# Génération d'une visualisation personnalisée

Mohamed Mouine
RALI-DIRO Université de montréal
mouinemo@iro.umontreal.ca

Résumé. Nous présentons une méthode permettant de calculer les besoins et les préférences d'un utilisateur sur lequel nous avons peu d'information. Nous utilisons ces résultats pour générer un bulletin météorologique personnalisé à chaque utilisateur. La visualisation générée est une combinaison de graphique et de texte. Nous présentons dans ce papier les méthodes utilisées pour calculer les besoins et les préférences des utilisateurs et pour générer la visualisation et le texte.

Mots-Clés : Visualisation personnalisée, Intelligence artificielle, Clustering, génération de texte.

## 1 Introduction

Le défi étudié dans cet article est la présentation personnalisée de l'information à l'utilisateur. Nous présentons une approche qui calcule les besoins et les préférences de l'utilisateur selon son profil détecté automatiquement et génère un rapport personnalisé. Cette approche s'applique dans les domaines où on génère une grande masse d'informations évoluant dans le temps avec des besoins d'utilisateurs diversifiés. Nous avons appliqué cette approche sur un problème réel dans le domaine de la météo. Environnement Canada (EC) produit une masse énorme d'information météorologique (26Mb deux fois par jour). Cette information est utilisée pour fournir aux Canadiens des renseignements à jour sur les conditions météorologiques. Nous devons générer des bulletins météorologiques à la demande. Chaque bulletin doit répondre aux besoins spécifiques de l'usager pour lequel il a été généré. Pour cela, en collaboration avec EC, nous avons créé un générateur de bulletins météorologiques contenant du texte et des graphiques. Cette génération de bulletin doit prendre en compte les besoins spécifiques des usagers.

# 2 Personnalisation de la visualisation

Environnement Canada fournit une grande masse d'informations météorologiques que nous ne pourrons présenter en intégralité dans un seul graphique. Nous proposons dans la section 2.2 une méthode pour présenter une visualisation qui répond le plus possible aux besoins de l'usager. Toutefois, l'utilisateur peut raffiner cette présentation en l'adaptant à ses besoins.

## 2.1 Mieux connaitre l'usager

Dans le domaine de la visualisation, nous nous intéressons généralement à bien présenter l'information que l'utilisateur veut avoir. Une des contributions de notre étude est le profilage automatique des utilisateurs (détaillé dans la section 2.2). Notre système archive l'interaction entre l'utilisateur et la présentation pour jauger les préférences des usagers.

Nous avons proposé dans (Mouine et Lapalme, 2012) une méthode permettant de personnaliser une visualisation par le clustering des profils des utilisateurs. Les visualisations sont générées dynamiquement par notre système qui enregistre chaque configuration finale choisie par l'utilisateur. L'idée de cette approche est de regrouper les utilisateurs similaires à l'utilisateur étudié dans un même cluster. Le clustering est utilisé pour déterminer les paramètres des utilisateurs. La distance entre l'utilisateur actuel et les autres dans un cluster sera utilisée pour calculer la similarité. Le taux de similarité nous donne une idée de ce que pourraient être les besoins et les préférences de l'utilisateur actuel en tenant compte de la similitude avec les autres. Le taux de similarité est utilisée pour pondérer les besoins et les préférences. Nous avons fixé un seuil minimum pour chaque critère (besoins et préférences). Pour décider quels sont les paramètres qui seront pris en considération dans la visualisation (voir figure 1), on ne tient compte que des résultats supérieurs au seuil.

## 2.2 Personnalisation des présentations

Nous voulons générer des bulletins météorologiques personnalisés et adaptés aux besoins et préférences des utilisateurs.

Notre système est interactif. Les utilisateurs pourront modifier la présentation selon leurs goûts, leurs préférences et leurs besoins. Nous pouvons améliorer la qualité de la présentation générée en apprenant ces préférences. Pour que cette

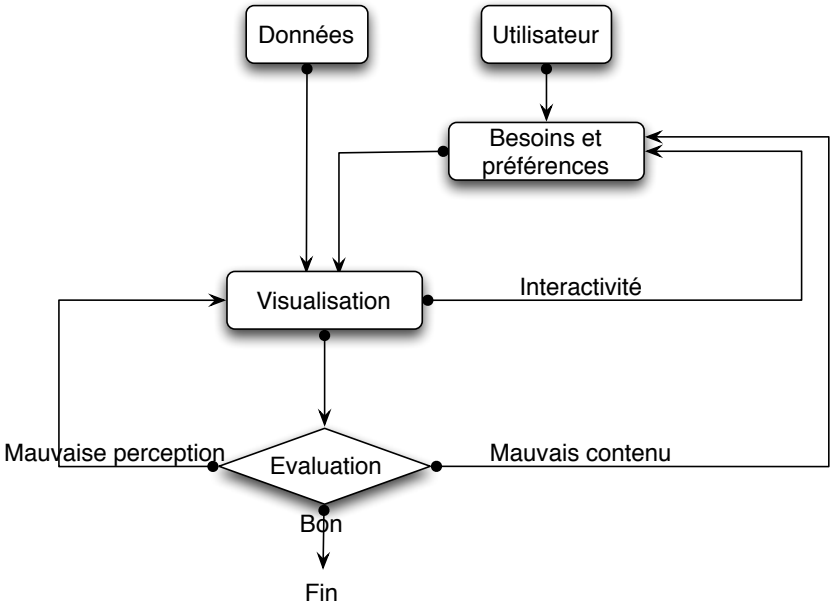


FIGURE 1 – Notre approche pour créer des visualisations personnalisées . L'approche consiste à préparer les données à partir des données brutes, déterminer les besoins et les préférences des utilisateurs, créer la visualisation selon les besoins et les préférences de l'utilisateur et enfin évaluer et améliorer la visualisation

présentation s'approche toujours le plus des préférences de chaque utilisateur, nous générons à chaque utilisateur une présentation basée sur les besoins et préférences des utilisateurs avec des profils similaires. Les informations que nous avons sur notre utilisateur (sans toucher à sa vie privée) sont :

– Son emplacement (à l'aide de la géolocalisation de son adresse IP).
– La langue qu'il préfère (français ou anglais) à l'aide des paramètres de son navigateur.
– L'heure actuelle (lors de sa connexion) selon son emplacement.
– La saison pendant laquelle nous sommes (automne, hiver, printemps ou été).

Ces informations ne nous permettent pas de prévoir les préférences de l'utilisateur. La première étape de la méthode que nous proposons consiste à archiver (anonymement) les interactions appliquées par les utilisateurs sur nos visualisations et surtout les résultats finaux. Nous considérons que, s'il y a interaction, le résultat final répond aux préférences de l'utilisateur. En seconde étape, nous regroupons les utilisateurs selon la similarité de leurs informations en nous intéressant maintenant au cluster dans lequel se trouve notre utilisateur. Les visualisations correspondantes aux utilisateurs appartenant au même groupe que

notre utilisateur seront utilisées pour dégager les préférences des utilisateurs « similaires » à notre utilisateur. Nous pondérons le taux de similarité en nous basant sur la distance entre les vecteurs caractéristiques.

## 2.3 Clustering

L'algorithme K-means (Hartigan et Wong, 1979) est un des plus simples algorithmes d'apprentissage non supervisé qui résolvent le problème du clustering. La procédure classe un ensemble de données parmi un certain nombre K de clusters fixé a priori. L'idée principale est de définir K centroïdes , un pour chaque cluster. Nous définissons ici K relativement grand pour plus de précision étant donnée la vaste surface du Canada. La prochaine étape est d'associer chaque utilisateur archivé dans notre base de données (pour qui nous avons déjà généré une visualisation) au centroïde le plus proche. Cet algorithme minimise une fonction objective :

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i^{(j)} - c_j\|^2$$

où $\|x_i^{(j)} - c_j\|^2$ est une mesure de distance choisie entre un point de données $x_i^{(j)}$ et le centre $c_j$ du cluster , c'est un indicateur de la distance des points de données $n$ à partir de leurs centres respectifs des clusters.

Dans ces calculs, nous « éliminons » les différences d'échelle (ordre de grandeur) des variables grâce à une transformation de normalisation. Cette normalisation est la suivante :

$$V_{norm} = \frac{V_i - V_{moyenne}}{EcartType}$$

Dans notre cas par exemple la variable heure varie de 1 à 24 et la variable saison varie de 1 à 4, les deux vont contribuer de la même manière aux distances à partir desquelles la solution du clustering sera déterminée.

**Algorithme**

1. Placer les K points dans l'espace représenté par les objets qui sont en cluster. Ces points représentent les centroïdes des groupes initiaux.
2. Attribuer à chaque objet le groupe qui a le plus proche centroïde.
3. Lorsque tous les objets ont été assignés, recalculer les positions des centroïdes K.

4. Répétez les étapes 2 et 3 jusqu'à ne plus avoir de déplacement de centroïdes.

Cela produit une séparation des utilisateurs en groupes à partir de laquelle la métrique à minimiser peut être calculée.

# 3 Génération de la visualisation

Basé sur l'étude des visualisations à partir des sites Web des concurrents d'EC (NOAA, Météo France, meteoblue et moteurs de recherche), les principes et les concepts présumés dans ce domaine, en utilisant les principes de perception (Ware, 2012) et techniques d'interaction (Yi *et al.*, 2007), notre système produit un bulletin météorologique [1] à l'emplacement initialement déduit de l'adresse IP (voir fig.2) avec les besoins et préférences de l'utilisateur à compter du regroupement des utilisateurs similaires.

Contrairement aux rapports, contenant des résumés statiques, préparés par EC, cette visualisation contient des informations plus détaillées. Cela est possible parce que nous extrayons dynamiquement toutes les informations des 26Mo de données relatives au profil de l'utilisateur. Dans la visualisation montrée sur la figure 2, un utilisateur peut voir ① la température : chaque point est la température de l'heure correspondante et indiquant les limites minimales et maximales de la ligne de la température pour la période choisie, ② couverture nuageuse, ③ type de précipitations (pluie ou neige), la probabilité de précipitations, l'humidité, le vent (non affiché sur cette figure) et / ou ④ accumulation d'un emplacement spécifique. À première vue, l'utilisateur peut voir et avoir une idée de la tendance générale de la température. Les lignes maximum et minimum montrent les limites de la température pour la période choisie. L'utilisateur comprend s'il y aura de la neige (ou pluie), quand, combien, combien de temps ... Si le vent est affiché (non représenté sur la figure 2), la vitesse et la direction du vent sont présentées par la direction et la taille d'une flèche.

Nous pouvons enrichir la visualisation en ajoutant des informations calculées (moyenne, maximale, minimale ...) afin de simplifier la perception par l'utilisateur.

1. www-etud.iro.umontreal.ca/~mouinemo/meteo

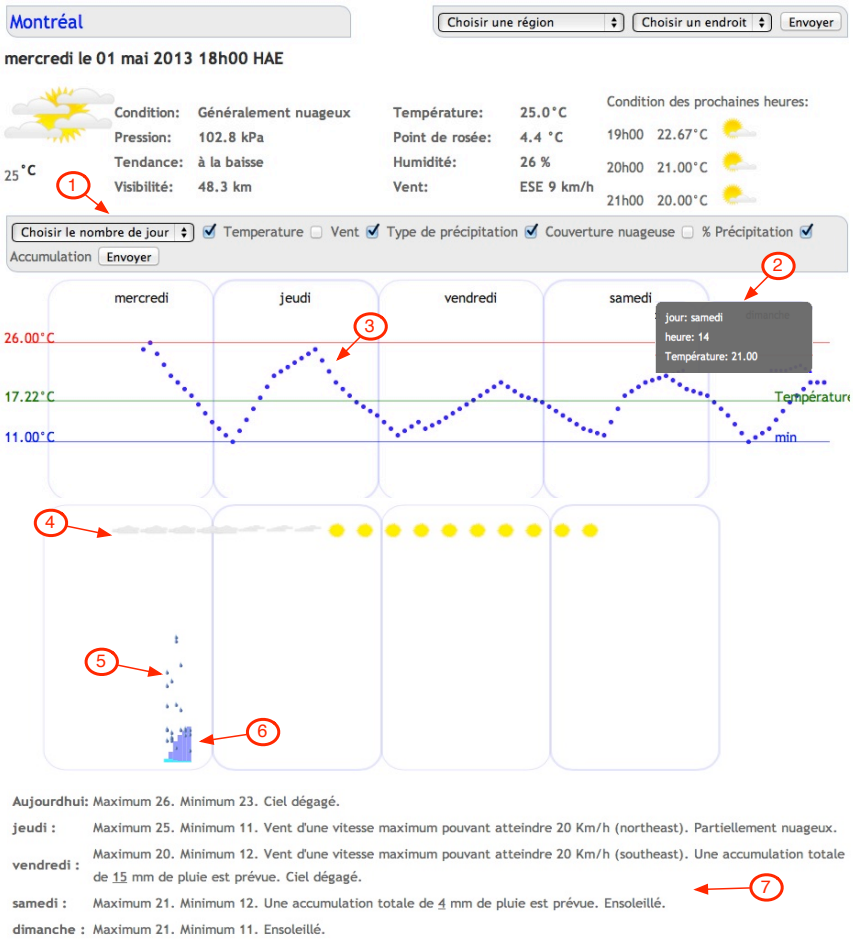


FIGURE 2 – Visualisation générée montrant certains types d'informations qui peuvent être sélectionnés par l'utilisateur. Les numéros encerclés sont ajoutés ici à titre de référence. Avec le menu des préférences ①, l'utilisateur peut sélectionner une province et une ville, modifier le nombre de jours à afficher et les paramètres à afficher. ② L'utilisateur peut avoir plus de détails sur les aspects en mettant le curseur dessus (info-bulle). ③ Température : la façon dont la température est affichée donne une idée de la tendance. Les lignes des maximum et minimum montrent les limites de la température pour la période choisie. ④ La couverture nuageuse . ⑤ type de précipitations : pluie ou neige. La quantité affichée est proportionnelle à la quantité prévue. ⑥ Accumulation : affiche la zone d'accumulation et les précipitations totales (mm) ou de la neige (cm). ⑦ est un texte décrivant l'état de la météo généré automatiquement.

# 4 Génération de texte

Nous avons préparé un module de génération de texte qui peut être appelé « gabarits de phrases ». Ce module génère des portions de phrases pré-préparés selon l'entrée. L'expression générée décrivant l'état de la météo est divisée en plusieurs parties, chaque partie désignant un paramètre de la visualisation. Le choix de la partie utilisée dépend de la valeur numérique de l'entrée. Les paramètres retenus pour la visualisation (résultat du clustering) jouent un rôle important dans la détermination du contenu pour la génération de texte. Le texte ne décrit que les paramètres retenus par le système. Le générateur de texte ignore aussi les paramètres dont les valeurs sont jugées de faible importance (exemple : vitesse du vent = 5 km/h). De même, les paramètres qui ne sont pas retenus par le système peuvent être utilisés si leurs valeurs sont jugées de haute importance et seront affichés sous forme d'avertissement (exemple : valeur totale prévue d'une averse de neige > 25 cm). Pour construire la phrase le système prépare des propositions pour chaque paramètre retenu selon les critères décrits précédemment et les affiche dans la zone de description du jour correspondant.

Pour amèliorer notre système nous prévoyons automatiser la génération de texte dans le future. Nous avons besoin de générer le texte dans les deux langues officielles du Canada. Pour cela, nous utiliserons la bibliothèque Java SimpleNLG-ENFR. Il s'agit d'une version adaptée par Pierre-Luc Vaudry de SimpleNLG v4.2 (Gatt et Reiter, 2009) qui permet de générer du texte en français et en anglais.

# 5 Conclusion

Le but de notre travail est de proposer des méthodes pour personnaliser la visualisation d'une grande quantité d'informations. Dans notre application spécialisée, il faut afficher une grande quantité d'informations météorologiques d'une manière simple et s'assurer qu'un usager ait toutes les informations dont il a besoin, et qu'il puisse les analyser. Pour cela, nous proposons de personnaliser la visualisation pour chaque usager en fonction de son profil que nous devrions détecter automatiquement. Nous calculons les besoins et les préférences de chaque utilisateur en se basant sur les choix des utilisateurs similaires. Nous utilisons également le texte et les graphiques dans cette visualisation parce qu'aucun d'eux ne peut à lui seul projeter exactement et d'une manière simple l'information. Afin de projeter l'information de manière simple et cohérente,

nous proposons de nouvelles méthodes combinant les textes et les graphiques dans une visualisation.

## Références


GATT, A. et REITER, E. (2009). Simplenlg : A realisation engine for practical applications. *In Proceedings of the 12th European Workshop on Natural Language Generation*, pages 90–93. Association for Computational Linguistics.

HARTIGAN, J. A. et WONG, M. A. (1979). Algorithm as 136 : A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

MOUINE, M. et LAPALME, G. (2012). Using clustering to personalize visualization. *In Information Visualisation (IV), 2012 16th International Conference on*, pages 258–263. IEEE.

REITER, E. et DALE, R. (2000). *Building natural language generation systems*. Cambridge university press.

WARE, C. (2012). *Information visualization : perception for design*. Morgan Kaufmann Pub.

YI, J. S., ah KANG, Y., STASKO, J. T. et JACKO, J. A. (2007). Toward a deeper understanding of the role of interaction in information visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1224–1231.