Sphinx 安装手册

LAMP 兄弟连 http://bbs.lampbrother.net

蔡达老师

Sphinx + MySQL5.1.x + SphinxSE + MMSeg(中文分词)

黄色 字体	要执行的命令，可直接粘贴复制。
青绿色字体	出现的错误。
绿色 字体	配置文件中标题性质选项。
粉红色字体	需要注意的不同之处。
红色背景字	安装步骤。

如粘贴复制的命令报错,请按命令手动输入一下,有时候可能多粘贴一个空格就会报错。

假设现在你已经安装了 LAMP 环境（CentOS5.5 版本、mysql-5.1.x 版本、Apache-2.2.9 版本、php-5.2.6 版本），在安装的过程中我们会碰到依赖包的问题，如果没有网络，把 yum 源修改成本地的，如果有网络直接在网上搜索即可。

源码包放在 /lamp 下，安装在 /usr/local/

Sphinx 安装与运行测试(一)

(1)、下载 Sphinx

从 sphinx 官网上找到 sphinx 的安装源码，我们下的是 0.9.9-release 版。

下载地址：http://www.sphinxsearch.com/downloads/sphinx-0.9.9.tar.gz

如果你的 linux 能联网的话,用 wget 命令直接下载.

```
Wget http://www.sphinxsearch.com/downloads/sphinx-0.9.9.tar.gz
```

```
解压：tar-zxvfsphinx-0.9.9.tar.gz
```

(2)、编译安装

进入到 sphinx 的源码文件夹里，运行下列命令就可以安装 sphinx 了：

```
cd /lamp/sphinx-0.9.9
```

```
./configure --prefix=/usr/local/sphinx --with-mysql=/usr/local/mysql
make && make install
```

Sphinx 中重要的三个命令，(Sphinx 安装的 bin 目录下)

Indexer 创建索引命令。Searchd 启动进程命令。Search 命令行搜索命令。

(3)、导入数据

我们需要一些数据，这里我们用安装 mysql 自带的 test 库进行测试

运行/usr/local/sphinx/etc/目录下的 example.sql 脚本，把数据导到数据库中：

```
/usr/local/mysql/bin/mysql -uroot -p***  < /usr/local/sphinx/etc/example.sql
```

这个是 sphinx 自带的测试数据。

(4)、进入 mysql 中查看并添加的数据

```
/usr/local/mysql/bin/mysql -uroot -p（密码）
```

进入 test 库，其中 documents 表是自动导进来的;

一共是 4 条记录。我们给它添加几条中文记录。

```
mysql> insert into documents values (null,1,10,now(),'兄弟连','感谢 LAMP 兄弟连的广大会员，感谢一直以来对兄弟连的关注与支持');
```

(5)、配置 sphinx.conf 配置文件。

进入到 sphinx 的 etc 目录找到配置文件

```
cd /usr/local/sphinx/etc
```

我们需要备份一下配置文件，防止改错不好处理。

```
cp sphinx.conf.dist sphinx.conf
```

进入配置文件。

```
vim sphinx.conf
```

配置文件的格式：

```
名称 {
配置选项;
。。。。。。。。。。。。。。。
}
```

```
# sphinx.conf 的基本配置 ：
数据源 src 是名字可以自己指定（意思就是说数据从哪里来得）
source src1                13 行
{    type = mysql                        # 数据库类型
sql_host = localhost                     # MySQL 主机 IP
sql_user = sphinxuser                    # MySQL 用户名
sql_pass = sphinxpass                    # MySQL 密码
sql_db    =  sphinx                      # MySQL 数据库
sql_port = 3306                          # MySQL 端口
```

```
sql_sock = /tmp/mysql.sock              # 如果是 linux 下需要开启，指定 sock 文件。      35 行
sql_query_pre = SET NAMES UTF8          # MySQL 检索编码 73 行
sql_query_pre = SET SESSION query_cache_type=OFF #关闭缓存 74 行
sql_query = \                           # 获取数据的 SQL 语句    79 行 （默认就可以）
SELECT id, group_id, UNIX_TIMESTAMP(date_added) AS date_added, title, content
\FROM documents
sql_attr_uint = group_id                # 无符号整型 107 行 根据 79 行指定的字段填写
sql_attr_timestamp = date_added         # 时间类型  131 行 根据 79 行指定的字段填写
# 用于命令界面端(CLI)调用的测试(一般来说不需要)   187 行
sql_query_info      = SELECT * FROM documents WHERE id=$id
}

xmlpipe settings                                211 行 这个是 XML 类型的，直接过去。

source srclthrottled:srcl    继承索引源。  256 行   我们在增量索引时用，使用篇会讲到。

主索引
index text1          271 行
{  source = src1                # 索引源声明  (根据我们指定的主索引源的名字)
charset_type = utf-8            # 数据编码(设置成 utf8)                  340 行
charset_table =                 # 上面指定了 utf-8，这里需要开启。       353 行
}

index testlstemmed:test1    增量索引        后面会讲到 (先关闭掉#号注释)481 行
index distl         分布式也注释掉          后面会讲到（暂时可以关闭,#号注释）492 行

索引器设置
indexer              529 行
{ mem_limit = 256M # 内存大小限制 默认是 32M, 推荐为 256M }   其他用默认即可

sphinx 服务进程 searchd 的相关配置       565 行
searchd { }  全部用默认的就可以了。
如果是本地测试，使用默认就可以，如果是多个服务器测试需要指定监听的 ip 即可。
```

Sphinx 配置文件结束

（6）、创建索引

Sphinx 的配置文件配置完成，数据也导进去了，接下来就用下面命令来创建索引：

创建索引命令：indexer

-c　　指定配置文件

--all　对所有索引重新编制索引。

--rotate 用于轮换索引，主要是再不停止服务的时候，增加索引。

--buildstops --buildfreqs　要一起使用，

例：indexer myindex --buildstops text.txt 1000 --buildfreqs

会在当前目录下生成，一个名字叫 text.txt,最多包含 1000 个词的词表。

--merge 合并索引 （后面会详细讲）

/usr/local/sphinx/bin/indexer -c /usr/local/sphinx/etc/sphinx.conf --all

创建索引是报了一个这样的错误：/usr/local/sphinx/bin/indexer: error while loading shared libraries: libmysqlclient.so.16: cannot open shared object file: No such file or directory，

可以用下面方法解决：

locate libmysqlclient　运行该命令找到关于 libmysqlclient.so.16 的文件

然后把该文件的一个连接复制到在环境变量的文件夹/usr/lib/下，

cp /usr/local/mysql/lib/mysql/libmysqlclient.so.16 /usr/lib/libmysqlclient.so.16

再次运行创建索引命令就能完成索引的创建了。

```
collected 8 docs, 0.0 MB
sorted 0.0 Mhits, 100.0% done
total 8 docs, 827 bytes
total 0.005 sec, 152077 bytes/sec, 1471.12 docs/sec
total 2 reads, 0.000 sec, 0.2 kb/call avg, 0.0 msec/call avg
total 7 writes, 0.000 sec, 0.2 kb/call avg, 0.0 msec/call avg
```

出现这样的提示就表示创建成功。

(7)、测试

检查数据命令：search （详细方法会调用 API 使用）

/usr/local/sphinx/bin/search -c /usr/local/sphinx/etc/sphinx.conf test（test 关键字）

```
[root@localhost etc]# /usr/local/sphinx/bin/search -c /usr/local/sphinx/etc/sphinx.conf test
Sphinx 0.9.9-release (r2117)
Copyright (c) 2001-2009, Andrew Aksyonoff

using config file '/usr/local/sphinx/etc/sphinx.conf'...
index 'test1': query 'test ': returned 3 matches of 3 total in 0.000 sec

displaying matches:
1. document=1, weight=2, group_id=1, date_added=Sat Oct  8 14:15:41 2011
        id=1
        group_id=1
        group_id2=5
        date_added=2011-10-08 14:15:41
        title=test one
        content=this is my test document number one. also checking search within phrases.
2. document=2, weight=2, group_id=1, date_added=Sat Oct  8 14:15:41 2011
        id=2
        group_id=1
        group_id2=6
        date_added=2011-10-08 14:15:41
        title=test two
        content=this is my test document number two
3. document=4, weight=1, group_id=2, date_added=Sat Oct  8 14:15:41 2011
        id=4
        group_id=2
        group_id2=8
        date_added=2011-10-08 14:15:41
        title=doc number four
        content=this is to test groups

words:
1. 'test': 3 documents, 5 hits
```

可以看到将数据中含有 test 关键字的数据查询出来，包括文档 id，权重，属性值等。

我们搜索下中文看下。

/usr/local/sphinx/bin/search -c /usr/local/sphinx/etc/sphinx.conf '兄弟连'

```
[root@localhost etc]# /usr/local/sphinx/bin/search -c /usr/local/sphinx/
Sphinx 0.9.9-release (r2117)
Copyright (c) 2001-2009, Andrew Aksyonoff

using config file '/usr/local/sphinx/etc/sphinx.conf'...
index 'test1': query '兄弟连 ': returned 0 matches of 0 total in 0.000 s

words:
```

找不到数据，为什么？

[中文分词 LibMMSeg(coreseek 中自带的) 安装(二)]

Coreseek 介绍：

Sphinx 默认不支持中文索引及检索，基于 Sphinx 开发了 Coreseek 全文检索服务器，Coreseek 应该是现在用的最多的 Sphinx 中文全文检索，它提供了为 Sphinx 设计的中文分词包 LibMMSeg 包含 mmseg 中文分词。

(1)、下载中文分词包

http://www.coreseek.cn 到官网去下载 Coreseek 相应的版本。

(2)、解压安装

```
cd /lamp
tar -zxvf coreseek-3.2.14.tar.gz
```

进入到 mmseg 所在文件夹，先安装中文分词 mmseg。

```
cd /lamp/coreseek-3.2.14/mmseg-3.2.14/
./configure --prefix=/usr/local/mmseg
```

编译过程中报了一个 `config.status: error: cannot find input file: src/Makefile.in`

这个的错误，然后运行下列指令再次编译就能通过了：

```
automake
```

然后再进行编译和安装：

```
make && make install
```

然后运行 mmseg，就能输入安装成功的信息了：

```
/usr/local/mmseg/bin/mmseg
```

出现下列信息，就证明 mmseg 中文分词已经安装好了。

```
[root@localhost mmseg-3.2.14]# mmseg
Coreseek COS(tm) MM Segment 1.0
Copyright By Coreseek.com All Right Reserved.
Usage: mmseg <option> <file>
-u <unidict>            Unigram Dictionary
-r          Combine with -u, used a plain text build Unigram Dictionary
-b <Synonyms>           Synonyms Dictionary
-t <thesaurus>          Thesaurus Dictionary
-h           print this help and exit
```

Coreseek 安装与运行测试(三)

接下来，我们要把 Sphinx 和 mmseg 结合起来。

(1)、检测安装

进入 coreseek 目录，进行安装。

```
cd /lamp/coreseek-3.2.14/csft-3.2.14/
```

```
./configure --prefix=/usr/local/coreseek --with-mysql=/usr/local/mysql --with-mmseg=/usr/local/mmseg --with-mmseg-includes=/usr/local/mmseg/include/mmseg/ --with-mmseg-libs=/usr/local/mmseg/lib/
```

```
make && make install
```

(2)、配置带有中文分词的 sphinx 配置文件

配置文件和上面的步骤一样，只不过是在 coreseek 中，有几个地方需要注意。

注意：coreseek 中得配置文件也是 sphinx.conf,而不是 coreseek.conf。

```
cd  /usr/local/coreseek/etc
```

```
cp  sphinx.conf.dist sphinx.conf
```

```
vim sphinx.conf
```

其他地方都一样，对照下面不一样的地方修改。

```
index test1
```

（只有索引这个里面的值要修改）

```
{
```

需要注释的地方：

```
#stopwords                              = G:\data\stopwords.txt
```

315 行

```
#wordforms                              = G:\data\wordforms.txt
```

321 行

```
#exceptions                             = /data/exceptions.txt
```

330 行

```
#charset_type                           = sbcs
```

340 行

添加下面这两行，意思是把中文分词加入到配置文件中。

```
charset_type            = zh_cn.utf-8
```

```
charset_dictpath     = /usr/local/mmseg/etc/
```

你安装 mmseg 的目录

```
}
```

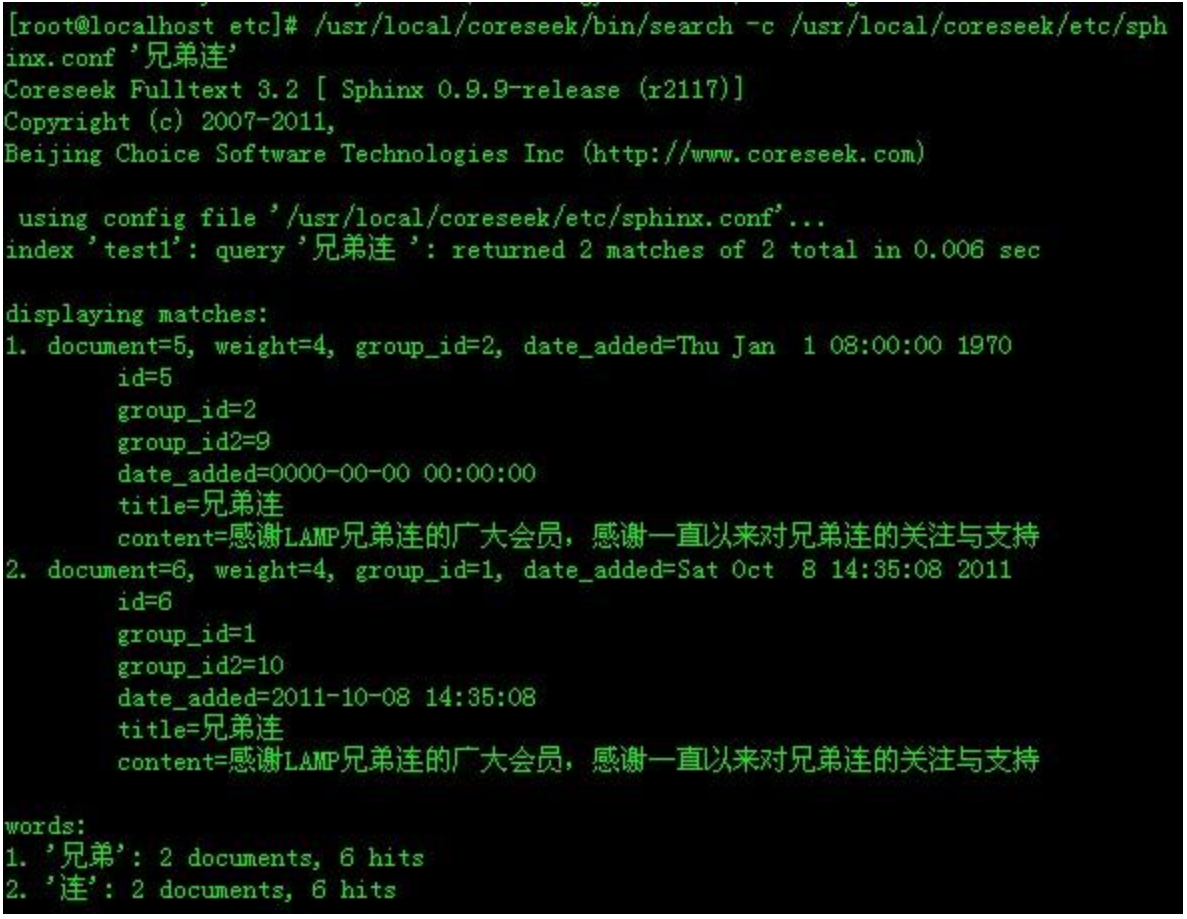
(3)、生成索引并测试

创建索引

```
/usr/local/coreseek/bin/indexer -c /usr/local/coreseek/etc/sphinx.conf --all
```

再次测试搜索中文

```
/usr/local/coreseek/bin/search [-a] -c /usr/local/coreseek/etc/sphinx.conf '兄弟连'
```



我们在 linux 下 Sphinx，中文分词已经安装完成了，并测试成功。

至于怎么在 PHP 程序中使用，Sphinx 使用篇详细道来。