Generation of Topic Tag Clouds

Jesús Antonio Soto Velázquez

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.

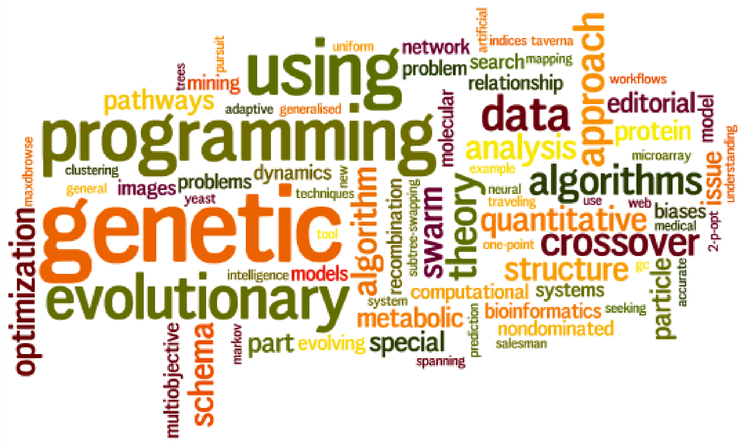

Figure 1: Artistic tag cloud using Wordle

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. Resolved with *String matching algorithms.*

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

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

The tag cloud Manage structure and appearance of a tag cloud. Resolved with the *OpenCloud 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.

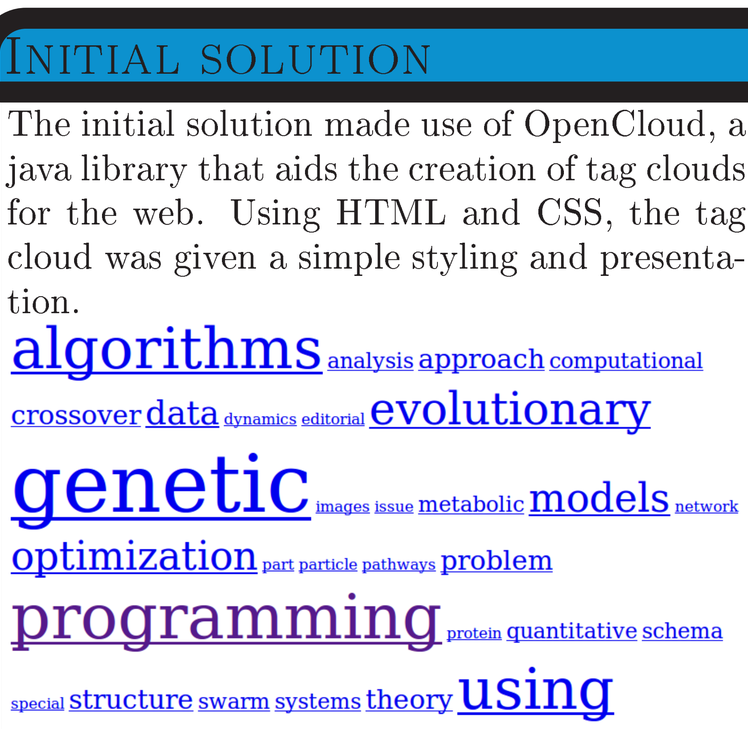

Figure 2: Tag cloud using OpenCloud

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 to represent how active each of the topics have been, or to know if its use has declined after some time. The brighter the color, the more active it is.

Interacting with users

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.

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.

Future Work

Although the project focused on the many components involved, there is much work to be done to integrate these parts into a system for use in the web. Key points to advance 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

- Porter, Martin; Boulton, Richard. Stemming Language *Snowball.* http://snowball.tartarus.org/
- Mcavallo. *Tag cloud Java library* http://opencloud.mcavallo.org/
- Apache Software Foundation. *Apache Solr* http://lucene.apache.org/solr/