Figure out your ideas: introducing Cluster Analysis in IdeaStream Analytics v1.7 release

admin - 7 Aug 2013 at 10:29:51

Following the past updates for our open-source Idea Management System IdeaStream, today we have some new features for the data analytics module that delivers insights into statistics and various analyses about ideas – IdeaStream Analytics.

The big new addition is called **Cluster Analysis** and allows to automatically group ideas based on similarities in term patterns and contextual information. For example: analysis of tags, how frequently are they assigned to ideas, if there are any typical groups of tags used together and how are they different from each other.

The key part is the contextual analysis that not only leads to grouping ideas with similar terms but also allows to highlight how particular group is different from the remaining ideas in other groups.

The Cluster Analysis starts to shine in systems of hundreds or even thousands of ideas, where it is difficult to understand the theme of gathered ideas and community preferences. Currently, we implemented 5 different types of analyses: tag, category, frequent keywords (taken from the Drupal full-text search index), Gi2MO Types characteristics (fixed vocabulary describing innovation types) and Gi2MO Links relationships (supplied by the IdeaStream Similarity module).

Cluster Analysis of Gi2MO Types annotations run for Ubuntu BrainStorm dataset (right click and open in new tab/window to view hi-res image).

To understand how it all works – let's see a simple example of an mockup system with ideas for a supermarket. The picture below shows three groups detected based on cluster analysis of the dataset. The first group appears to be generic fruit ideas, the second group are ideas about fruits of a particular kind (peaches), while the third are ideas about vegetables being onions specifically. This already gives a nice overview of the entire system but further insights show that although generic fruit ideas are favoured by the reviewers and get accepted more often (workflow analysis), the detailed ideas about peaches win the community by a significant degree (top amount of comment and comment per idea ratio).

Explanation of metrics and an example of Cluster Analysis based on Tags.

To aid such rapid analysis of clusters IdeaStream Analytics delivers some calculations and highlights about all clusters. Firstly, we output the theme for each cluster: a full list terms used in those ideas ordered according to use frequency but also highlighted in bold if the term was detected as distinctive for the group in comparison to other groups. Furthermore, each group is described by idea count, comment count etc. supplied with indicators if the group has the top or minimal value in comparison to other clusters.

Under the hood all of this is implemented using Singular Value Decomposition (SVD) used in a similar way as in an algorithm called Latent Semantic Analysis (LSA). In short, SVD allows to replace idea term annotations with new concepts that represent combinations of old terms, thus simplifying the analysis of idea similarity. Furthermore, SVD output allows to project terms and ideas into the same multi-dimensional space – in practice this allows us to draw a scatter plot to observe semantic distances between ideas and terms visually.

Furthermore, on top of SVD, IdeaStream Analytics uses an unsupervised clustering algorithm called k-Means to automatically group the similar ideas together. Also, to make this meaningful for Idea Management, we add a number of custom calculations to deliver some interesting insights and analytics about the discovered idea groups.

The drawback of the entire solution is that it's quite hungry for memory and computational power. As a result, the current implementation allows a number of settings to limit the amount of terms analysed and blacklist some unwanted terms (specially useful for keyword analysis that often will have many unwanted words).

To check all this in action go to the apps section for IdeaStream Analytics, download the version and find out yourself!

Since this is a first release of this feature we welcome any feedback and information about its use in practice (e.g. which statistics are missing, how to better describe cluster theme or any sort of complaints).