

CNSearch 1.5.1

© 2008 CN-Software

Оглавление

Часть I	Общая информация	5
Часть II	Установка	7
1	Обновление версии	7
Часть III	Работа с системой	9
1	Индексация	9
2	Дефрагментация	11
3	Поиск	11
Часть IV	Настройка системы	15
1	Индексатор	15
	search.conf	15
	Список параметров	15
	Использование прокси-сервера	22
	Поддержка поиска с учётом словоформ	23
	Стоп-слова	24
2	Модуль поиска	24
	cnsearch.conf	25
	Параметры конфигурирования	26
	Настройка шаблонов	27
	Использование различных шаблонов	30
	Поиск по выбранным сайтам	31
	Группировка результатов поиска по сайтам	34
Часть V	Дополнительные возможности	36
1	Оптимизация поиска	36
	Оптимизация скорости поиска	36
	Оптимизация размера индексного файла	37
2	Статистика	37
3	Плагины	39
	Создание плагина	40
4	Стандарт блокирования поисковых роботов	41
	Введение	41
	Функция	41
	Структура	42
	Примеры	43
5	Использование META-тэгов "Robots"	43
6	Автоматическая генерация Google Sitemap	44

Часть



1 Общая информация

CNSearch представляет собой поисковую систему для Web-сайтов с расширенными возможностями. Программа проста в установке, настройке и управлении и обладает следующим набором функций:

- Поиск по HTML, TXT, PDF, DOC, RTF, XLS и MP3 файлам.
- Поиск по одному или нескольким сайтам.
- Автоматическая генерация Google SiteMap.
- Выделение найденных фраз.
- Сортировка по релевантности или дате.
- Изменяемая логика запросов - И, ИЛИ, КОМБИНИРОВАННАЯ.
- Поддержка морфологии русского языка
- Сбор статистики по поисковым запросам
- Легко и на 100% изменяемый внешний вид страницы с результатом поиска.

Для работы CNSearch необходим лишь доступ к '/cgi-bin'; база данных не требуется.

Часть



2 Установка

Для установки CNSearch выполните следующие действия:

1. Скачайте с [официального сайта](#) версию CNSearch, предназначенную для Вашей операционной системы;
2. Распакуйте архив. Архив содержит три типа файлов:
 - **Модуль индексации** - находится в каталоге /os/indexer/. Модуль индексации должен быть скопирован в каталог, не доступный из интернета.
 - **Модуль поиска** - находится в каталоге /os/frontend/. Модуль поиска должен быть скопирован в каталог /cgi-bin/ веб-сервера
 - **Страница с поисковой формой** - находится в файле /index_example.htm. Эта страница должна быть скопирована в каталог веб-сервера, предназначенный для размещения HTML страниц.

***Примечание:** Индексатор может функционировать как на веб-сервере, так и на локальном компьютере имеющим другую операционную систему. В этом случае необходимо скачать два дистрибутива для разных операционных систем и взять индексатор из одного дистрибутива, а модуль поиска из другого.*

[Обновление версии](#)

2.1 Обновление версии

Для обновления версии следует выполнить следующие действия:

1. Скачайте с [официального сайта](#) новую версию CNSearch;
2. Скопируйте файлы индексатора новой версии поверх старых.
3. Создайте новый поисковый индекс - произведите повторную индексацию.
4. Удалите старый поисковый индекс.
5. Скопируйте файлы модуля поиска новой версии поверх старой.
6. Скопируйте новый поисковый индекс на место старого.

***Примечание:** В случае если в новой версии потребуются дополнительные опции для установки, они будут описаны в прилагающемся к версии руководстве.*

Часть



3 Работа с системой

Поисковая система CNSearch Pro состоит из модуля индексации и модуля поиска. Индексатор производит анализ сайта (или группы сайтов) и создает индексные файлы (индекс); модуль поиска осуществляет быстрый поиск по созданному индексу.

[Индексация](#)

[Дефрагментация](#)

[Поиск](#)

3.1 Индексация

Для запуска индексации следует выполнить следующие действия:

1. Предварительно в файле search.conf необходимо указать условное название задачи (Job) и адрес сайта, на котором будет проводиться индексация:

```
search.conf [----] 0 L:[ 1+ 0 1/ 9] *(0 / 215b)= [
[Job localhost]
[Index]
URL          http://localhost/my_site
Statistic    Append
Charset      ByHTTPHeader
MaxFiles     10000
StopWordsFile stopwords.txt
Exclude      search/,mail/,.zip,.gif,.jpg
```

1Help 2Save 3Mark 4Replac 5Copy 6Move 7Search 8Delete

Задание параметров индексации

В данном примере localhost - название задачи, а http://localhost/my_site - адрес Вашего сайта.

2. Запустить файл indexer.exe в командной строке, указав следующие параметры:

- Название задачи;
- Имя конфигурационного файла и путь к нему (в случае если файл расположен в другом каталоге).

Пример для Windows:

```
C:\indexer.exe localhost
```

или

```
C:\indexer.exe --config=D:\www\search.conf localhost
```

Пример для Unix/Linux:

```
./indexer name_of_task
```

или

```
./indexer.exe --config=/home/www/search.conf name_of_task
```

Для осуществления индексации нескольких сайтов следует в файле search.conf указать адреса данных сайтов в рамках одной задачи:

```
search.conf  [----]  0 L:[ 1+ 0  1/ 22] *(0 / 490b)= [
[Job localhost]
[Index]
URL          http://127.0.0.1/
Statistic    Append
CharSet      ByHTTPHeader
MaxFiles     10000
StopWordsFile stopwords.txt
Exclude      search/,mail/,.zip,.gif,.jpg
[Job disk]
[Index]
URL          /usr/home/www/
Statistic    Append
CharSet      ByMetaTag
Extensions   html,htm
MaxFiles     100
[Job spam]
[Index]
URL          http://127.0.0.1/
CharSet      ByMetaTag
MaxFiles     10000
ShowURL      No
ShowEmail    Yes
1Help  2Save  3Mark  4Replac  5Copy  6Move  7Search  8Delete
```

Задание индексации нескольких сайтов

В случае если копия Вашего сайта размещена на Вашем компьютере, возможна локальная индексация файлов сайта с последующим переносом индексного файла на сервер (подробнее см. [search.conf](#)).

По завершении процесса индексации система создает следующий комплекс индексных файлов:

По завершении процесса индексации система создает следующий комплекс индексных файлов:

- **files.cns** - описание всех документов сайта;
- **index.cns** - собственно индексный файл;

- **docs.cns** - перечень всех текстов сайта;
- **fulltxt.cns** - полнотекстовый индекс;
- **stats.log** - статистический отчет (для сохранения полученной информации в базе данных; подробнее см. [Статистика](#)).

Примечание: На данный момент в системе реализованы два типа индексации

- **HTTP-режим** - стандартный метод поиска по сайту, находящемуся непосредственно на веб-сервере. Для запуска HTTP-индексации следует указать URL сайта в конфигурационном файле (подробнее см. [search.conf](#)).
- **Индексация локального диска** - дополнительный вариант, предусмотренный для индексации файлов копии сайта, хранящейся на локальном диске Вашего компьютера. Доступ к сети Internet не требуется. Для запуска локальной индексации следует указать в конфигурационном файле URL и расширения файлов сайта (подробнее см. [search.conf](#)).

3.2 Дефрагментация

Для оптимизации процесса поиска по индексным файлам в системе предусмотрена возможность дефрагментации индекса: результаты индексации систематизируются, что позволяет значительно ускорить последующий поиск. Для этого следует скопировать файл `idefrag.exe`, расположенный в каталоге индексатора, в папку с индексными файлами (`index.cns`, `docs.cns` и `files.cns`) и запустить на исполнение. По окончании процесса файл `docs.cns` будет заменен файлом `results.cns`.

Примечание: Дефрагментированные индексные файлы не подлежат изменениям.

3.3 Поиск

Для осуществления поиска по созданному индексу следует выполнить следующие действия:

1. Скопировать полученные индексные файлы (см. [Индексация](#)) в каталог с модулем поиска. Как правило, это каталог `/cgi-bin/` веб-сервера, на котором размещен Ваш сайт ;
2. Указать в браузере путь к интерфейсу поиска (к файлу `search.exe`);
3. В открывшейся форме ввести запрос и нажать кнопку **Искать**; при условии корректно заданного объекта поиска система отобразит список результатов поиска:

The screenshot shows a search interface with a search bar containing the text 'christmas' and a 'Search' button. Below the search bar, it indicates 'Documents found: 185, christmas: 185'. There are two links: 'Sort by date' and 'Sort by relevancy'. The search results are displayed as a list of three items, each with a bullet point and a sub-list of details:

- 1. [Most popular postcards](#) [Relevancy: 3883]
 - P-Cards.ru virtual postcards for any holidays English Русский Birthday Love New year's day Cities Travel Sport Flowers Important events Random postcard Populars Holidays Links Links exchange movie wallpapers best wallpapers \ Virtual postcards \ MOST
 - Thu Dec 14 15:46:11 2006
 - Unknown
 - <http://en.p-cards.ru/popular.php?pg=3>
- 2. [Christmas / Virtual postcards - E-Cards](#) [Relevancy: 3574]
 - P-Cards.ru virtual postcards for any holidays English Русский Birthday Love New year's day Cities Travel Sport Flowers Important events Random postcard Populars Holidays Links Links exchange movie wallpapers best wallpapers \ Virtual postcards \
 - Thu Dec 14 15:45:51 2006
 - Unknown
 - http://en.p-cards.ru/pc/New_Year/Christmas/
- 3. [Most popular postcards](#) [Relevancy: 3276]
 - P-Cards.ru virtual postcards for any holidays English Русский Birthday Love New year's day Cities Travel Sport Flowers Important events Random postcard Populars Holidays Links Links exchange movie wallpapers best wallpapers \ Virtual postcards \ MOST

Результаты поиска

Для удобства просмотра списка возможна сортировка результатов поиска по дате или по релевантности.

Настройка интерфейса списка результатов поиска осуществляется с помощью шаблонов (см. [Настройка шаблонов](#)).

Файл 'fulltxt.cns' содержит тексты всех индексируемых документов: эта информация позволяет отображать образцы текста, содержащие выделенный поисковый запрос, в результатах поиска. Например:

- 1. [Most popular postcards](#) [Relevancy: 3883]
 - 5 | 5 | > >> **Christmas** and New Year's day sent: 1 times Newborn sent: 1 times **Christmas** and New Year's day sent: 1 times Tropics sent: 1 time **Christmas** and New Year's day sent: 1 times **Christmas** and New Year's day sent: 1 time Winter sent: 1
 - Thu Dec 14 15:46:11 2006
 - Unknown
 - <http://en.p-cards.ru/popular.php?pg=3>

Однако, файл 'fulltxt.cns' может достигать больших размеров. В этом случае его можно удалить либо отменить его создание при помощи параметра **Type** (см. [search.conf](#)) в процессе индексации; в этом случае результаты поиска будут выглядеть следующим образом (без выделения и цитирования, отображаются только первые 256 символов документа):

- 1. [Most popular postcards](#) [Relevancy: 3883]
 - P-Cards.ru virtual postcards for any holidays English Русский Birthday Love New year's day Cities Travel Sport Flowers Important events Random postcard Populars Holidays Links Links exchange movie wallpapers best wallpapers \ Virtual postcards \ MOST
 - Thu Dec 14 15:46:11 2006
 - Unknown
 - <http://en.p-cards.ru/popular.php?pg=3>

Часть

IV

4 Настройка системы

[Индексатор](#)

[Модуль поиска](#)

4.1 Индексатор

[search.conf](#)

[Поддержка поиска с учётом словоформ](#)

[Стоп-слова](#)

4.1.1 search.conf

Все настройки индексатора содержатся в конфигурационном файле search.conf. Данный файл имеет следующую структуру:

```
[Job name_of_task]
[Index]
Parameter1      Value1
Parameter2      Value2
Parameter3      Value3
[Index]
Parameter1      Value1
Parameter2      Value2
Parameter3      Value3
```

Для каждого действия заданы параметры и их значения, разделяемые пробелом или табуляцией.

Примечание: В конфигурационном файле возможно использование однострочных комментариев. Каждый комментарий начинается с символа "#".

[Список параметров](#)

[Использование прокси-сервера](#)

4.1.1.1 Список параметров

Для процесса индексации возможно использование следующих параметров:

- [URL](#)

- [Extensions](#)
- [Type](#)
- [Path](#)
- [CharSet](#)
- [MaxFiles](#)
- [MinWords](#)
- [Exclude](#)
- [ExcludeVar](#)
- [AddOption](#)
- [StopWordsFile](#)
- [Language](#)
- [AFrom](#)
- [ATo](#)
- [StartWord](#)
- [Sleep](#)
- [ShowURL](#)
- [ShowEmail](#)
- [ShowFTP](#)
- [Compress](#)
- [MetaDescription](#)
- [MetaRobots](#)
- [UseRobotsTxt](#)
- [ConnectCount.](#)

URL <url>

URL url

Адрес, начинающийся с 'http://...' в HTTP-режиме индексации, либо путь к копии сайта на локальном диске в режиме локальной индексации.

Пример:

Для HTTP:

```
URL http://www.novgorod.ru/frisbee/
```

Для диска (Windows):

```
URL c:/pub/home/frisbee/
```

Для диска (Unix):

```
URL /pub/home/frisbee/
```

Extensions <ext>

Extensions ext1,ext2,ext3

Параметр задает список расширений файлов, включенных в индексацию; может использоваться только в режиме локальной индексации. Расширения файлов разделяются запятой ",".

Пример:

```
Extensions htm,html,shtml,shtm
```

Type <typ>

Type typ

Параметр задает тип поискового индекса:

- **Обычный;**
- **Сокращенный** - индексный файл меньшего размера, не поддерживающий отображение части текста, содержащей выделенные поисковые слова. (См. [Модуль поиска](#)).

По умолчанию - обычный.

Пример:

```
Type Strict
```

Path <path>

Path path

Параметр задает путь к каталогу, в котором сохраняются индексные и лог-файлы.

Пример:

```
Path c:\www\site.com
```

либо

```
Path /home/www/site.com
```

CharSet <cset>

```
CharSet cset
```

Параметр устанавливает метод определения кодировки индексируемых файлов. Возможны следующие методы:

- **ByMetaTag** - идентификация кодировки с помощью тэга META (по умолчанию).
- **ByHTTPHeader** - идентификация кодировки с помощью HTTP заголовка. В случае если данная идентификация не срабатывает, система предпринимает попытку определить набор символов с помощью тэга META. Если оба варианта не проходят, система предполагает, что данный документ имеет кодировку windows-1251.
- **win-1251** - не определяет кодировку; win-1251 по умолчанию.
- **koi8-r** - не определяет кодировку; koi8-r по умолчанию.

Пример:

```
CharSet ByHTTPHeader
```

MaxFiles <num>

```
MaxFiles num
```

Параметр задает максимальное количество индексируемых файлов; по умолчанию 10000. Будьте осторожны: многие сервера содержат огромное количество закливающихся ссылок.

Пример:

```
MaxFiles 50
```

MinWords <num>

```
MinWords num
```

Параметр задает минимальное количество слов в индексируемом документе. Документы с меньшим количеством слов не будут добавлены в поисковый индекс. Этот параметр позволяет повысить качество результатов поиска путем выбрасывания маленьких и неинформативных документов. Значение по умолчанию - 1

Пример:

```
MinWords 30
```

Statistic <stat>

```
Statistic stat
```

Параметр задает способ сохранения отчетов, которые генерируются в завершающей стадии процесса индексации и сохраняются в stats.log. Возможные опции:

- **No** - не сохранять отчет;
- **Append** - добавить к существующему файлу (по умолчанию);
- **Overwrite** - заменить существующий файл.

Пример:

```
Statistic Append
```

Exclude <excl>

```
Exclude excl1,excl2,excl3
```

Параметр задает список слов, исключаемых из индексации. Адреса, содержащие, как минимум, одно из исключаемых слов, также не включаются в индексацию. Слова разделяются запятой ",".

Пример:

```
Exclude editpost.php?,reply.php?,admin/
```

ExcludeVar <var>

```
ExcludeVar var1,var2,var3
```

Параметр задает список переменных, исключаемых из URL сайта. Переменные разделяются запятой ",".

Пример:

```
ExcludeVar PHPSESSID,order
```

AddOption <opt>

```
AddOption opt
```

Параметр задает метод индексации и используется только в HTTP-режиме. Доступны следующие варианты:

- **Page** - индексируется только текущая страница;
- **SubPages** - индексируются все страницы, URL которых содержит адрес стартовой страницы;
- **Server** - индексируется весь сервер.

Пример:

```
AddOption SubPages
```

StopWordsFile <file>

```
StopWordsFile file
```

Параметр задает имя файла, содержащего стоп-слова (см. [Стоп-слова](#)).

Пример:

```
StopWordsFile stop.txt
```

Language <lng>

Параметр задает язык. Если данный параметр указан, поле 'Accept-Language' включается в HTTP-заголовок. Эта переменная может влиять на содержимое документов на некоторых сайтах.

Пример:

```
Language ru
```

AFrom <path>

```
AFrom path
```

Параметр задает подстроку, которая в URL будет замещена строкой, указанной в параметре ATo.

Пример:

```
AFrom /home/dir/mysite/  
ATo http://search.codenet.ru/
```

ATo <url>

```
ATo url
```

Параметр задает подстроку, заменяющую [AFrom](#) в URL; используется вместе с параметром [AFrom](#).

Пример:

```
AFrom http://127.0.0.1/  
ATo http://www.codenet.ru/
```

или

```
AFrom c:/documents/www/www.codenet.ru/  
ATo http://www.codenet.ru/
```

StartWord <word>

```
StartWord word
```

Параметр определяет слово, с которого начинается процесс индексации страницы сайта. Описание страницы составляется из слов, следующих за стартовым словом. Таким образом, возможно исключение меню и т.п. из описания.

Пример:

```
StartWord about
```

Sleep <seconds>

```
Sleep seconds
```

Параметр определяет задержку между индексированием страниц сайта. Задается в секундах.

Пример:

```
Sleep 5
```

ShowURL <yesno>

```
ShowURL yesno
```

Отображать адреса страниц в процессе индексации. По умолчанию - "yes".

Пример:

```
ShowURL no
```

ShowEmail <yesno>

```
ShowEmail yesno
```

Отображать найденные адреса электронной почты (mailto:) в процессе индексации. По умолчанию - "no".

Пример:

```
ShowEmail no
```

ShowFTP <yesno>

```
ShowFTP yesno
```

Отображать найденные FTP адреса в процессе индексации. По умолчанию - "no".

Пример:

```
ShowFTP no
```

Compress <yesno>

```
Compress yesno
```

Просить сервер сжимать содержимое ответа, если сервер поддерживает такую возможность. По умолчанию - "yes". Некорректное сжатие страниц может привести к сбою в работе индексатора.

Пример:

```
Compress no
```

MetaDescription <yesno>

`MetaDescription yesno`

Параметр определяет метод описания страницы. Описание может отображаться в результатах поиска с помощью специального символа %E (см. [cnsearch.conf](#)). Возможные значения - "Yes"/"No". По умолчанию - "No". Если используется "Yes", система пытается получить описание из тэга '<META name="description...>'. Если тэг невозможно найти, либо задано значение "No", описание составляется из первых слов документа (см. StartWord).

Пример:

`MetaDescription Yes`

MetaRobots <yesno>

`MetaRobots yesno`

В случае если проставлен параметр "No", тэг '<META name="robots"...>' игнорируется; в противном случае тэг анализируется на наличие NOINDEX, NOFOLLOW, NONE. Более подробно см. в разделе Поисковые роботы. По умолчанию - "Yes".

Пример:

`MetaRobots No`

UseRobotsTxt <yesno>

`UseRobotsTxt <yesno>`

В случае если проставлен параметр "Yes", алгоритм индексации заимствуется из файла 'robots.txt', хранящегося в корневом каталоге веб-сервера. По умолчанию - "No". Более подробно см. в разделе Поисковые роботы. Имя робота - "CNSearch".

Пример:

`UseRobotsTxt yes`

ConnectCount <num>

`ConnectCount <num>`

Параметр задает количество запросов удаленного файла. По умолчанию - 5.

Пример:

`ConnectCount 10`

4.1.1.2 Использование прокси-сервера

Начиная с версии 0.91 в системе поддерживается возможность использования прокси-сервера; добавлены следующие директивы:

- [ProxyServer](#)

- [ProxyPort](#)
- [ProxyLogin](#)
- [ProxyPassword](#)

ProxyServer <serv>

```
ProxyServer server
```

Параметр задает адрес прокси-сервера. Индексатор подключается к прокси-серверу, взаимодействуя с параметром [ProxyPort](#).

Пример:

```
ProxyServer proxy.domain.ru
```

ProxyPort <port>

```
ProxyPort port
```

Параметр определяет порт прокси-сервера.

Пример:

```
ProxyPort 3128
```

ProxyLogin <login>

```
ProxyLogin login
```

Параметр задает логин для соединения с прокси-сервером; используется только в случае если прокси-сервер запрашивает авторизацию. Взаимодействует с параметром [ProxyPassword](#).

Пример:

```
ProxyLogin alex
```

ProxyPassword <password>

```
ProxyPassword password
```

Параметр задает пароль для соединения с прокси-сервером; используется только в случае если прокси-сервер запрашивает авторизацию.

Пример:

```
ProxyPassword qwerty
```

4.1.2 Поддержка поиска с учётом словоформ

Для учета в процессе поиска морфологии (грамматических форм слова) следует создать файл 'lang.cns' и сохранить его в каталоге, предназначенном для хранения индексных файлов. Данный файл не включен в дистрибутив из-за большого размера - 16 Mb.

В системе разработан специальный инструмент, позволяющий сгенерировать 'lang.cns' из словарей **ispell**, которые можно найти по адресу: <http://fmg-www.cs.ucla.edu/geoff/ispell-dictionaries.html>. Словарь **ispell** состоит из двух файлов - первый представляет собой список слов (lang.dict), второй - набор грамматических правил (lang.aff). В архивах данные файлы могут иметь другие имена; в этом случае необходимо переименовать их в 'lang.dict' и 'lang.aff' соответственно.

***Примечание:** В случае если Вы уже сгенерировали индекс с учетом различия словоформ, то при последующем процессе поиска следует также принимать во внимание словоформы и использовать тот же словарь.*

4.1.3 Стоп-слова

Начиная с версии 1.3 в CNSearch Pro реализована возможность пропуска часто используемых служебных слов (артиклей, местоимений, предлогов) в процессе индексации для увеличения скорости поиска и уменьшения объема информации, хранящегося в поисковом файле. Такого рода слова в системе условно именуются "стоп-словами".

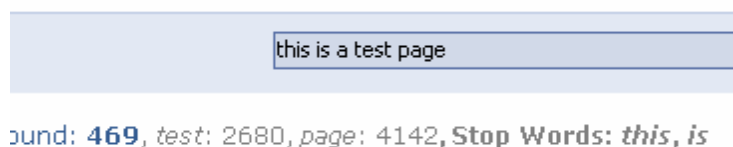
Сток-слова определяются на стадии индексации при помощи файла, в котором они размещены списком, например:

```
- file: stopwords.txt -----  
a  
an  
is  
the  
this  
-----
```

Имя файла, содержащее стоп-слова, указано в search.conf в параметре **StopWordsFile**:

```
StopWordsFile stopwords.txt
```

Посетителей Вашего сайта можно уведомить о том, какие именно слова в их запросе будут проигнорированы в процессе поиска, путем использования специального символа %P, включающего опцию отображения стоп-слов в результатах поиска:



```
und: 469, test: 2680, page: 4142, Stop Words: this, is
```

Словосочетание "Stop Words" можно заменить другим эквивалентом (например, при переводе на другой язык) путем изменения параметра StopWords в конфигурационном файле клиентской части (см. [cnsearch.conf](#)).

4.2 Модуль поиска

[cnsearch.conf](#)

4.2.1 cnsearch.conf

Конфигурационный файл модуля поиска (по умолчанию cnsearch.conf) размещается в одном каталоге с файлом 'search.exe' (search.cgi для Unix) и представляет собой текстовый файл, оптимизированный для ускоренного процесса поиска.

Файл cnsearch.conf состоит из двух частей:

- **Конфигурирования** - настройки модуля поиска;
- **Шаблонов страниц**, отображающих результаты поиска.

Структура конфигурационного файла выглядит следующим образом:

```
::CONFIG regcode = Enter Your registration code here
::CONFIG stats = password
::CONFIG content-type = text/html

::HTMLTOP
<HTML>
  <TITLE>This is the top part of the HTML document</TITLE>
</HEAD>
<BODY>

::HTMLRESULT
<P>This the description of the found page.
There will be displayed 10 such descriptions.

::HTMLNOTFOUND
<P>This text will be displayed if no search
results will be found

::HTMLBOTTOM
This is the bottom part of the HTML document
</BODY>
</HTML>
```

В конфигурационном файле возможно использование однострочных комментариев; каждый комментарий начинается с символа "#".

[Параметры конфигурирования](#)

[Настройка шаблонов](#)

[Использование различных шаблонов](#)

[Поиск по выбранным сайтам](#)

[Группировка результатов поиска по сайтам](#)

4.2.1.1 Параметры конфигурирования

Часть файла, отвечающая за конфигурирование модуля поиска, содержит следующие параметры, размещенные построчно:

Path

Параметр задает путь к поисковому индексу и может использоваться в случае если Вы не собираетесь хранить индекс в каталоге 'cgi-bin' либо планируете использовать несколько поисковых индексов.

Пример:

```
::CONFIG path=/home/www/search/en/
```

Для MS Windows:

```
::CONFIG path=d:\www\search\en\
```

Content-Type

Параметр определяет поле Content-type заголовка. По умолчанию - "text/html". Результаты поиска могут также формироваться в виде XML-файла.

Пример:

```
::CONFIG content-type = text/xml
```

SearchType

Параметр определяет логику поиска:

- **And** - будут отображены страницы, содержащие все слова поисковой фразы;
- **Or** - будут отображены страницы, содержащие хотя бы одно из слов поискового запроса;
- **Combined** - сперва отображаются результаты с учетом параметра "And", затем результаты с учетом параметра "Or" с пометкой "нечеткое совпадение".

Логическая операция "And" - наиболее быстрая; рекомендуется к использованию в случае если размер поискового индекса превышает 100Mb.

Логическая операция "Combined" рекомендуется к использованию на небольших сайтах с общим количеством страниц не более 50.

Пример:

```
::CONFIG SearchType = Combined
```

Stats

Параметр задает пароль для доступа в интерфейс статистических данных (см. [Статистика](#)).

Пример:

```
::CONFIG stats = secret
```

RegCode

Параметр задает регистрационный код (подробнее о регистрации на официальном сайте приложения).

Пример:

```
::CONFIG regcode = JF7KF-KFJEP-4KSFT-K49GN-FJ40F
```

StopWords

Параметр определяет термин, отображаемый в результатах поиска при условии использования опции %P (найденные стоп-слова).

Пример:

```
::CONFIG StopWords =, Ignored Words :
```

MaxRelevance

Параметр задает максимальную релевантность страниц отображаемых в результатах поиска. Страницы с релевантностью большей чем *MaxRelevance* игнорируются. Этот параметр позволяет повысить качество поиска путем "выбрасывания" страниц с подозрительно высокой релевантностью. Обычно это страницы не содержащие много текста или содержащие слишком часто повторяющиеся ключевые слова.

Пример:

```
::CONFIG MaxRelevance = 4000
```

NonStrictMatch

Параметр задает термин, отображаемый в результатах поиска при условии использования опции %S (соответствие поисковому запросу). Используется только совместно с логической операцией "[Combined](#)".

Пример:

```
::CONFIG NonStrictMatch = [non strict match]
```

4.2.1.2 Настройка шаблонов

Часть конфигурационного файла, содержащая шаблоны, состоит из HTML кода, генерирующего HTML-документ с результатами поиска. В данном коде следует использовать специальные символы, заменяемые на соответствующий текст по окончании генерации HTML-документа:

- **%Q** - Текст запроса;
- **%G** - Текст запроса (urlencoded);
- **%O** - Количество найденных страниц;
- **%N** - Номер страницы;
- **%U** - URL страницы;
- **%T** - Название страницы;
- **%S** - Соответствие (отображается только при полном соответствии найденной информации поисковому запросу);
- **%R** - Релевантность страницы;
- **%E** - Описание страницы;
- **%D** - Дата последнего обновления страницы;
- **%C** - Кодировка символов страницы;
- **%F** - Название поискового скрипта;
- **%I** - Номер сайта в поисковом индексе;
- **%P** - Стоп-слова, найденные в запросе;
- **%W** - Описание поискового запроса;
- **%L** - Включение опции сортировки по релевантности
- **%A** - Включение опции сортировки по дате обновления документа;
- **%B** - Навигация по найденным страницам (< << 1 2 3 4 5 6 >> >)

Пример:

```
-- cnsearch.conf -----  
# This is a cnsearch configuration file  
  
::CONFIG regcode = Enter Your registration code here  
::CONFIG stats = password  
::CONFIG content-type = text/html  
::CONFIG NonStrictMatch = [non strict match]
```

```

::CONFIG StopWords =, Ignored Words :
::CONFIG SearchType = Combined

::HTMLTOP
<HTML>
<HEAD>
<TITLE>Search results - %Q</TITLE>
</HEAD>
<BODY>
<table width=400 height=40 align=center bgcolor=#C0C0C0>
<form action="%F" method=get><tr><td align=center>
<input type=text name=q size=40 maxlength=64 value="%Q">
<input type=submit value="Search">
</td></form></tr></table>
Documents found: %O
  <B>%O</B><font color=gray>%W<B>%P</B></font><br>
<br>
<div align=right>
Sort by: <a href="%A">date</a> | <a href="%L">relevancy</a>
</div>

::HTMLRESULT
<HR>
<UL>
<LI>%N. <a href="%U" target=_new>%T</A> <small>
  <font color=red>%S</font> [Relevancy: %R]</small>
<UL>
<LI>%E
<LI>%D
<LI>%C
<LI><a href="%U" target=_new>%u</A>
</UL>
</UL>

::HTMLNOTFOUND
<P><font color=red>%Q not found</font>

::HTMLBOTTOM
%B
</BODY>
</HTML>
-- end cnsearch.conf -----

```

4.2.1.3 Использование различных шаблонов

В системе предусмотрена возможность использования различных вариантов шаблонов для отображения различных модификаций поискового интерфейса и использования различных индексных файлов в процессе поиска. Для использования нескольких шаблонов следует задать параметр 'template' в исходном коде поисковой формы. Если параметр 'template' не задан, по умолчанию используется стандартный шаблон 'cnsearch.conf'.

В качестве названия шаблона может фигурировать любое произвольное наименование. Название шаблона должно содержать только латинские буквы (верхнего либо нижнего регистра) и арабские цифры; необязательно добавлять 'conf.' к названию.

Правильный вариант:

```
<input type="hidden" name="template" value="black">
```

Неправильный вариант:

```
<input type=hidden name="template" value='../black'>
```

```
<input type=hidden name="template" value='red.htm'>
```

Ниже представлен пример использования шаблона, позволяющего пользователю:

- выбрать нужный индексный файл из нескольких для последующего поиска. Этот же результат можно достигнуть путем указания нужного шаблона в параметре Path (см. [Параметры конфигурирования](#)).

В шаблоне указан следующий путь к индексным файлам:

```
::CONFIG path=/home/www/search/en
```

- выбрать нужный конфигурационный файл для использования в процессе поиска (при помощи параметра 'template'). В данном примере пользователю предоставляется возможность выбора между шаблонами en.conf, es.conf, и ru.conf (в первичной поисковой форме будет отображен список данных шаблонов).

Пример:

```
-- en.conf -----
::CONFIG path=/home/www/search/en
::CONFIG regcode = Enter Your registration code here
::CONFIG stats = password
::CONFIG content-type = text/html
::CONFIG NonStrictMatch = [non strict match]
::CONFIG StopWords =, Ignored Words :
::CONFIG SearchType = Combined

::HTMLTOP
<HTML>
<HEAD>
<TITLE>Search results - %Q</TITLE>
</HEAD>
```

```
<BODY>
<table width=400 height=40 align=center bgcolor=#C0C0C0>
<form action="%F" method=get><tr><td align=center>
<input type=text name=q size=40 maxlength=64 value="%Q">
<input type=submit value="Search">
<select name=template>
<option value="en">English
<option value="es">Spanish
<option value="ru">Russian
</select>
</td></form></tr></table>
Documents found: %O
  <B>%O</B><font color=gray>%W<B>%P</B></font><br>
<br>
<div align=right>
Sort by: <a href="%A">date</a> | <a href="%L">relevancy</a>
</div>

::HTMLRESULT
<HR>
<UL>
<LI>%N. <a href="%U" target=_new>%T</A> <small>
  <font color=red>%S</font> [Relevancy: %R]</small>
</UL>
<LI>%E
<LI>%D
<LI>%C
<LI><a href="%U" target=_new>%u</A>
</UL>
</UL>

::HTMLNOTFOUND
<P><font color=red>%Q not found</font>

::HTMLBOTTOM
%B
</BODY>
</HTML>
-- end of en.conf -----
```

4.2.1.4 Поиск по выбранным сайтам

Начиная с версии 1.3 в системе доступна опция поиска по выбранным сайтам. Каждому сайту на стадии

индексации назначается порядковый номер, начинающийся с нуля, например:

```
[job localhost]
[Index]
URL          http://www.mysite.com/
Statistic    Append
CharSet      ByHTTPHeader
MaxFiles     10000
StopWordsFile stopwords.txt
Exclude      search/,mail/, .zip, .gif, .jpg
[Index]
URL          http://www.second.com/
Statistic    Append
CharSet      ByHTTPHeader
[Index]
URL          http://www.test.com/
Statistic    Append
CharSet      ByHTTPHeader
```

Номера сайтов назначаются следующим образом:

```
0 - http://www.mysite.com/
1 - http://www.second.com/
2 - http://www.test.com/
```

Обратите внимание, что после осуществления реиндексации у двух разных сайтов может оказаться один номер. Например при реиндексации с использованием следующего конфигурационного файла:

```
[job addon]
[Index]
URL          http://www.newsite.com/
Statistic    Append
CharSet      ByHTTPHeader
MaxFiles     10000
StopWordsFile stopwords.txt
Exclude      search/,mail/, .zip, .gif, .jpg
сайты http://www.newsite.com/ также присваивается номер "0":
0 - http://www.mysite.com/
0 - http://www.newsite.com/
1 - http://www.second.com/
2 - http://www.test.com/
```

Для реализации поиска по выбранным сайтам следует использовать параметр "d"; если данный параметр не указан (по умолчанию), поиск будет произведен по всем сайтам.

Пример:

```
-- cnsearch.conf -----
::CONFIG regcode = Enter Your registration code here
::CONFIG stats = password

::HTMLTOP
<HTML>
<HEAD>
<TITLE>Search results - %Q</TITLE>
</HEAD>
<BODY>
<table width=400 height=40 align=center bgcolor=#C0C0C0>
<form action="%F" method=get><tr><td align=center>
<input type=text name=q size=40 maxlength=64 value="%Q">
<input type=submit value="Search">
<br>
<select name=d>
<option value="0">www.mysite.com, www.newsite.com
<option value="1">www.second.com
<option value="2">www.test.com
</select>
</td></form></tr></table>
Documents found: %O
  <B>%O</B><font color=gray>%W<B>%P</B></font><br>
<br>
<div align=right>
Sort by: <a href="%A">date</a> | <a href="%L">relevancy</a>
</div>

::HTMLRESULT
<HR>
<UL>
<LI>%N. <a href="%U" target=_new>%T</A> <small>
  <font color=red>%S</font> [Relevancy: %R]</small>
<UL>
<LI>%E
<LI>%D
<LI>%C
<LI><a href="%U" target=_new>%u</A>
</UL>
</UL>

::HTMLNOTFOUND
```

```

<P><font color=red>%Q not found</font>

::HTMLBOTTOM
%B
</BODY>
</HTML>
-- end cnsearch.conf -----

```

4.2.1.5 Группировка результатов поиска по сайтам

Часто, при поиске по большому количеству сайтов результаты поиска могут засорять страницы только одного сайта. Например, при поисковой фразе "новости" будут найдены все страницы новостного сайта, заканчивающиеся на " // Местные новости", а результаты с других сайтов будут сдвинуты на сотни, а иногда на тысячи позиций.

Для того, чтобы такой ситуации не возникало, крупные поисковые системы, такие как Google, Yandex и Rambler, выводят только по одному результату с каждого сайта. С версии 1.5 такая возможность появилась в CNSearch

Для того чтобы включить группировку по сайтам, нужно добавить скрытое поле *group* в форму поискового запроса:

```

-- cnsearch.conf -----
....
<BODY>
<table width=400 height=40 align=center bgcolor=#C0C0C0>
<form action="%F" method=get><tr><td align=center>
<input type="text" name="q" size="40" maxlength="64" value="%Q">
<input type="hidden" name="group" value="1">
<input type="submit" value="Search">
</td></form></tr></table>
....
-- end cnsearch.conf -----

```

Для того чтобы дать пользователям произвести более подробный поиск по одному сайту из результатов поиска, обычно используют ссылку "еще с сайта". Реализовать ее можно с помощью [специального символа %d](#):

```

-- cnsearch.conf -----
....
::HTMLRESULT
....
<LI>%N. <a href="%U" target=_new>%T</A> <small>
    <font color=red>%S</font> [Relevancy: %R]</small>
    [ <a href="%F?d=%I&q=%G">еще с сайта</a> ]
<UL>
....
-- end cnsearch.conf -----

```

Часть



5 Дополнительные возможности

[Оптимизация поиска](#)

[Статистика](#)

[Плагины](#)

[Стандарт блокирования поисковых роботов](#)

[Использование META-тэгов "Robots"](#)

[Автоматическая генерация Google Sitemap](#)

5.1 Оптимизация поиска

В системе предусмотрена возможность оптимизации двух параметров:

- **Скорость поиска** - рекомендуется оптимизировать в случае если процесс поиска осуществляется часто либо результаты поиска отображаются медленно;
- **Размер индексного файла** - рекомендуется оптимизировать в случае если провайдер хоста устанавливает лимит места на диске.

[Оптимизация скорости поиска](#)

[Оптимизация размера индексного файла](#)

5.1.1 Оптимизация скорости поиска

Для оптимизации скорости поиска рекомендуется применить следующие опции:

- Использовать дефрагментированный индекс (см. [Дефрагментация](#));
- Использовать опцию "стоп-слова" (см. [Стоп-слова](#));
- Использовать логическую операцию "And" (см. [Параметры конфигурирования](#)) - при этом количество обращений к диску снижается (только при условии использования дефрагментированного индекса);
- Отключить поддержку морфологии - процесс поиска осуществляется быстрее без использования словаря lang.cns.

5.1.2 Оптимизация размера индексного файла

Для оптимизации скорости поиска рекомендуется применить следующие опции:

- Использовать сокращенный вариант поискового индекса - отменить создание файла "fulltxt.cns", что приведет к уменьшению размера индекса в 1.5 - 2 раза. Данную опцию можно осуществить при помощи параметра `Ture` в конфигурационном файле индексатора (см. [search.conf](#));
- Отключить поддержку морфологии;
- Использовать опцию "стоп-слова".

5.2 Статистика

CNSearch осуществляет полнотекстовый поиск и предоставляет статистические отчеты для оценки содержимого сайта и релевантности его составляющих.

Статистические данные хранятся в файле 'stats.cns', который следует содержать в том же каталоге, что и поисковый индекс. В случае если cgi-скрипт, осуществляющий поиск, не сможет иметь доступа к файлу либо не будет иметь разрешение на его создание (типичная практика), то статистические данные сохраняться не будут.

Вы можете создать файл 'stats.cns' вручную и задать права доступа для него.

Доступ к статистическим данным защищен паролем, который устанавливается в конфигурационном файле клиентской части при помощи параметра **Stats** (см. [Параметры конфигурирования](#)), например:

```
-- cnsearch.conf -----  
::CONFIG stats = thisispass  
::HTMLTOP  
<HTML>  
<HEAD>  
<TITLE>Search results - %Q</TITLE>  
</HEAD>  
...  
-- end of cnsearch.conf -----
```

Для просмотра статистических данных следует указать параметр 'stats' в URL сайта, например:

```
http://www.site.com/cgi-bin/search.cgi?stats=1&password=thisispass
```

или:

```
http://www.site.com/cgi-bin/search.exe?stats=1&password=thisispass
```

На данный момент в системе доступны два отчета:

1. Поисковые запросы;

Этот отчет отображает поисковые фразы, наиболее часто используемые посетителями Вашего сайта, и

количество найденных результатов. С помощью этого отчета Вы сможете проанализировать наиболее популярные объекты поиска, процент их находжений. Возможен просмотр статистики за любой период времени:

Date: Month: Year: Date: Month: Year:

<< < | 1 | [26](#) | [51](#) | [76](#) | > >>

Search phrase	Pages found	count
	0	21
java	1307	10
flash	27	8
hook	17	7
php	1721	6
ssi	40	6
mail	1762	5
картинки	107	5
информатизация	1	5
ulead photo express	22	5
perl	1497	5
all	354	5

2. Распределение поисковых запросов по времени.

В данном отчете отображено распределение количества поисковых запросов по времени; возможна настройка просмотра данных за любой период:

Date: Month: Year: Date: Month: Year:

2002-8-14	93	
2002-8-13	172	
2002-8-12	71	
2002-8-11	68	
2002-8-10	162	
2002-8-9	207	
2002-8-8	210	
2002-8-7	191	
2002-8-6	51	

В случае если Вам понадобятся другие виды статистических отчетов, пожалуйста, обращайтесь к нам; возможно, они будут включены в следующие версии программы.

5.3 Плагины

Плагины представляют собой специальные модули, позволяющие расширить функционал приложения. CNSearch использует плагины к индексным файлам различных типов.

Плагины следует хранить в том же каталоге, что и индексатор. В версиях UNIX и Linux эти файлы имеют расширение .so, в Windows - .dll. Для отключения плагина потребуется всего лишь переместить его в другой каталог.

В текущую версию приложения включены плагины, позволяющие индексировать файлы следующих типов:

Имя файла для версии UNIX/ Linux	Имя файла для версии Windows	Тип обрабатываемого документа
libtxt.so	libtxt.dll	*.TXT - текстовые файлы
librtf.so	librtf.dll	*.RTF - текстовые файлы формата RFT
libdoc.so	libdoc.dll	*.DOC - файлы Microsoft Word
libxls.so	libxls.dll	*.XLS - файлы Microsoft Excel
libmp3.so	libmp3.dll	*.MP3 - MPEG Layer 3 аудио-файлы

Плагины версии 0.92 не определяют кодировку, поскольку для большинства файлов это не нужно.

Поле 'encoding' в документах, обрабатываемых плагинами, заменяется текстом, заданным в плагине; это позволяет создавать шаблоны с отображаемым типом найденного документа.

При запуске индексатор загружает все активные плагины, например:

```
F:\1\bin\indexer>searchctl.exe localhost
CNSearch ver.0.92 [build 2073]
Compiled 07.04.2002 under MS Windows 2000 [Version 5.00.2195]
Rebuilding URL list...Ok.
Loading library: RTF (Rich text format)
Loading library: TXT (Plain text)
Loading library: DOC (Microsoft Word document format)
http://www.test.ru/
```

Главным преимуществом плагинов является возможность создания новых, позволяющих индексировать файлы особых форматов, например, изображений.

[Создание плагина](#)

5.3.1 Создание плагина

Для создания плагина следует воспользоваться архивом 'plugin.zip', расположенным в папке '/manual' инсталляционного пакета. Данный файл содержит исходный код плагина, обрабатывающего текстовые файлы.

Для корректной работы в системе плагин должен иметь надлежащее расширение и обладать следующим набором функций:

Название функции	Описание функции
char *get_info(void)	Возвращает строку - информацию о плагине (название)
char *get_mime(void)	Возвращает строку - список MIME TYPEs, обрабатываемых данным плагином и разделенных вертикальной линией " "
char* get_shortdesc(void)	Возвращает строку - краткое название типа файла
char* get_range(void)	Возвращает строку - поле "Range" HTTP заголовка (см. RFC2068); если поле не используется возвращает значение NULL.
char* get_title(void)	Возвращает строку - название документа. При значении NULL отображается URL документа.
TPluginWord* get_word(unsigned char *d, unsigned long filesize)	<p>Основная функция - возвращает указатель на структуру 'TpluginWord', содержащую слово, добавляемое к поисковому индексу. Данная функция должна возвращать слова, содержащиеся в документе последовательно.</p> <ul style="list-style-type: none"> d - указатель на индексируемый документ, оканчивающийся кодом \0x0 filesize - размер индексируемого документа; используется в случае если документ содержит код \0x0 (например, Microsoft Word Document)

Структура TpluginWord выглядит следующим образом:

```
typedef struct {
    char word[32];
    int rel;
    bool end;
} TPluginWord;
```

где:

- word** - слово с присоединенными значениями \0x00 справа; таким образом, максимальная длина слова равна тридцати двум символам.

- **rel** - релевантность слова; может варьироваться от 1 до 256. Рекомендуемые значения - от 1 до 4. В данном примере релевантность каждого слова равна 1, за исключением слова, состоящего из заглавных букв - его релевантность равна 2.
- **end** - имеет значение 'true', если в документе больше нет слов; в этом случае 'word' и 'rel' игнорируются.

Методы, используемые программой для генерации функций плагина:

- Функции *get_info()*, *get_mime()*, и *get_shortcode()* вызываются один раз, при загрузке плагина;
- Функция *get_title()* вызывается единожды для каждого документа, после чего вызывается функция *get_word()* для соответствующих документов до тех пор, пока поле 'end' структуры TwordPlugin не приобретает значение 'true'.

5.4 Стандарт блокирования поисковых роботов

[Введение](#)

[Функция](#)

[Структура](#)

[Примеры](#)

5.4.1 Введение

Поисковые роботы (search bots) представляют собой программы, индексирующие веб-документы в сети Internet.

В 1993-94 годах было обнаружено, что поисковые роботы зачастую индексируют сайты против воли их владельцев. Иногда, вследствие различных причин, роботы могут проиндексировать одни и те же файлы несколько раз. В некоторых случаях поисковые роботы индексируют ненужные документы - виртуальные каталоги, временные данные либо CGI-скрипты. Для решения подобных проблем был разработан Стандарт Блокирования.

5.4.2 Функция

Для решения проблемы необходимо создать файл, содержащий информацию об управлении поведением робота с целью заблокировать запрос робота к веб-серверу либо его составляющим. Данный файл должен находиться в корневом каталоге сайта '/robots.txt'.

Суть решения заключается в предоставлении роботу возможности нахождения алгоритмов, описывающих его действия по запросу только одного файла. Файл '/robots.txt' можно создать на любом действующем веб-сервере.

Выбор подобного имени файла продиктован следующими обстоятельствами:

- Имя файла должно быть одинаковым для любой операционной системы;

- Расширение файла не должно повлечь за собой необходимость реконфигурации сервера;
- Имя файла должно быть описательным и запоминающимся;
- Возможность совпадения с существующими файлами должна быть минимальной.

5.4.3 Структура

Структура и семантика файла '/robots.txt' заключается в следующем:

Файл должен содержать одну или несколько записей, разделенных одной или более строками (оканчивающимися кодом CR, CR/NL, or NL). Каждая запись должна содержать строки "<field>: <optional_space><value><optional_space>".

Поле <field> не зависит от регистра.

Комментарии задаются стандартным способом, используемым в UNIX: символ '#' обозначает начало комментария, конец линии обозначает завершение комментария.

Запись следует начинать со строки 'User-Agent' (одной или нескольких), за которой следует одна или более строк Disallow. Нераспознаваемые строки игнорируются.

User-Agent:

- Значение этого поля должно представлять собой имя поискового робота. В этой записи задаются права доступа данного робота;
- Несмотря на то, что стандарт позволяет обозначать имена нескольких роботов, приложение CNSearch распознает только одного, поскольку здесь не реализован метод разделения роботов по именам;
- Регистр не имеет значения;
- В случае если значение этого поля равно '*', то права доступа, заданные в записи, действительны для любого поискового робота, запрашивающего файл '/robots.txt'.

Disallow:

- Значение данного поля должно представлять собой частичный неиндексируемый URL. Путь к файлу может быть полным либо частичным. Например, 'Disallow: /help' блокирует доступ как к файлу '/help.html', так и к файлу '/help/index.html', тогда как 'Disallow: /help/' блокирует доступ лишь к файлу '/help/index.html'.
- Любая запись должна содержать как минимум одну строку 'User-Agent' и одну - 'Disallow'.

Если файл '/robots.txt' пуст, не соответствует вышеозначенной структуре и семантике либо отсутствует, поисковые роботы действуют согласно своим настройкам.

5.4.4 Примеры

Пример 1:

```
# robots.txt for http://www.site.com
User-Agent: *
# this is an infinite virtual URL space
Disallow: /cyberworld/map/
Disallow: /tmp/ # these will soon disappear
```

В данном примере содержание '/cyberworld/map/' и '/tmp/' защищено.

Пример 2:

```
# robots.txt for http://www.site.com
User-Agent: *
# this is an infinite virtual URL space
Disallow: /cyberworld/map/
# Cybermapper knows where to go
User-Agent: cybermapper
Disallow:
```

В данном примере поисковому роботу 'cybermapper' предоставлен полный доступ, тогда как остальные роботы не имеют доступа к содержимому '/cyberworld/map/'.

Пример 3:

```
# robots.txt for http://www.site.com
User-Agent: *
Disallow: /
```

В данном примере доступ к серверу запрещен любому поисковому роботу.

5.5 Использование META-тэгов "Robots"

Помимо вышеописанного стандарта блокирования поисковых роботов в системе также представлена возможность управления поведением роботов при помощи HTML-тэга 'META'.

В отличие от файла 'robots.txt', описывающего процесс индексации сайта, как единого целого, тэг 'META' позволяет управлять процессом индексации конкретной веб-страницы. Кроме того, возможна отмена индексации не только документа в целом, но и ссылок, содержащихся в нем.

Параметры индексации следует указывать в атрибуте 'content' исходного кода каждой страницы веб-сайта. Возможно использование следующих параметров:

- **NOINDEX** - отменить индексацию документа;
- **NOFOLLOW** - отменить индексацию ссылок, найденных в документе;

- **INDEX** - осуществить индексацию документа;
- **FOLLOW** - осуществить индексацию ссылок, найденных в документе;
- **ALL** - аналогично INDEX, FOLLOW
- **NONE** - аналогично NOINDEX, NOFOLLOW

Значение по умолчанию: `<meta name="Robots" content="ALL">`.

Примечание: не следует перечислять значения через запятую.

Пример некорректного варианта:

```
<META name="ROBOTS" content="noindex, nofollow">
```

Правильный вариант:

```
<META name="ROBOTS" content="none">
```

В данном примере индексатор позволяет анализировать документ без последующей индексации ссылок, найденных в нем:

```
<META name="ROBOTS" content="nofollow">
```

Имя тэга, а также названия и значения полей не зависят от регистра. В действительности, индексатор проверяет наличие лишь трех значений: NOINDEX, NOFOLLOW и NONE, поскольку FOLLOW и INDEX являются значениями по умолчанию.

5.6 Автоматическая генерация Google Sitemap

Примечание: Генерация Google Sitemap не является основной функцией CNSearch: в первую очередь CNSearch - это система для организации полнотекстового поиска по сайту.

Для генерации Google Sitemap необходимо произвести индексацию ([см. модуль индексации](#)) и установить модуль поиска ([см. модуль поиска](#)) на сайт. После этого генерация Google Sitemap может быть произведена посредством обращения к адресу:

```
http://www.site.com/cgi-bin/search.cgi?sitemap=1&password=secretword
```

где,

- `http://www.yourserver.com/cgi-bin/search.cgi` - полный путь к модулю поиска;
- `secretword` - пароль для просмотра статистики ([см. статистика](#))

Если поисковый индекс содержит несколько сайтов, то с помощью дополнительного параметра `d` можно задать номер сайта для которого будет сгенерирован Google Sitemap.

Способ присвоения номеров описан в разделе "Поиск по выбранным сайтам"

Например:

```
http://www.site.com/cgi-bin/search.cgi?sitemap=1&password=secretword&d=5
```

Генерацию Google Sitemap наиболее логично производить сразу после обновления поискового индекса. В Unix/Linux системах для этого удобно использовать программы fetch (установлена в FreeBSD по умолчанию) или wget:

```
fetch -q -o /path/to/www.site.com/sitemap.xml \  
http://www.site.com/cgi-bin/search.cgi?sitemap=1&password=secretword
```

или

```
wget -q -O /path/to/www.site.com/sitemap.xml \  
http://www.site.com/cgi-bin/search.cgi?sitemap=1&password=secretword
```

где

- **/path/to/www.site.com** - путь к корневому каталогу сайта
- **http://www.yourserver.com/cgi-bin/search.cgi** - полный путь к модулю поиска
- **secretword** - пароль для просмотра статистики (см. [статистика](#))

Ссылки по теме

1. [Информация о Google Sitemap на сайте Google](#)
2. [BSD fetch utility на сайте www.freebsd.org](#)
3. [Официальный сайт GNU Wget](#)