врачи даже не нашли у него переломов.
66-летний мужчина упал с обрыва Сочи, поскользнувшись на дороге. В больницу пострадавший был
доставлен в состоянии комы, однако сейчас он уже очнулся. По словам врачей, после падения с 70-
метровой высоты он не получил серьезных травм.</span></a><br>
<table widtn=100% border=0 cellspacing=0 cellpadding=2><tr>

<td width=16><a href=/russia/22jan2004/sochi_print.html target=_blank><img border=0
src=/img/ico/2.gif align=top width=16 height=16 alt='версия для печати'></a></td><td width=16>
<a href=/russia/22jan2004/sochi_save.html><img border=0 src=/img/ico/3.gif align=top width=16
height=16 alt='сохранить в виде файла'></a></td><td width=16>
<a href=#
onClick="window.open('/arch/russia/22jan2004/sochi_email.html','mailto','width=400,height=420,re
sizable=1');"><img border=0 src=/img/ico/4.gif align=top width=16 height=16 alt='отправить по
почте'></a></td><td valign="top"><A href=#
onclick="window.open('/pict/big/621107.html','photo','width=360,height=480,resizable=1');return
false;"><img border=0 src=/img/ico/7.gif align=top width=16 height=16 alt='просмотр
фотоиллюстраций'></a></td>

<td width=100% align=right><span class=explaindate>17:32</span></td></tr></table>
```

Нам необходимо создать шаблон новости, выделив:
- гиперссылку, указывающую на текст новости (~url~)
- заголовок новости (~title~)
- время новости (~time~)

```
$NewsShablon <td width=126><nobr><a href=~url~>~something~
class=headcolumn>~title~</font></a>~something~
class=explaindate>~time~</span>
```

Основной принцип такой – находите кусок html-кода перед нужной информацией, кусок - после, удаляете нужную информацию и вставляете вместо нее служебный тег (~url~, ~title~,…). Единственное, за чем нужно следить – это чтобы от новости к новости ваш опорный html-код сохранялся. Если нет – выбирайте другой опорный код, а вместо изменяющегося пишите ~something~.

Установите флажок «собирать с новостных лент» и «не производить чистку». На этом настройка новостной ленты завершена.

5. Создание шаблона страниц самих сообщений полностью аналогично, за исключением того, что обязательным служебным тегом должен быть тег ~text~, вместо которого паук со страницы забирает текст новости. Могут использоваться также и все остальные теги, кроме ~url~. Чаще всего Вам нужно будет найти неизменяемый от новости к новости html-кусок до текста и после текста и сформировать, например, такой шаблон:

```
$TextShablon </table>
<p class=maintext>~text~<p class=maintext><b>
```

Обратите внимание, что при формировании шаблона html-текст должен повторяться символ в символ, как на исходной странице, учитывая переносы строк и пробелы. Если Вы хотите сделать шаблон покороче, используйте

~something~, но аккуратно. «~something~текст» означает «пропустить все, пока не встречу ‘текст’»[4].

6. Итак, в поле тонкая настройка Вы должны получить[5]:

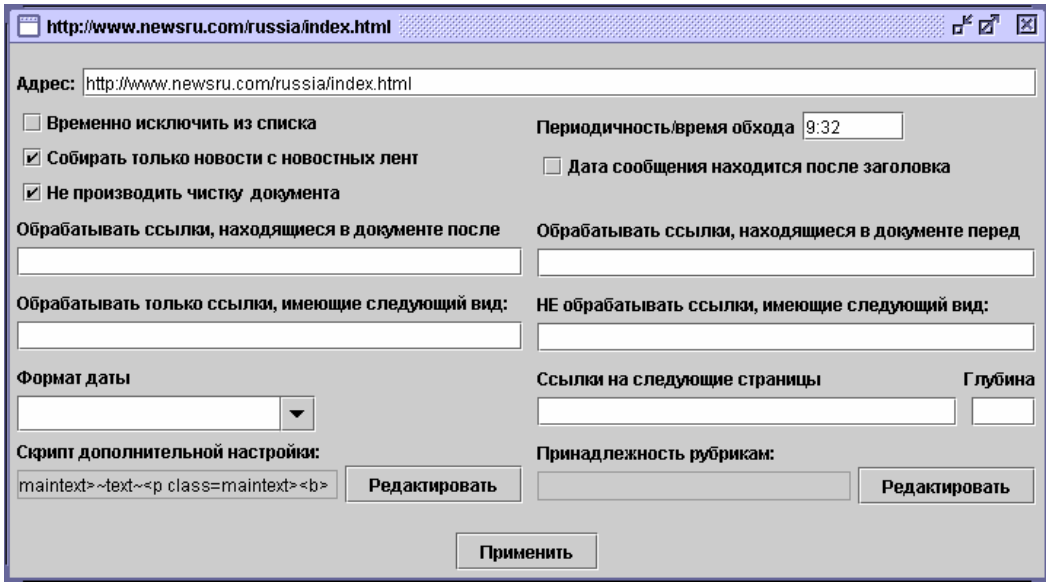
```
$NewsShablon <td width=126><nobr><a href=~url~>~something~
class=headcolumn>~title~</font>~something~ class=explaindate>~time~</span>
$TextShablon </table>
<p class=maintext>~text~<p class=maintext><b>
```

[4] При этом, если произошла неудача, отката и перебора других вариантов не происходит (так шаблоны работают надежнее). То есть используется модель поиска соответствия регулярному выражению без отката (детерминированный конечный автомат).

[5] В силу автопереноса строк при наборе текста возможны лишние переносы.

Справочная информация

Настройка ссылок



- В поле «адрес» можно изменить URL страницы с лентой новостей.
- Можно «временно исключить из списка» данную ссылку, чтобы она не участвовала в сборе сообщений
- «Периодичность/время обхода» задает либо интервал (в минутах), либо время в формате чч:мм, определяющее то, когда в следующий раз будет обрабатываться ссылка в режиме сбора «по таймеру»
- «Собирать новости только с новостных лент» - опция, переключающая алгоритмы распознавания лены новостей. Общая рекомендация – включать опцию, если используется шаблон ленты $NewsShablon, и выключать, если он не используется.
- Иногда «Дата сообщения находится после заголовка». Если это так, укажите эту опцию – возможно это позволит автоматически распознать ленту.
- «Не производить чистку документа» необходимо отмечать практически всегда, особенно если Вы указываете шаблоны.
- «Обрабатывать ссылки, находящиеся в документе после…» позволит отсечь ненужное начало HTML-кода страницы.
- «Обрабатывать ссылки, находящиеся в документе перед…» аналогично отсекает все после встреченного заданного фрагмента.
- «Обрабатывать только ссылки, имеющие следующий вид». Здесь укажите маску ссылки (например, <a class=someclass href=*). Звездочка обозначает произвольную цепочку. При анализе ленты новостей будут обрабатываться только ссылки, соответствующие указанной маске.
- «Не обрабатывать только ссылки, имеющие следующий вид». Аналогично, но удовлетворяющие маске ссылки наоборот не обрабатываются.

- «Формат даты» нужен для облегчения поиска даты и автоматического распознавания ленты пауком.
- «Ссылки на следующие страницы» определяемые задаваемой маской будут рассматриваться как ссылки на продолжение ленты. Формат маски такой же, как и раньше.
- «Глубина» определяет максимальное количество обрабатываемых дополнительно лент, ссылки на которые удовлетворяют указанной выше маске.
- Скрипт дополнительной настройки может содержать элементы $NewsShablon (задает шаблон новостной ленты) и $TextShablon (задает шаблон текста сообщений). Для того чтобы шаблон в встроенном редакторе сохранялся необходимо нажимать кнопку «сохранить» на панели инструментов редактора.
- Принадлежность рубрикам позволяет отметить те рубрики, в которые автоматически будут попадать все сообщения с настраиваемой ссылки.

Формат шаблона новостной ленты

$NewsShablon <значение> - позволяет полностью определять новость на ленте.

Пример:

$NewsShablon <table border=0 cellpadding=0 cellspacing=5>

<tr valign=top><td width=15>**~time~**</td>

<td width=1><img border=0 src="img/2.5/dot.gif" width=1></td>

<td width=615> **~title~** **~text~**</td>

</tr><tr valign="top"><td colspan="2"> </td><td align="right">

<table cellspacing=0 cellpadding=0 border=0 width="100%"><tr valign="top">

<td align="left" width="50%">**~something~**</td><td align="right" width="50%">**~something~**</td></tr></table></td></tr></table>

HTML-код, описывающий несколько новостей на одной странице, практически идентичен, поэтому в HTML- коде необходимо выявить повторяющиеся куски, скопировать его и заменить изменяющиеся участки на соответствующие параметры:

~something~ меняющийся текст который пропускается до слудующего за ним элемента в шаблоне.

~title~ текст, стоящий на этом месте, является заголовком сообщения.

~subtitle~ текст, стоящий на этом месте, является подзаголовком сообщения.

~date~ текст, стоящий на этом месте, является датой сообщения.

~time~ текст, стоящий на этом месте, является временем сообщения.

~url~ текст, стоящий на этом месте, является адресом страницы, на которой находится текст сообщения.

~text~ текст, стоящий на этом месте, является текстом сообщения.

~source~ источник сообщения.

~author~ автор сообщения.

Формат текстового шаблона

$TextShablon <значение> - позволяет полностью определять вид документа, содержащего текст сообщения. Параметры те же, что и в новостном шаблоне кроме **~url~** и **~subtitle~**(их нет)

$# - комментарии.

Внимание! При использовании скрипта тонкой настройки необходимо «не производить чистку документов»!