Biotechvana Search Engine 1.0

Lloréns, C.[1,2] Futami, R.[1] Vicente-Ripolles, M.[1,3] and Moya, A.[2,4]

1 - Biotechvana, Valencia, Spain
2 - Instituto Cavanilles de Biodiversitat i Biología Evolutiva Universitat de València, Spain
3 - Departament de Sistemas Informátics i Computació Universitat Politécnica de València
4 - CIBER de Epidemiología y Salud Pública (CIBERESP), Spain

Corresponding author: carlos.llorens@biotechvana.com

Availability: Available online February 1, 2008. BSE is distributed under the terms of the Biotechvana Open Source License (URL 1).

Biotechvana Search Engine is a cross-platform customizable web search engine script written in PHP that indexes in a database all the words contained in your web site or intranet to provide faster searches to your visitors. Results are sorted by relevance and a list of similar words is given when no matches are found.

Keywords: PHP | search engine | web crawling | web indexing

OVERVIEW

Biotechvana Search Engine (BSE) allows web developers to include a highly customizable search engine in any PHP-developed web site just by adding a handful of scripts. This tool acts like a web crawler, parsing and indexing all web pages and keywords of a web site. The workflow is divided in two main processes:

a) Indexing: Starting at the index page of the web site, BSE explores the source code of the current document and captures all hyperlinks present in the document, recording them in the URL database to be processed later. Once the document is analyzed, BSE looks for the next non-analyzed available location of the URL database, until all locations are parsed. Only internal hyperlinks are recorded, avoiding locations external to the target web site. Once all URLs are processed, BSE begins the keyword recording process and all significant words of the document are recorded in a keyword database. Finally, each document recorded in the URL database is parsed looking for the keywords contained in the keyword database. If a keyword is found in the document, the document key and the keyword key are recorded in a URL-keyword pair match database recording also the number of appearances of the keyword in that URL. The entire process is shown in figure 1.

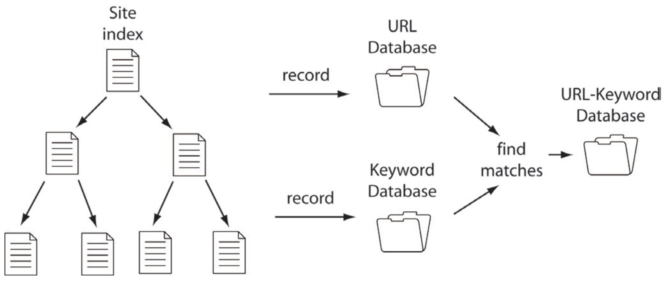

Figure 1. BSE indexing a web site flow diagram.

b) Searching: BSE looks for the keyword in the keyword database. If it does not exist, a list of similar words is presented to the user. The engine then looks for the keyword key in the URL-keyword pair match database to retrieve all URLs containing this keyword, and organizes results by the number of appearances of the keyword in each. This process is shown in figure 2.

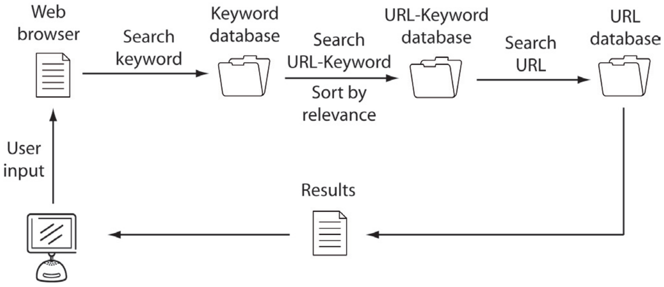

Figure 2. BSE searching keywords flow diagram.

INSTALLATION

First, confirm that a web server, a PHP application server, and a DBMS are properly installed on your system. You can download and install a web server like Apache at URL 2 for Windows and Linux platforms; or an IIS (Internet Information Services) web server for Windows platforms, which comes included in Windows server versions. Next, install the PHP server, which can be downloaded at URL 3. Installation instructions are provided on its corresponding web sites. Finally, install the DBMS. There are many DBMS available, such as MySQL at URL 4, Firebird (URL 5), and PostgreSQL (URL 6). A list of DBMS supported by this application is available at the AdoDB web site at URL 7. This script has been successfully tested in a MySQL DBMS, we thus recommend this choice. Once a web server engine, the PHP application server, and a DBMS are properly installed and working, proceed with the installation of BSE as following. Unpack the zip package into your web server's public folder and two folders will be created: 'indexing', which contains the scripts required performing the web site indexing; and 'bse', which contains the scripts required for database querying

Database setup: create a user named 'bse' and create a database named 'bse_bse'. Execute the SQL backup file called 'bse.sql', included with this application. For MySQL users, enter in your command line the following instruction: `'mysql -u root -p bse_bse < bse.sql'`.

Indexing: Copy the folder ‘indexing’ in a folder in which the web server can execute it, but not publicly accessible to all users. Open your web browser and type (http://localhost/indexing/index.php) in the URL. In the list of options, select ‘Start indexing’. The process will take some minutes, depending on the contents of your web site. Once indexing is complete, a message will appear and your database will be ready for working.

Testing: open your web browser and type the following URL (http://localhost/indexing/index.php). In the list of options, select ‘Search keyword’. You will see a web form for performing tests in your database.

Web site integration: First, copy the folder ‘bse’ to a folder of your web site. Then embed in your Web page a form HTML tag:

```
<form name=‘formSearch’ action=‘bse/search/process.
php’ method=‘post’>
<input type=‘text’ name=‘key’ />
<input type=‘submit’ name=‘action’ value=‘Search’
/>
<input type=‘hidden’ name=‘start’ value=‘0’ />
<input type=‘hidden’ name=‘end’ value=‘10’ />
</form>
```

Results pagination: To customize pagination, two parameters encoded with hidden type fields must be modified. ‘Start’ is the starting record and ‘end’ is the number of records to show in each page. Finally, open your web page and test the search engine.

REQUIREMENTS

To install BSE, the following are required: a web server, a PHP application server, and a database management system (DBMS). A web server is a computer program responsible for accepting HTTP requests from the user’s web browsers and delivering HTML web pages, images, and other files. An application server is software that helps a web server to process web pages containing server-side scripts that cannot be processed by a regular web server. When a dynamic page is requested by a visitor’s browser, the web server calls the application server for processing of the scripts, prior to sending the page to the browser. A DBMS is computer software designed for the purpose of managing databases. It controls the organization, storage, management, and retrieval of data in a database. Some knowledge of HTML language is also required, principally in web forms management, as well as intermediate knowledge of PHP programming concepts, and some basic knowledge of DBMS management (creating databases and tables, import-export SQL files, etc.).

ACKNOWLEDGMENTS

We thank Rachel Epstein for language revision and the Servei Central de Suport a la Investigació Experimental (SCSIE) (URL 8) at UVEG for technical support. Biotechvana Bioinformatics has been awarded the NOVA 2006 by IMPIVA and Conselleria d`Empresa, Universitat I Ciència of Valencia. The research has been partly supported by grants IMCBTA/2005/45, IMIDTD/2006/158 and IMIDTD/2007/33 from IMPIVA, and by grant BFU2005-00503 from MEC to AM.

URLS

1. Open Source License: http://biotechvana.com/loader.php?section=contents&page=terms_ocl
2. PHP language: http://php.net
3. Apache: http://www.apache.org
4. MySQL: http://www.mysql.com.
5. Firebird: http://www.firebirdsql.org
6. PostgreSQL: http://www.postrgresql.com.
7. AdoDB: http://adodb.sourceforge.net.
8. SCSIE, Universitat de València: http://scsie.uv.es

SPONSORS

VNIVERSITAT DE VALÈNCIA

IMPIVA

GENERALITAT VALENCIANA
CONSELLERIA D'EMPRESA, UNIVERSITAT I CIÈNCIA

CEEI VALENCIA

BIOVAL
ASOCIACIÓN DE EMPRESAS
BIOTECNOLÓGICAS
DE LA COMUNIDAD VALENCIANA

nova06
Premis de l'empresa valenciana

MINISTERIO
DE EDUCACIÓN
Y CIENCIA

UNIÓN EUROPEA
Fondo Europeo de Desarrollo Regional
Una manera de hacer Europa