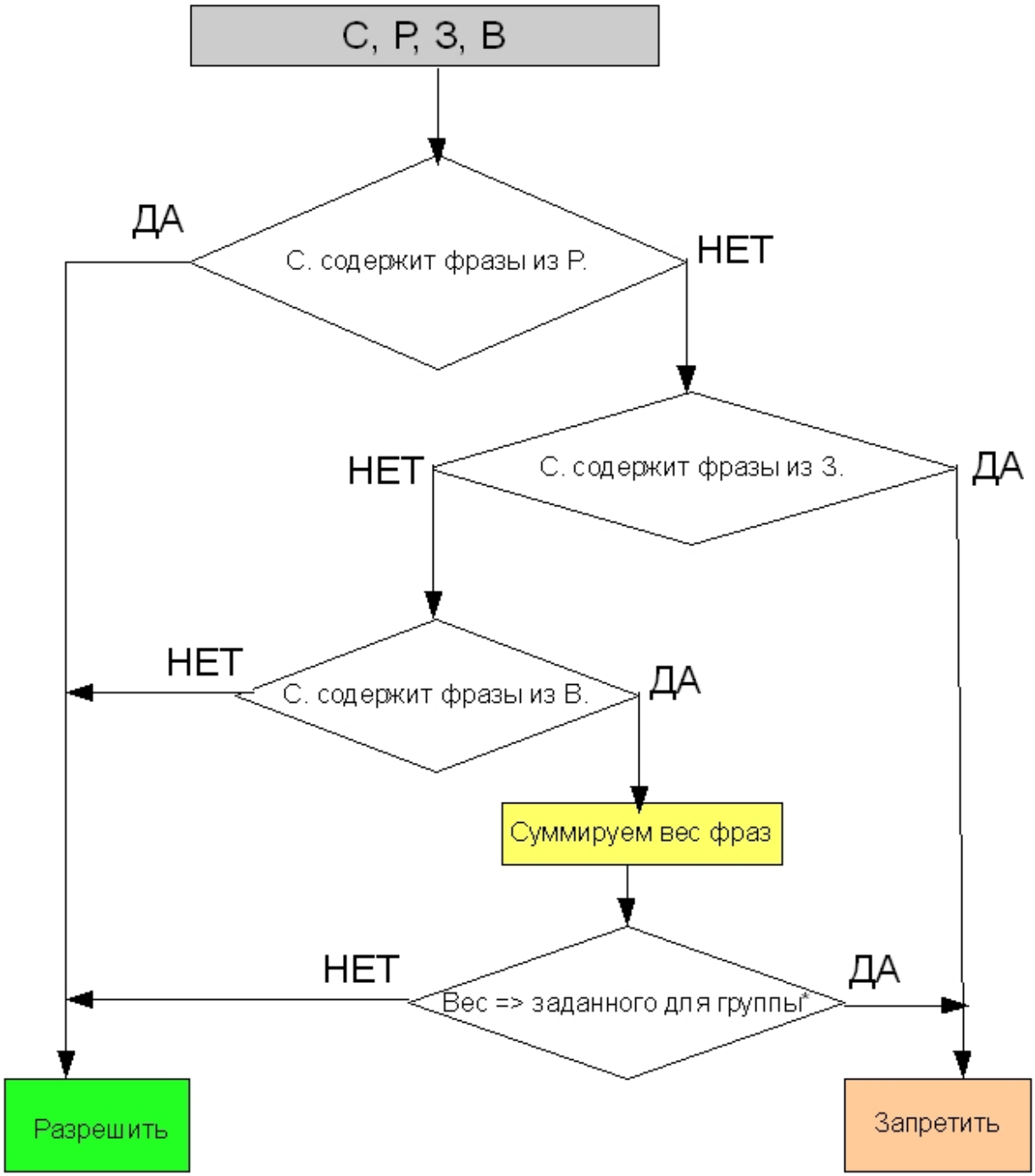
Работаем с файлами фраз для фильтра Dansguardian.

Алгоритм работы фильтрации по ключевым словам Dansguardian.



Обозначения.

С. сканируемая страница
Р. список разрешенных фраз
З. список запрещенных фраз
В. список с весами фраз
=> больше или равно

*** для каждой группы задается свое значение, см. параметр naughtynesslimit в файле dansguardianf1.conf и/или dansguardianf*.conf, где * цифра от 1 до 9.**

Версия Dansguardian - 2.10.0.3
Автор Adminden2007@mail.ru для www.linformatika.ru, 2009

Все файлы со списками фраз находятся в директории phraselists, в соответствующих категории директориях.

Файлы, содержащие фразы с указанием веса (на рисунке это "B."), обычно имеют имена **weighted** и **weighted_***, где вместо * - язык фраз,
например **weighted_russian**.
Они подключаются в файле weightedphraselist (директория lists).

Файлы, содержащие разрешенные фразы (на рисунке это "P."), имеют имена **exception** и **exception_***,где вместо * - язык фраз, или обобщающий критерий,
например **exception_russian**.
Они подключаются в файле exceptionphraselist (директория lists).
Необходимо очень аккуратно и тщательно подбирать фразы для этих файлов во избежание доступности нежелательных сайтов, так как при нахождении фразы из этих файлов, остальные проверки страницы на фразы Dansguardian не делает и страница будет доступна.(см. алгоритм выше)

Файлы, содержащие запрещенные фразы (на рисунке это "З."), имеют имена **banned** и **banned_***,где вместо * - язык фраз, или обобщающий критерий,
например **banned_russian**.
Они подключаются в файле bannedphraselist (директория lists).

Подключение файла с фразами в процесс фильтрации, осуществляется в соответствующем файле (см. выше) директивой **.Include</полный_путь_к_файлу>** ,например
так **.Include</etc/dansguardian/lists/phraselists/porno/weighted_russian>**
Для отключения файла с фразами из процесса фильтрации, достаточно в соответствующем файле закомментировать или удалить строку с директивой его включения.

Формат файлов weighted отличается от формата banned и exception только наличием указания веса фразы, в остальном формат этих файлов совпадает.
Например, строка с фразой в weighted **<порно><50>**, строка в banned и exception указания веса не имеет, например **< министерство >** .

Рассмотрим, **формат файла Weighted**.

Любой файл этого формата начинается со строки, указывающей категорию к которой относятся фразы. Это название категории может отображаться на странице запрета, выводимой клиенту Dansguardian, если в dansguardian.conf процесса фильтрации параметр reportinglevel установлен в 1 или 2.
Для облегчения тестирования, рекомендую на время тестирования использовать reportinglevel=2 при этом в странице запрета, указываются какие фразы были найдены и общий их вес.
Строка должна быть первой в файле и имеет вид
#listcategory: "название_категории"
Например,
#listcategory: "porno_RU"

Далее следуют фразы.
Каждая фраза состоит из одного или нескольких ключевых слов.
Каждое ключевое слово заключается между < и >.
Во фразах может использоваться логическое "И" выраженное запятой между ключевыми словами.
В конце ставится вес фразы в десятичном виде. Это число также заключается между < и >.

Примеры.
< эротик ><20>
< фото>,< Семенович >,< скачать ><50>

Есть одна маленькая тонкость.
Дело в том, что следующие строки абсолютно по разному работают:
<порно><30>
< порно><30>
<порно ><30>

< порно ><30>

Почему?
Тут все очень просто. Символ пробела перед > или после < указывает на точность совпадения ключевого слова.

Рассмотрим примеры с ключевым словом "порно".

<порно><30>
под действие данной фразы попадет и **о*порно*-двигательный**, и ***порно***, и ***порно*графия**, и **с*порно***.

< порно><30>
под действие данной фразы попадет только ***порно*** и ***порно*графия.**

<порно ><30>
а эта фраза присуммирует к весу 30, найдя только **с*порно и порно***, да-м не такой эффект нам нужен.

< порно ><30>
собственно эта фраза добавит к весу 30, найдя только **порно**.

Как видно самый оптимальный вариант данной фразы **< порно><30>**

Но отдельные слова не всегда отрабатывают как хотелось бы.
Вот для таких случаев и используется логическое "И".
Например, **<фото><10>** может срабатывать и на пейзажах, и на порнофото и т.д.
Тогда лучше использовать
< порно>,<фото><40>
Т.е. **порно*фото***, **порно и любительское фото ню** будут учтены, а **фотография бурого медведя** нет.

Нежелательно использовать более 3-х ключевых слов в одной фразе, объединенных «И», так как это замедлит работу фильтра.

Вес фраз может иметь и отрицательное значение. Хорошие фразы имеют отрицательное значение.
Оно уменьшает общий вес при суммировании и снижает вероятность блокирования страницы.

В dansguardian.conf есть важные опции, влияющие на процесс фильтрации по фразам, но об этом в следующих материалах.

Материал создан adminden2007@mail.ru для сайта **www.linformatika.ru** с целью помощи школам в фильтрации контента сети Интернет.

Прим. автора. Все данные проверены опытным путем на Dansguardian версии 2.10.0.3 и Linux Mandriva 2009.1, в более старых версиях Dansguardian может быть все по-другому.
При возникновении проблем с фильтрацией в разных кодировках, смотрите материал
http://linformatika.ru/content/reshenie_problemy_filtratsii_saitov_v_raznykh_kodirovkakh_dansguardian
Новые версии программы и документация к ней всегда доступна на оф.сайте Dansguardian www.dansguardian.org.