Generation of Topic Tag Clouds

Jesús Antonio Soto Velázquez
Undergraduate Program in Software Technology Engineering
Faculty of Mechanical and Electrical Engineering, Universidad Autónoma de Nuevo León
May 2013

Topic Visualization

Tag clouds aid users to recognize at a first glance what a group of various documents is about by displaying the most relevant words or topics.

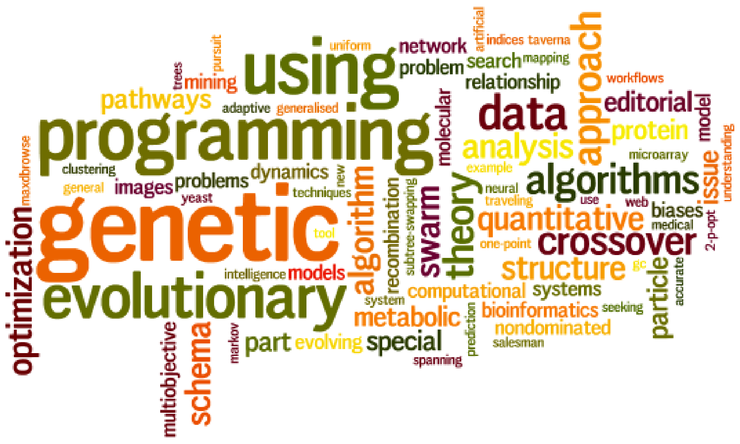

Figure 1: Artistic tag cloud using Wordle

The **objective** is to deploy a web platform which generates tag clouds with meaningful information extracted from a collection of documents. This work focuses on such an insightful *visualization* of the topics in the documents.

Tackling the problem

Several aspects were taken into account when choosing the right information to put in the tag cloud, such as:

Stopword filtering Unimportant words in the given context were discarded. Solved with *String matching algorithms [2]*.

Word stemming Words with the same root were grouped together. Resolved with the *Snowball [6] library*.

Language detection Articles in other languages were to be excluded from the tag clouds. Resolved with the *language-detection [4] library*.

The tag cloud Manage structure and appearance of a tag cloud. Resolved with the *OpenCloud [5] Java library*

Portability As the intention was to reach as many users as possible, a web environment was chosen. Technologies used:*HTML5, CSS, Javascript, Servlets*

Initial solution

The initial solution made use of OpenCloud, a Java library that aids the creation of tag clouds for the web. Using HTML and CSS, the tag cloud was given a simple styling and presentation.

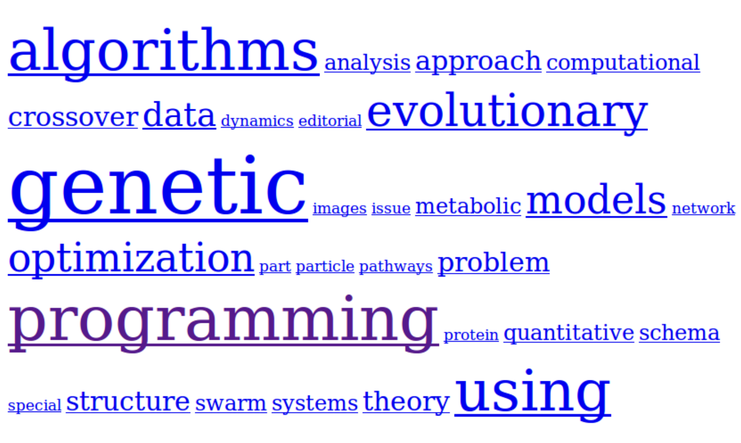

Figure 2: Tag cloud using OpenCloud

Semantic approach

Rather than focusing on the artistic side of tag clouds, like most tag cloud tools do, an approach on **semantic similarity** between topics was taken. As such, the position of each topic is determined by the semantic similarity of itself and its surroundings.

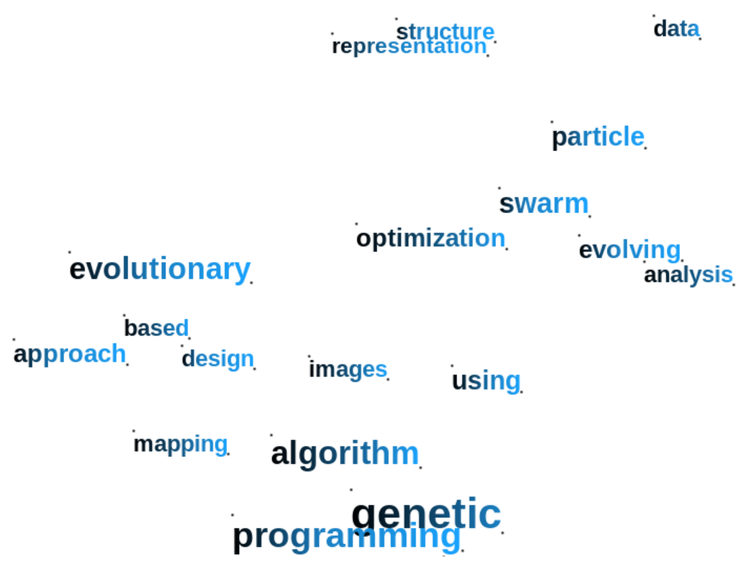

Figure 3: Topic grouping with gradients

It might be of interest knowing how active or inactive the topics have been throughout the years. For such cases, a **two-color gradient** is used, such that the brighter the color, the more active it is.

Indexing documents

In order to quickly search through the documents by typing a keyword, a structure known as an index was used. The tool used to index the groups of articles, Solr, provides a simple interface between the data stored and the means of returning the desired information in a web environment.

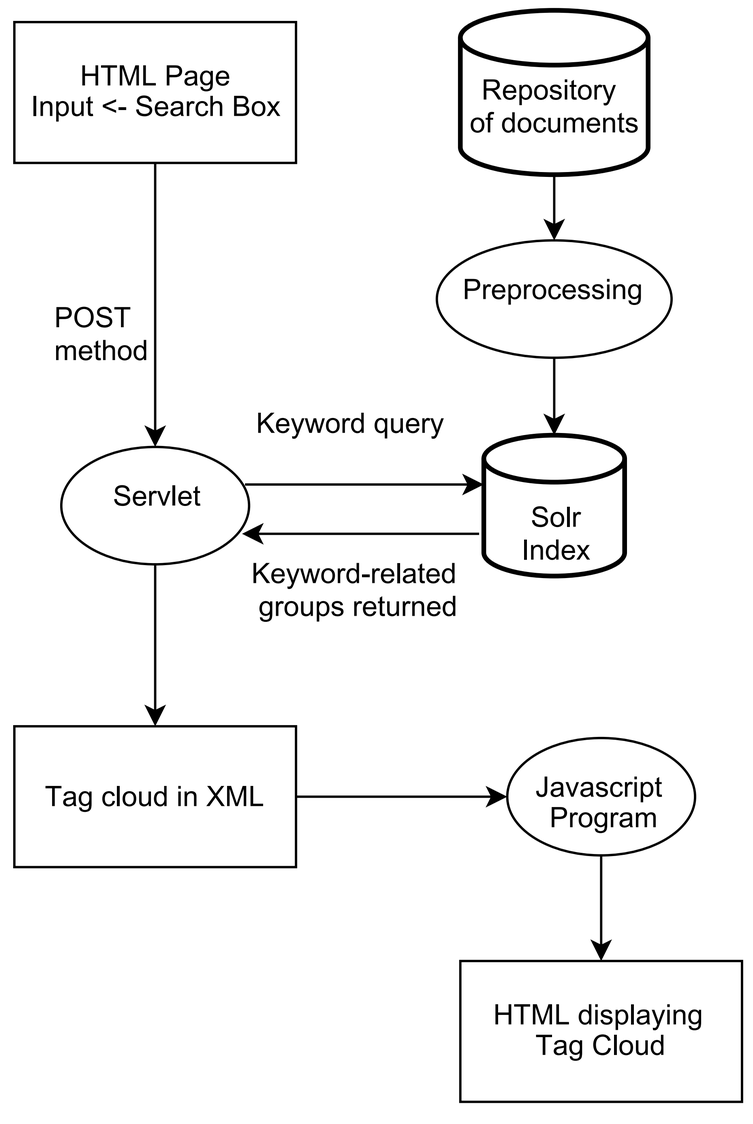

Figure 4: Overall Software Architecture

User interaction

The first step to allow interaction with the tag cloud was through the functionality of clicking a word in the tag cloud and firing an event for that particular word. A match between the cursor's coordinates [3] and each of the words' coordinates is considered.

Future Work

The **motivation** for creating *insightful* topic tag clouds has been shown, and a candidate **architecture** to deploy a web platform with such functionality has been presented.
Key points for further development are:

Data retrieval from tag cloud Through user interaction, more information about the selected topic can be obtained, such as researchers involved.

Generation on the web Given a document from the index in XML, a tag cloud should be generated.

References

[1] Apache Software Foundation. *Apache Solr* http://lucene.apache.org/solr/
[2] Charras, Christian; Lecroq, Thierry. *Exact String Matching Algorithms* http://www-igm.univ-mlv.fr/ lecroq/string/
[3] Lopez, Manuel. *Get the coordinates of a mouse click on Canvas in Javascript*: http://miloq.blogspot.mx/2011/05/coordinates-mouse-click-canvas.html
[4] Nakatani, Shuyo © Cybozu Labs, Inc. *language-detection Java Library* https://code.google.com/p/language-detection/
[5] Cavallo, Marco. *Tag cloud Java library* http://opencloud.mcavallo.org/
[6] Porter, Martin; Boulton, Richard. Stemming Language *Snowball*. http://snowball.tartarus.org/

Acknowledgements

This poster was made under the supervision of Prof. Sara E. Garza and Prof. Elisa Schaeffer. They also provided the clusters of scientific articles grouped by their topics, as well as the functionality of setting the location of each of the tag words in a canvas.
The author also wishes to thank SEP-PROMEP for support (grant 103.5/12/7884).