2012 16th International Conference on Information Visualisation

# Using clustering to personalize visualization

Mohamed Mouine
Guy Lapalme
*RALI-DIRO Université de Montréal*
*Quebec, Canada*
*Email: mohamed.mouine@umontreal.ca*
*Email: lapalme@iro.umontreal.ca*

***Abstract*—The goal of our work is to propose models or methods to personalize the visualization of a large amount of weather information in a simple way and to make sure that a user can analyze all needed information. We personalize this visualization for each user according to an automatically detected profile based on clustering. Clustering is used to group users who are similar to current user and then set the visualization variables according to the visualizations of those users.**

***Keywords*-clustering; visualization; personalized; weather; user profile**

## I. Introduction

In recent years, we witnessed an explosion of data generated in all the fields of knowledge. The most difficult task being its analysis and exploration. Data mining allows to locate the necessary information. The visual exploration of data can help better assimilate information when combined with a textual description. The process of visualization of information and scientific data seeks to solve the problem of representation of the various types of data for the user, so that the data can be easily communicated and interpreted. In this paper, we present a method to personalize the presentation of a large amount of environmental information produced daily by Environment Canada (EC). This information provides Canadians with up to date information on weather conditions. We aim to present users with weather reports on demand, meeting specific needs of the targeted user.

In order to summarize and analyze large amounts of information, we present a method to generate a visual report automatically (chart, picture, text ...). We want the user to retrieve easily the information without having to scan the whole mass of information. Given the extent of the Canadian territory, EC cannot prepare beforehand a specific bulletins for each need in various formats. We want to create a generator of climatic bulletins to produce reports on request. This generator must summarize a great quantity of information. To enable to the user to choose information to display, the user must have the possibility of interacting with the interface.

The generator must display custom visualizations to users. For that, we must take into account the needs and the preferences of each user. But we have only little information on the users because we do not want to use cookies for privacy concerns. Our method predicts the preferences and the needs of a user based on the history of the preferences of the similar users.

In the following section, we will explain the role of the interactivity in our system. Section III will be devoted to the method to identify the preferences compared to the profile. And in the last section, we will explain how, based on these preferences, we will be able to generate a personalized visualization.

## II. Interaction

Interaction techniques allow the data analyst to interact directly with visualizations and to dynamically change visualizations according to the objectives of exploration. They also make it possible to connect and combine independent visualizations. [1] proposes a classification of the techniques for information visualization and data mining based on the type of data, the techniques of visualization, interaction.

In our field of application (the weather), the user must manage a large amount of information to have a visualization, so he must be able to interact with in order to meet her needs. One must be able to ask more or less details, to filter the data, to reorganize visualization according to preferences, to request data to be displayed differently, etc.

In addition to satisfying the user, we take advantage of the results of this interactivity to save the resulting visualizations with the user profile.

## III. Approach

The information stored according to the interactivity will be used to train the system to predict the preferences of users and generate custom visualizations.

We propose to create user groups based on similarity of their profiles. Each cluster will contain users with a similar profile. A new user will be assigned to one cluster. We consider that users in the same cluster are similar based on their preferences and should influence the result more. For that we set a weight $w = 1/d$ to each user of the same cluster. $d$ is the distance between our user and the other elements of same the cluster.

1550-6037/12 $26.00 © 2012 IEEE
DOI 10.1109/IV.2012.51



CPS Conference Publishing Services

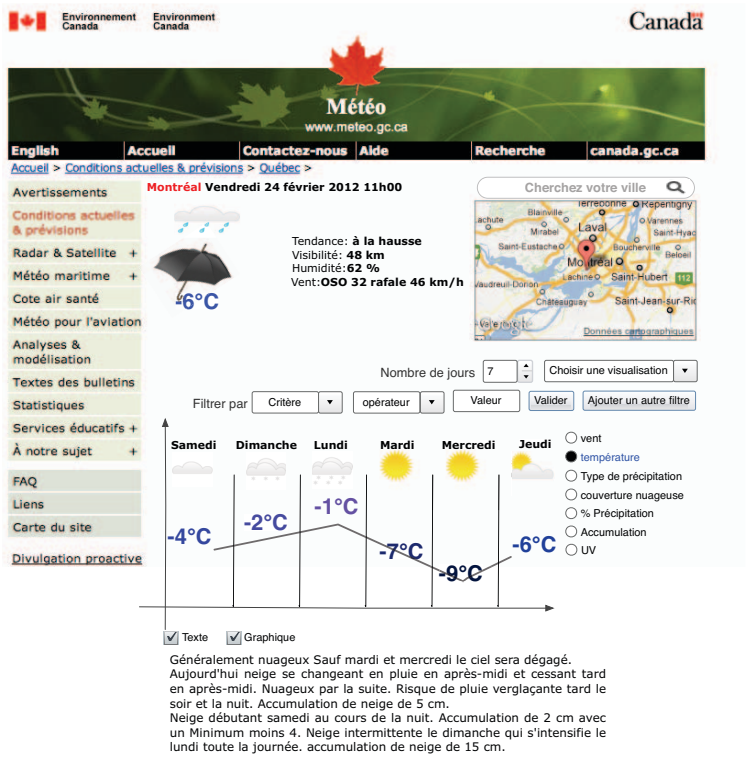


Figure 1. Example of an interactive visualisation: The user can visualize data by choosing visual encodings, he can also filter out data to focus on relevant items. The user is able to select some parameter to highlight, filter or manipulate the visualization.

### A. *User profile*

Our goal is to personalize the presentation of information according to the user. To achieve this, we need to know the preferences of this user. Research on the profiles of users and more precisely the personalization of information was dealt mainly in the field of information retrieval [2].

There are two types of user profiles: determined by a user who sets preferences and needs or built automatically using several techniques such as history, behavior, rules of associations, classification techniques and algorithms of clustering. [3] presents two clustering techniques of user profiles and Web page viewed in order to provide personalized recommendations in real time.

PACT (Profile Aggregations based on Clustering Transactions) gathers in clusters similar transactions of pages consulted by user. This technique can consider other factors to determine the weight of the elements within each profile. These include the distance of page views at the current location of the user on the site or the rank of the profile according to its importance.

Another profile generation method, ARHP(Association Rule Hypergraph Partitionning), computes groups of pages based on the number of times they co-occur in the logs rather than grouping the transaction themselves.

### B. *Visualization*

We must prepare a visualization as a starting point and with which our users will be able to interact. This visualization should follow the standard norms of creation of visualization and is created according to the three stages approach of [4]. The first stage is to identify the principles of design by analyzing existing visualizations. Second is to create a visualization according to the rules and principles of the first stage. The last stage evaluates the visualization.

In the first stage, our analysis is based on the Gestalt principles [5] (proximity, similarity, continuity, closure...) and the categories of interactivity described in [6]. Figure 1 shows one visualization designed following this process. This visualization allows to the user to interact with all its components. He can modify the forcast horizon (in days), change type of the plots and he can choose between graphical visualization, text visualization or both. The user can also explore data directly from visualization and select the desired level of detail. This interaction is important for the personalization method.

## IV. Customizing presentations

We want to generate weather reports personalized and adapted to the needs and preferences of the users.

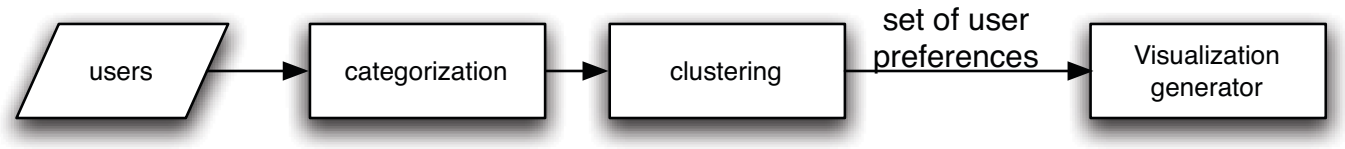


Figure 2. Our approach consists of four steps: 1) categorized the user depending on device used. 2) group users according to profiles similarity 3) set the user preferences 4) use these preferences to generate a customized visualization

Users will be able to modify the presentation according to their tastes, their preferences and their needs. We can improve quality of the presentation generated by learning these preferences. In order that this presentation corresponds to the preference of each user, we begin with a users categorization then an aggregation of the user types and determine for each user a presentation which corresponds to her wishes. Information available about a user, taking into account the privacy concerns, is the following:

- Location based on the geolocalisation of the IP address IP ( longitude and the latitude).
- The prefered language (French or English) based the parameters of his browser.
- The time of connection according to his location.
- The season (autumn, winter, spring or summer).

Other user information, will not be used for clustering, but for categorization:

- Type of operating system and browser.
- The type of the user device.

This information alone is not enough to determine the preferences of the user. The first stage of the method (see figure 2) is to archive (anonymously) interactions of users with the visualizations and especially the type of visualizations they decide to keep last. We consider that the final settings correspond to the preferences of the user. We then categorize users according to the similarity of their information by focusing on the cluster to which the user belongs. Visualizations corresponding to the users in to the same group will be used to determine the preferences. We will balance the rate of similarity based on the distance between the characteristic vectors.

### A. Categorization

Information about users differentiates them from others. We consider categories of users before starting to learn their preferences and predict them for new users. Users will be categorized according to the type of device they use. We chose to categorize users based on this criterion because visualizations vary depending on the device.

### B. Clustering

K-means [7] is a simple algorithm for clustering. It classifies a set of data among a number of clusters fixed a priori. The principal idea is to define K centroids, one for each cluster. We define here K relatively large for more precision because of the vast area of Canada. The next stage is to take each user in our database for which a visualization has already been generated and to associate her with the nearest centroid. This algorithm minimizes an objective function:

$$J = \sum_{j=1}^{k} \sum_{p \in C_i} \|p - m_i\|^2$$

where $\|p-m_i\|^2$ is a measure of the distance chosen between a data point $p$ and the center $m_i$ of the cluster $C_i$, it is an indicator of the distance from the $n$ data points starting from their respective center of the clusters.

In these calculations, we eliminate the differences in scale (order of magnitude) of variables with a normalization transformation. This normalization is:

$$V_{norm} = \frac{V_i - \bar{V}}{V_\sigma}$$

With $\bar{V}$ and $V_\sigma$ are respectively the average and the standard deviation of values of the same parameter for all users.

In our case, the hour variable varies from 1 to 24 and the variable season varies from 1 to 4, both will contribute in the same manner to the distances from which the solution of the clustering will be given. Variables of user profiles belonging to the same category represents the features for the clustering process.

The problem of the clustering by K-means is NP-difficult [8] in the general case. For that, we decided to use this algorithm with a starting corpus (of small size) to determine the classes, then to use the supervised learning by using the method of classification.

### C. Setting of the user preferences

The clustering process allows to group similar users, each one associated with visualization variables. The new visualization for the current user will be influenced by the visualization variables corresponding to the user in the same cluster. The degree of influence is affected by the parameter $d$.

## V. Generation of graphics and text

The graph and the text generated must be synchronized to be complementary and descriptive. For this, we build upon

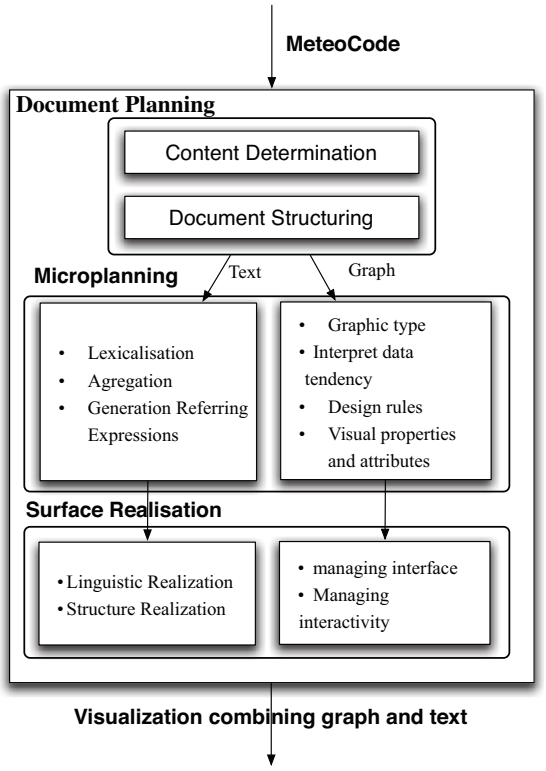


Figure 3. Method of generating combined text and graphics. Meteocode is the input file containing the weather forecasting data in XML format.

the method of [9] describing text generation from a set of data. To create a combined text and graphics generation method (see Figure 3).

Document planning deals with the determination of the content of the document (text and graphics). The information is then structured to decide if it will be produced as text, graphic or a combination of these.

### A. Text generation

Methodology presented in [9] is based on a particular architectural decomposition process of natural language generation in three modules: document planning, microplanning, and the surface realization. Document planning is what is often called "text planning", and includes two subtasks: content determination and document structuring. Microplanning includes aggregation, referring expression generation, and some aspects of lexicalization. The surface realization includes linguistic and structure realization.

In terms of intermediate representations, the output document planning in this model is a document specification, This is a tree made up of information bearing units called messages, often with discourse relations between specified parts of the tree. The output of the microplanning is a specification of the text, It is a tree whose external nodes specify the characteristics of the sentence and whose internal nodes specify the document's logical structure in terms of paragraphs, sections,...

We need to generate the text in both official languages of Canada. For this we use the Java library SimpleNLG-ENFR. This is an adapted version by Pierre-Luc Vaudry from SimpleNLG v4.2 [10] that allows for text in French and English. Figure 4 is an example of the structure for generating "Heavy rain fell on the 27th and 28th" using SimpleNLG-ENFR.

### B. Visual generation

The first stage is to decide the type of the graph to be generated. It depends on several parameters: the preferences of the user deduced from the clustering, the choice of the user or the type of device used. Then, we have to interpret the trend of the data to be able to highlight warnings about potential bad weather. It would allow the users to detect immediately bad meteorological conditions. We also have to specify the rules of design and the properties and visual attributes brought out following to the second stage of the approach of [4].

The last stage is the interface. In this work, we plan to use the SVG technology (Scalable Vector Graphics). The source code of the images SVG is defined in a file XML. All the files of data generated by EC are also in XML. The XML

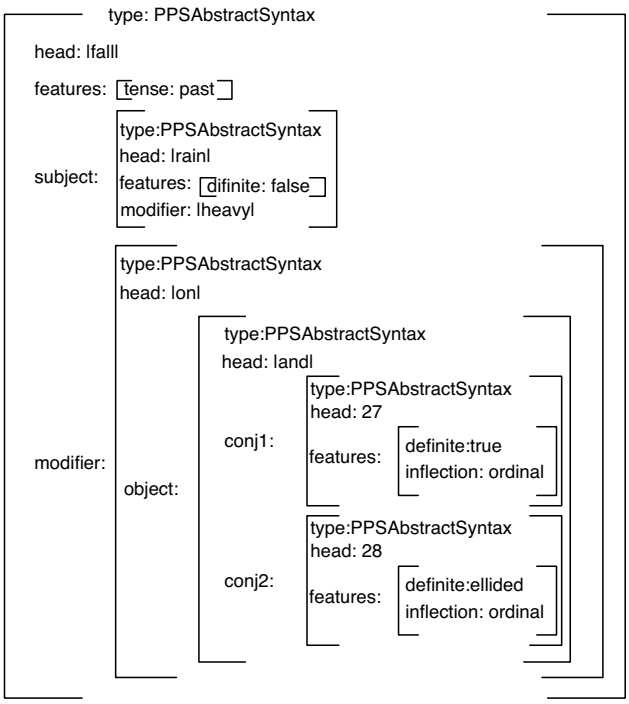


Figure 4. Example of application of the method [9]. The sentence generated is "Heavy rain fell on the 27th and 28th"

family offers a possibility to transform XML documents into other XML documents with different elements, attributes, content, and structure. W3C's recommendation to do so is XSLT (eXtensible Stylesheet Language Transformation). This uniformity of notation is an advantage that allows us to generate visualizations directly from the data using a simple process.

## Conclusions

The goal of our work is to propose models or methods to personalize the visualization of a large amount of information. In our case study, it is necessary to display a great amount of weather information in a simple way and to make sure that a user can analyze all needed information. For that, we personalize this visualization for each user according to his profile detected automatically. We also take into account the evolution of information through time. The new approaches and methods used in this work could also find their application in other fields where a large quantity of information change continuously through time.

The visual presentation and the automatic generation reports can also be applied in the financial negotiation, biocomputing, medicine

## Acknowledgements

We want to thank Alessandro Sordoni for his work in this project that we use as a starting point. We want also to thank Pascal Vincent for his advices concerning the machine learning part.

## References

[1] D. A. Keim, "Information visualization and visual data mining," *IEEE transactions on visualization and computer graphics*, vol. 8, no. 1, p. 1, 2002.

[2] D. Kostadinov, "Personnalisation de l'information et gestion des profils utilisateurs," *Mémoire de DEA PRiSM, Versailles*, 2003.

[3] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa, "Discovery and evaluation of aggregate usage profiles for web personalization," *Data Mining and Knowledge Discovery*, vol. 6, no. 1, pp. 61–82, 2002.

[4] M. Agrawala, W. Li, and F. Berthouzoz, "Design principles for visual communication," *Communications of the ACM*, vol. 54, no. 4, pp. 60–69, 2011.

[5] C. Ware, *Information visualization: perception for design*. Morgan Kaufmann, 2004.

[6] J. Yi, Y. ah Kang, J. Stasko, and J. Jacko, "Toward a deeper understanding of the role of interaction in information visualization," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 13, no. 6, pp. 1224–1231, 2007.

[7] J. Hartigan and M. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[8] S. Dasgupta, "The hardness of k-means clustering," *Techn. Rep. no. CS-2007-0890 (Univ. California, 2007)*, 2008.

[9] E. Reiter and R. Dale, *Building natural language generation systems*. Cambridge Univ Pr, 2000.

[10] A. Gatt and E. Reiter, "Simplenlg: A realisation engine for practical applications," in *Proceedings of the 12th European Workshop on Natural Language Generation*. Association for Computational Linguistics, 2009, pp. 90–93.