Université de Montréal

Présentation personnalisée des informations environnementales

par
Mohamed Mouine

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Rapport pour la partie orale
de l'examen pré-doctoral

Avril, 2012

Université de Montréal
Faculté des études supérieures

Cet examen pré-doctoral intitulé:

Présentation personnalisée des informations environnementales

présenté par:

Mohamed Mouine

a été évalué par un jury composé des personnes suivantes:

Esma Aïmeur,	président-rapporteur
Guy Lapalme,	directeur de recherche
Philippe Langlais,	membre du jury

Examen accepté le: . . . . . .

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

CHAPITRE 1

PRÉSENTATION DU PROBLÈME

Ces dernières années, il y a eu une explosion du volume de données générées dans tous les domaines de la connaissance, la tâche la plus difficile étant devenue leur analyse et leur exploration. La fouille de données nous permet de localiser les informations nécessaires. La visualisation de l'information et l'exploration visuelle des données peuvent aider à mieux assimiler l'information quand elle est combinée avec une description textuelle. Le processus de visualisation d'information et de données scientifiques cherche à résoudre le problème de représentation des différents types de données pour l'utilisateur, de sorte que les données peuvent être facilement communiquées et interprétées. Dans cette thèse, nous développons des méthodes pour automatiser l'exploration d'informations de type environnemental et leur présentation de la façon la plus simple (Mouine, 2011).

Notre partenaire dans ce projet est Environnement Canada (EC) qui produit une masse énorme d'information météorologiques de façon continue. Cette information est utilisée pour fournir aux Canadiens des renseignements à jour sur les conditions météorologiques. Nous devons pouvoir présenter aux usagers des bulletins météorologiques à la demande. Chaque bulletin doit répondre aux besoins spécifiques de l'usager pour lequel il a été généré. Pour cela, nous allons créer un générateur de bulletins météorologiques contenant du texte et des graphiques. Cette génération du bulletin doit prendre en compte le type de périphérique de sortie et les besoins spécifiques des usagers. Dans le domaine de la météo, nous sommes confrontés à une autre contrainte, nous devons présenter l'information à jour pour l'usager le plus rapidement possible.

Les données météorologiques et leur visualisation indiquant les intempéries comme la neige ou le verglas, les tornades et les ouragans ont besoin d'être personnalisés pour les différents types d'usagers.

Dans le but de résumer et d'analyser une grande quantité d'information, nous comptons présenter une méthode qui génère automatiquement un rapport visuel (graphe, image,

texte...). Nous voulons, par cette approche, permettre à l'usager en toute simplicité de récupérer l'information utilisée dans la génération de ce rapport sans avoir besoin de consulter toute la masse d'information. Étant donné l'étendue du territoire canadien, EC ne peut préparer au préalable des bulletins spécifiques pour chaque besoin d'autant plus que ces bulletins doivent être dans différents formats. Pour cela, nous voulons créer un générateur de bulletins climatiques qui produira des bulletins sur demande. Ce générateur doit résumer une grande quantité d'information. Le contenu du bulletin résultant sera une combinaison de texte et de graphique. Pour permettre à l'usager de choisir l'information à afficher, notre système doit être interactif et l'usager doit avoir la possibilité d'interagir avec notre interface.

1.1 MétéoCode

Le RALI[1] est impliqué dans un projet[2] en collaboration avec Environnement Canada, qui publie déjà une grande quantité d'informations météorologiques sous forme XML, ce type d'information est appelée MeteoCode (figure1.1). Ce fichier contient, en plus des avertissements, toutes les valeurs des paramètres météo à toutes les heures prévues (la température, le vent, la quantité et le type de précipitation...).

Un affichage sélectif de ces informations personnalisées permettrait à EC de fournir au public des meilleures prévisions ciblées dans le temps et l'espace que celles produites actuellement (figure 2.7) qui présentent certains problèmes que nous allons essayer de résoudre :

- Ces prévisions sont limitées à quelques dizaines de mots trouvés dans les prévisions météorologiques régionales.
- Les icones utilisées sont génériques pour toute la journée.
- L'information présentée est la même pour tous les usagers.

[1] Le RALI réunit des informaticiens et des linguistes d'expérience dans le traitement automatique de la langue. Il est le plus important laboratoire universitaire dans le domaine au Canada.

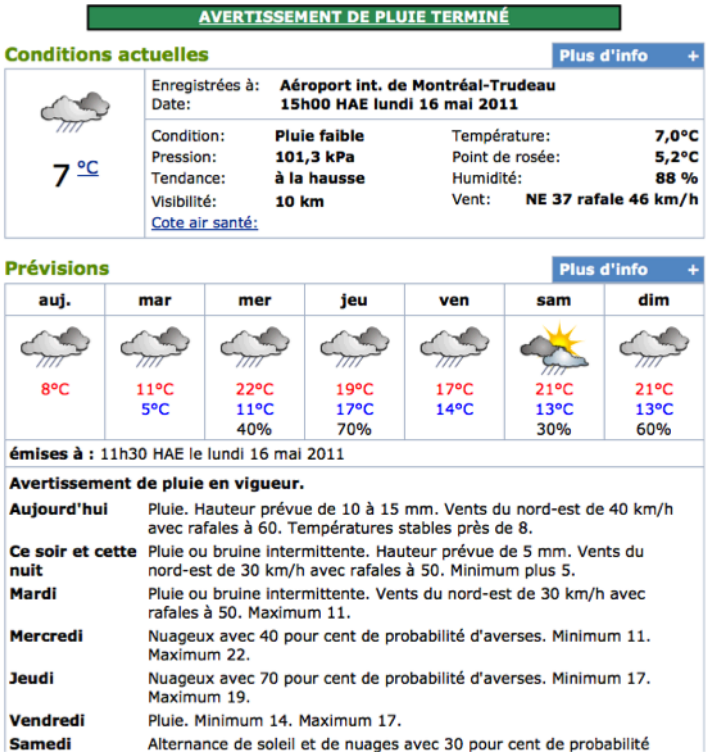
[2] http ://rali.iro.umontreal.ca/EnvironmentalInfo/index.fr.html

Figure 1.1 – Exemple de fichier MétéoCode : les information qui aparaissent dans cette exemples sont : les details de la date (<dateTime >), les details de l'emplacement (<location>), les valeurs de la codition courante de la météo (<currentConditions>) et les prévisions des jours suivants (<forecastGroup>).

```
– <siteData xsi:noNamespaceSchemaLocation="http://dd.weatheroffice.gc.ca/citypage_weather/schema/site.xsd">
  – <license>
      http://dd.weatheroffice.gc.ca/doc/LICENCE_GENERAL.txt
    </license>
  + <dateTime name="xmlCreation" zone="UTC" UTCOffset="0"></dateTime>
  – <dateTime name="xmlCreation" zone="HAE" UTCOffset="-4">
      <year>2011</year>
      <month name="septembre">09</month>
      <day name="vendredi">23</day>
      <hour>11</hour>
      <minute>14</minute>
      <timeStamp>20110923111400</timeStamp>
      <textSummary>23 septembre 2011 11h14 HAE</textSummary>
    </dateTime>
  – <location>
      <continent>Amérique du Nord</continent>
      <country code="ca">Canada</country>
      <province code="qc">Québec</province>
      <name code="s0000635" lat="45.52N" lon="73.65O">Montréal</name>
      <region>Montréal métropolitain - Laval</region>
    </location>
    <warnings/>
  – <currentConditions>
      <station code="yul" lat="45.47N" lon="73.75O">Aéroport int. de Montréal-Trudeau</station>
    + <dateTime name="observation" zone="UTC" UTCOffset="0"></dateTime>
    + <dateTime name="observation" zone="HAE" UTCOffset="-4"></dateTime>
      <condition>Généralement ensoleillé</condition>
      <iconCode format="gif">01</iconCode>
      <temperature unitType="metric" units="C">20.1</temperature>
      <dewpoint unitType="metric" units="C">13.2</dewpoint>
      <pressure unitType="metric" units="kPa" change="0.03" tendency="à la hausse">102.1</pressure>
      <visibility unitType="metric" units="km">24.1</visibility>
      <relativeHumidity units="%">64</relativeHumidity>
    – <wind>
        <speed unitType="metric" units="km/h">0</speed>
        <gust unitType="metric" units="km/h"/>
        <direction/>
        <bearing units="degrees">0</bearing>
      </wind>
    </currentConditions>
  – <forecastGroup>
    + <dateTime name="forecastIssue" zone="UTC" UTCOffset="0"></dateTime>
    + <dateTime name="forecastIssue" zone="HAE" UTCOffset="-4"></dateTime>
    + <regionalNormals></regionalNormals>
    – <forecast>
        <period textForecastName="Aujourd'hui">vendredi</period>
      – <textSummary>
          Ensoleillé. Devenant alternance de soleil et de nuages en mi-journée. Maximum 22. Indice UV de 6 ou élevé.
        </textSummary>
      – <cloudPrecip>
        – <textSummary>
            Ensoleillé. Devenant alternance de soleil et de nuages en mi-journée.
```


Compte tenu de la taille du Canada, ces bulletins doivent rester généraux et uniformes et ne peuvent pas présenter tous les détails disponibles dans le MeteoCode. Déjà, plus de

Figure 1.2 – Le site officiel d'Environnement Canada : http ://www.meteo.gc.ca. Pour arriver à cette page (une ville précise), l'usager doit sélection la langue (français ou anglais) puis selectionner la ville de son choix sur une carte ou dans une liste.



AVERTISSEMENT DE PLUIE TERMINÉ

Conditions actuelles — Plus d'info +

7 °C

Enregistrées à: **Aéroport int. de Montréal-Trudeau**
Date: **15h00 HAE lundi 16 mai 2011**

Condition:	**Pluie faible**	Température:	**7,0°C**
Pression:	**101,3 kPa**	Point de rosée:	**5,2°C**
Tendance:	**à la hausse**	Humidité:	**88 %**
Visibilité:	**10 km**	Vent:	**NE 37 rafale 46 km/h**

Cote air santé:

Prévisions — Plus d'info +

auj.	mar	mer	jeu	ven	sam	dim
8°C	11°C 5°C	22°C 11°C 40%	19°C 17°C 70%	17°C 14°C	21°C 13°C 30%	21°C 13°C 60%

émises à : 11h30 HAE le lundi 16 mai 2011

Avertissement de pluie en vigueur.

Aujourd'hui	Pluie. Hauteur prévue de 10 à 15 mm. Vents du nord-est de 40 km/h avec rafales à 60. Températures stables près de 8.
Ce soir et cette nuit	Pluie ou bruine intermittente. Hauteur prévue de 5 mm. Vents du nord-est de 30 km/h avec rafales à 50. Minimum plus 5.
Mardi	Pluie ou bruine intermittente. Vents du nord-est de 30 km/h avec rafales à 50. Maximum 11.
Mercredi	Nuageux avec 40 pour cent de probabilité d'averses. Minimum 11. Maximum 22.
Jeudi	Nuageux avec 70 pour cent de probabilité d'averses. Minimum 17. Maximum 19.
Vendredi	Pluie. Minimum 14. Maximum 17.
Samedi	Alternance de soleil et de nuages avec 30 pour cent de probabilité

1000 bulletins météorologiques qui présentent la météo du Canada sont émis deux fois par jour.

Selon les informations dans le MeteoCode, nous voulons développer un générateur de bulletins climatiques pour une adresse ou un code postal donné par l'utilisateur. En outre, les informations météorologiques régionales doivent également être mises à disposition dans différents modes : graphique, web, radio-météo et répondeurs automatiques. Un objectif important de notre projet est d'étudier le développement d'approches novatrices pour la communication d'informations météorologiques pertinentes à l'utilisateur tout en tenant compte de l'heure du jour et de l'agrégation géographique.

Etant donné que le Météocode est déjà au format XML validé par un schéma XML, nous sommes convaincus que l'entrée est facilement analysable. Ainsi, nous nous concentrerons sur la détermination de la manière la plus appropriée de présenter les données

d'une façon significative selon le type de périphérique de sortie. Étant donnée la taille des données, nous devrons mettre au point des techniques à usage spécial pour l'agrégation des données dans l'espace et dans le temps.

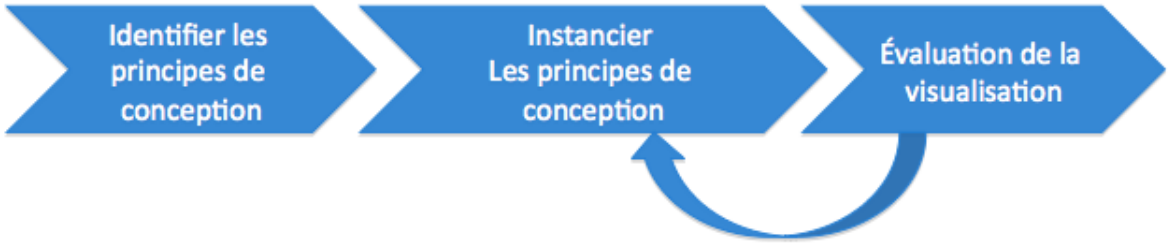
1.2 Approche

Nous avons choisi de suivre les étapes proposées dans Agrawala et al. (2011). Comme indiqué dans la figure 1.3, la première étape est l'analyse des visualisations conçues à la main. Nous pensons que le mieux est de commencer à analyser le site actuel d'Environnement Canada et de ses concurrents ainsi que d'autres sites internationaux. Les sites que nous analysons pour cette première étape sont :

- Météo Canada figure 2.7 (le site actuel d'Environnement Canada).
- MétéoMédia (Un site privé canadien de la météo).
- NOAA (Le site officiel de la météo au États-unis).
- Météo France (Le site officile de la météo en France).
- Météo Blue (Un site suisse privé de la météo mondiale).

Après notre analyse de ces visualisations conçues à la main, nous avons essayé d'extraire et de dégager des règles et des principes pour la visualisation de la météo. Parmi les principes déduits, nous pouvons citer que l'affichage des avertissements est fondamental pour un système d'information météorologique. Un autre principe qui pourra faciliter l'interprétation de notre visualisation est la combinaison du texte et des graphiques. Le principe que nous considérons comme le plus important est la personnalisation de la visualisation pour chaque usager en tenant compte de son profil qui doit être détecté automatiquement et sans atteinte à sa vie privée et à ses informations personnelles. Pour cette raison, nous n'utiliserons pas de témoin de connexion (cookies). Nous pouvons toutefois se baser sur plusieurs astuces pour profiler notre usager. Par exemple, nous pouvons détecter son emplacement à l'aide de son adresse IP. Nous pouvons aussi déterminer la

Figure 1.3 – Les étapes de l'approche de Agrawala et al. (2011) pour la conception d'une visualisation



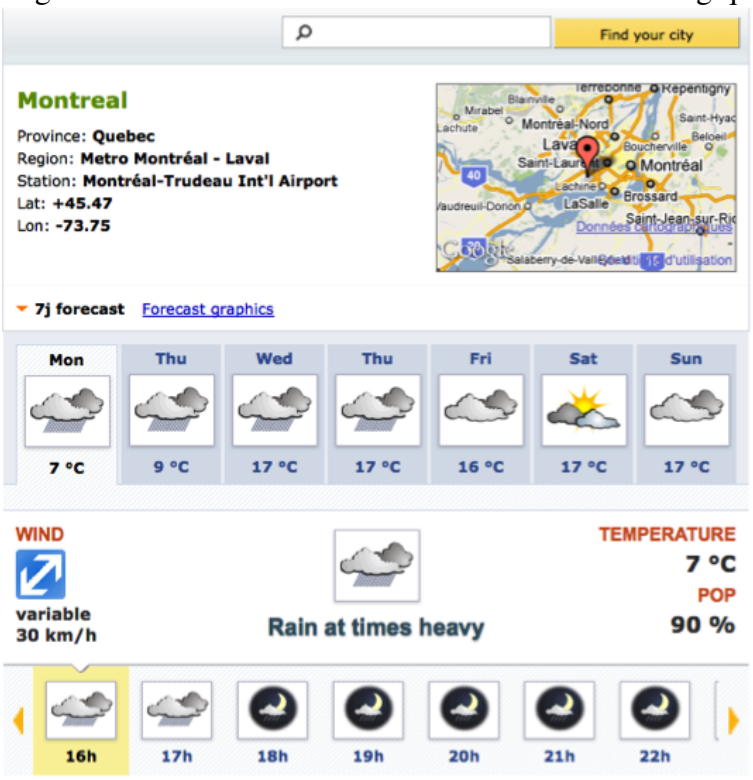
langue d'affichage en détectant la langue qu'il utilise dans son navigateur . Nous pouvons se baser aussi sur son comportement sur le site en se basant sur l'étude faite dans Cadez et al. (2000).

Pour la deuxième étape, nous devons appliquer les principes de conception identifiés à la première étape. Nous pourrons prendre comme point de départ des expériences menées au cours des deux dernières années par les membres du RALI pour illustrer le type d'information disponible chez EC, des prototypes web ont été développés pour afficher l'information météorologique graphiquement en utilisant Protovis qui est basé sur *Scalar Vector Graphics* (SVG) et un autre en utilisant des informations alphanumériques, mais placé géographiquement à l'aide de Google Maps. Une troisième expérience a été réalisée en utilisant jqPlot. Ce dernier est un plugin jQuery pour créer des graphiques. Ces expérimentations offrent de nombreuses représentations ainsi qu'une interactivité poussée avec l'utilisateur. La figure 1.4[3] montre un exemple de résultat des expérimentations réalisées au RALI. Dans cette visualisation, nous trouvons les conditions météorologiques et les prévisions pour chaque heure de la semaine. Nous pouvons aussi choisir l'emplacement directement sur une carte ou en indiquant le nom de la ville.

Cela a permis d'expérimenter différentes manières de combiner les informations publiées quotidiennement par Environnement Canada avec d'autres approches basées sur le Web. Bien que ces prototypes ne soient pas mis en production, ils ont montré la possibi-

[3]Ce site web permettant la visualisation des informations météorologiques a été conçu par Alessandro Sordoni pour le projet météo en collaboration avec Environnement Canada (http ://www-etud.iro.umontreal.ca/ sordonia/deploy/prototypeV2/)

Figure 1.4 – Visualisation des informations météorologiques



lité d'intégrer l'information environnementale avec les applications Web de sorte qu'elle devienne plus accessible et utile.

La troisième étape de l'approche de Agrawala et al. (2011) que nous essayons d'appliquer à notre problème est l'évaluation des visualisations conçues en se basant sur les principes obtenus suite à la première étape. Nous pouvons mesurer la rétroaction qualitative des usagers avec des interviews et la rétroaction quantitative à l'aide des statistiques d'utilisation. Nous pouvons également réaliser des études plus formelles des usagers afin de vérifier dans quelle mesure nos visualisations améliorent l'interprétation de l'information. Les critères d'évaluation doivent quantifier l'efficacité de certains aspects de la visualisation. Notre visualisation doit être expressive et donc présenter toute l'information dont nous avons besoin et seulement cette information. Elle doit être aussi efficace et donc peut être interprétée avec précision et rapidité.

1.3 Graphique

Pour utiliser des graphiques de manière efficace dans la génération automatique de rapports, nous songeons nous inspirer du générateur de rapports PostGraphe (Fasciano et Lapalme, 2000) qui génère des rapports statistiques contenant du texte et des graphiques en se servant d'une description annotée des données à présenter. Ainsi l'utilisateur peut indiquer au système ses intentions, les types de données à présenter et les relations entre les données.

Compte tenu de la grande variété de périphériques disponibles (Web, texte, télévision, ordinateurs de poche, etc.), adapter le Météocode pour chaque périphérique de sortie devient prohibitif. D'autre part, la même information ne devrait pas être présentée exactement de la même façon sur tous les périphériques. Chaque type de dispositif apporte ses propres contraintes et offre de nouvelles possibilités. Ce faisant, l'information doit être accessible pour tous les types de périphériques tout en s'assurant que le sens de l'information reste intact.

Nous aurons également besoin de développer de bonnes techniques pour produire des résumés en langage naturel et pour cela nous nous appuierons sur les résultats du projet SumTime Yu et al. (2007). Ce projet a mis au point une architecture pour la production de courts résumés des données de séries chronologiques. Nous voulons nous inspirer du modèle utilisé pour choisir les mots qui seront utilisés dans le résumé.

1.4 Résumer : bonne prévision et précision de l'emplacement

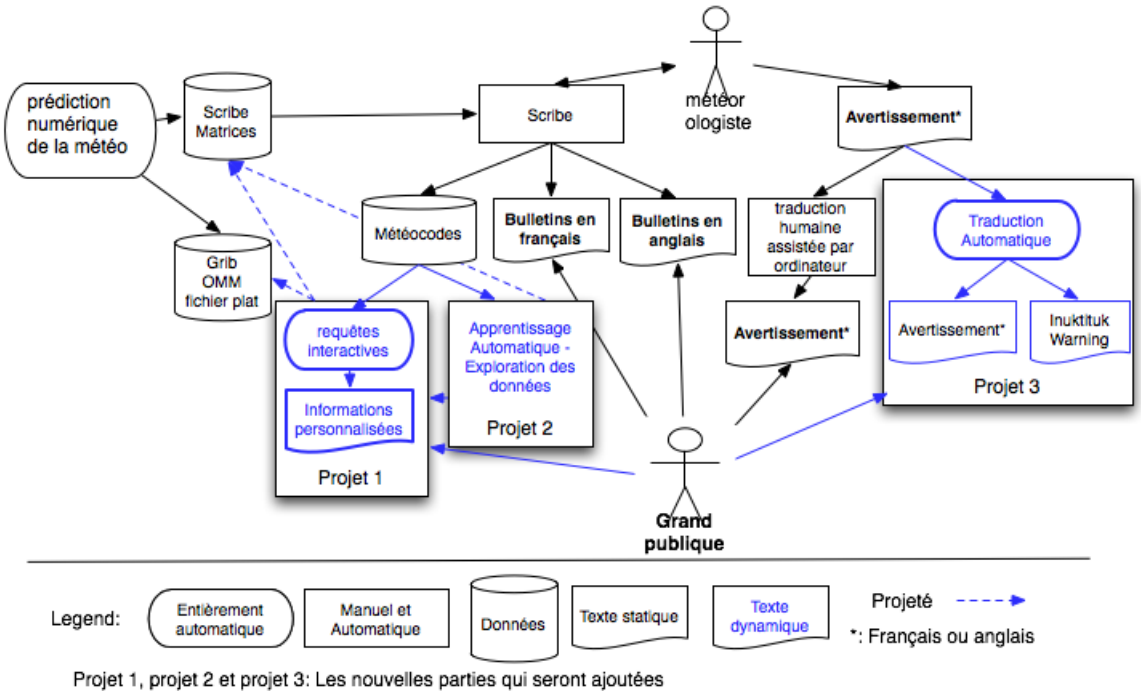
Notre projet cherche à s'intégrer au système d'information d'Environnement Canada illustré à la figure 1.5. La partie que nous traitons dans cette thèse est marquée par *projet1*.

Dans notre projet, deux types de données[4] peuvent être utilisés. Le premier type est SCRIBE[5] et contient des prévisions sous forme brute de matrices. L'information est

[4] la différence entre les données est la précision de la localisation et la qualité de la prévision

[5] Ce fichier contient des bonnes prévisions à court terme et à une résolution moyenne (stations météorologiques)

Figure 1.5 – Flux d'information simplifié à Environnement Canada



générée automatiquement par un modèle de prévision numérique du temps. Cette sortie est ensuite envoyée à un autre système qui permet aux météorologues de commenter et de modifier les prévisions. Le résultat de ce changement est le fichier appelé MeteoCode dans le format XML. Le deuxième type de données est le fichier GRIB (est un fichier plat) qui contient une prédiction brute du modèle à très haute résolution.

Afin de mieux résumer les données, nous aurons besoin d'effectuer un regroupement spatio-temporel de ces données en fonction de la similitude des conditions météorologiques, et la relation de la grappe (les algorithmes spectraux de clustering (Luxburg, 2007) sont connus pour agréger les données des matrices en fonction de leurs similitudes) et les conditions trouvées. Pour répondre à ce besoin, nous nous baserons sur les résultats du *projet2 : Apprentissage automatique - exploration des données* qui est réalisé au sein du laboratoire LISA[6]. Ce projet vise à apprendre à "interpoler" les corrections des ex-

[6]Laboratoire d'Informatique des Systèmes Adaptatifs : Le LISA travaille à comprendre les principes de l'intelligence et de l'apprentissage, afin de faire progresser les algorithmes d'apprentissage et l'intelligence artificielle.

perts, ou les observations, sur la grille haute résolution. Dans une deuxième étape, nous allons effectuer le même travail, mais en utilisant le fichier SCRIBE directement au lieu d'utiliser le fichier MeteoCode qui est le résultat des corrections appliquées par les météorologues sur le fichier SCRIBE. Le but de ce regroupement est de réduire le nombre de descriptions possibles des conditions météorologiques et de minorer l'intervention humaine. Chaque état peut être décrit comme le noyau local le plus proche. Enfin, un bon rapport devrait attirer l'attention de l'usager à des phénomènes et des conditions inhabituelles.

1.5 Conclusion

En utilisant les méthodes et les approches standard de visualisation et en tenant compte de la quantité d'information environnementale que produit EC, l'affichage de toutes ces informations est impossible. La tâche principale de notre travail est de créer des méthodes et des approches novatrices permettant la visualisation d'une grande masse d'information. Cette présentation doit être personnalisée pour chaque usager. Nous appliquerons ces méthodes dans notre domaine d'application. Nous allons créer un générateur de bulletins météorologiques qui génèrera des bulletins sur la demande des usagers. Nous essayerons de personnaliser ces bulletins pour chaque usager. Cette génération tiendra compte des meilleurs techniques de visualisation. Dans le prochain chapitre, nous présenterons des techniques, des approches et des méthodes de visualisations existantes. Nous essayerons dans ce chapitre de mettre en balance la visualisation actuelle présentée par EC et la fine pointe des techniques de visualisation.

CHAPITRE 2

ÉTAT DE L'ART : TECHNIQUES, MÉTHODES ET APPROCHES DE VISUALISATION

L'objectif de cette thèse est d'élaborer des nouvelles méthodes et approches pour résoudre le problème de visualisation afin de permettre au public canadien de consulter toute l'information environnementale dont il a besoin.

Parmi les travaux antérieurs pertinents à notre recherche, nous retrouvons des travaux théoriques sur la conception de présentations visuelles, les techniques de visualisation, l'étude de la perception, des études psychologiques et des travaux dans le domaine de la vision. Dans ce chapitre, nous passerons en revue les travaux les plus pertinents dans ces domaines et nous soulignerons leurs points importants ainsi que leur influence sur notre projet.

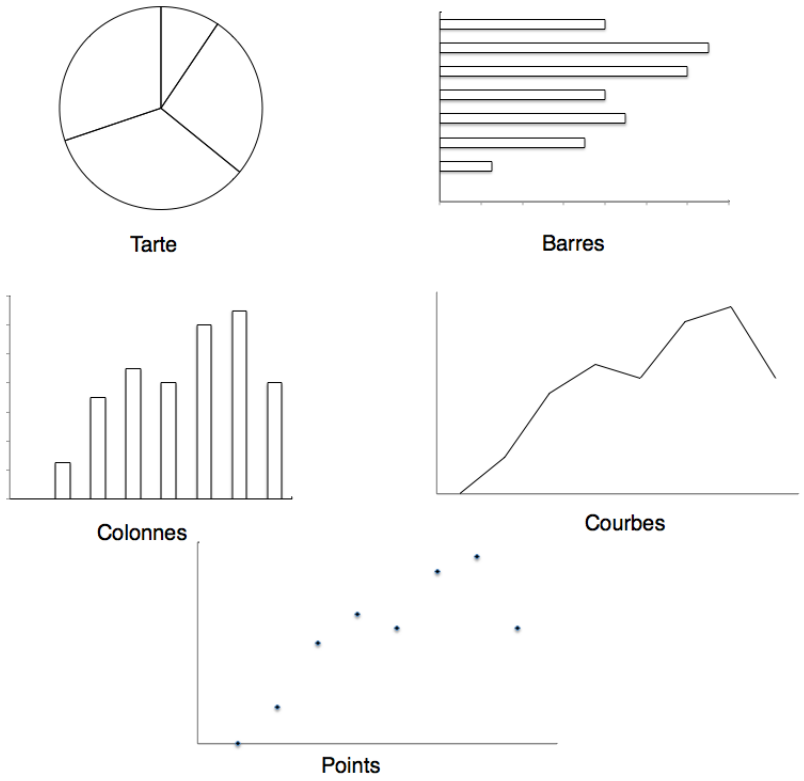
Ma thèse est dans la ligne des travaux de Fasciano (1996) qui a développé un modèle pour résoudre le problème de la génération de texte et de graphiques intégrés dans les rapports statistiques. Il a considéré plusieurs critères tels que l'intention du rédacteur, les types de variables, les relations entre ces variables et les valeurs des données. L'inspiration vient du fait que nous voulons aussi générer des bulletins contenant des graphiques et du texte. Un critère très important dans notre projet est le fait que notre visualisation doit être interactive et que l'usager ait la possibilité de paramétrer le résultat et l'adapter à ses besoins. Ce point sera étudié dans la section 2.4 qui traite de l'interactivité.

2.1 Génération de texte et de graphes

2.1.1 Génération des graphes

L'affichage des données est crucial pour leur analyse. La visualisation de ces données nous permet de mieux les explorer et de constater leurs tendances générales et leur comportement. La tâche de construction d'un graphe est l'encodage d'une quantité d'information catégorique et quantitative par le moyen de méthodes d'affichage. Cleveland

Figure 2.1 – Les 5 graphiques les plus utilisés



(1994) étudie les principes de construction des graphes. Cette étude détermine les principes qui peuvent améliorer la capacité d'un graphique pour montrer la structure des données. Ces principes sont basés sur l'étude de la perception graphique. Cette étude concerne les 5 graphes les plus connus (voir figure 2.1) qui sont décrits dans Zelazny (2001). Ces graphes sont :

La tarte

Son but principal est de montrer l'importance relative des composants l'un par rapport à l'autre et à l'ensemble. Pour faire comprendre une comparaison de décomposition, le mieux est d'utiliser une tarte. Une tarte est un diagramme circulaire constitué d'un cercle divisé en segments en forme de coin. La superficie de chaque segment (parfois appelée tranche) est le même pourcentage du cercle total.

Les barres

Le graphique barre est celui qui convient le mieux pour démontrer une position. Dans un graphique à barres, la taille de chaque barre est proportionnelle à la valeur qu'il représente. Le graphique à barres n'a pas des marques de graduation. Le but principal d'un graphique à barres est d'orienter visuellement le spectateur à la taille relative des divers éléments d'une série de données avec une échelle quantitative sur l'axe horizontal. La dimension verticale n'est pas une échelle ; elle est entièrement consacrée aux intitulés des éléments mesurés.

Les colonnes :

Les colonnes sont utilisées pour montrer l'évolution au fil du temps. Une évolution sera efficacement représentée au moyen soit d'une colonne, soit d'une courbe. La courbe est privilégie si nous avons un grand nombre d'unités de temps (les jours d'une année). Les colonnes sont préférées si notre laps de temps est divisé en quelques parties (6 ou 7) par exemple les trimestres d'une année. Contrairement aux barres, les colonnes sont dans une position verticale.

La courbe :

Ce graphique est le plus facile à élaborer, le plus souple et celui qui montre avec le plus de clarté l'augmentation, la dimunition, la fluctuation ou la stabilité d'une tendance.

Points :

Ce graphe affiche des informations quantitatives par le biais de points de données représentés par point ou d'autres symboles. Il est fréquemment utilisé pour analyser les relations entre deux ou plusieurs variables. Il est considéré comme l'un des meilleurs types de graphiques pour enquêter sur la corrélation potentielle entre deux ensembles de données.

Ces graphiques sont les plus connus et les plus utilisés. Cependant, il en existe une bonne centaine d'autres. Harris (1999) est une référence dans le domaine des graphiques.

Il peut être considéré comme étant un dictionnaire de graphes possibles. Dans ce dernier, nous distinguons quatre représentations graphiques utilisées dans le domaine de la météo.

Température : des bandes de couleur sont utilisées pour indiquer les zones de plages de température égale. Cet usage est parfois appelé une isoplèthe [1] (bandes de valeurs égales), ou une isotherme[2] (bandes d'égale température.). La gamme de valeur qui s'applique est généralement indiquée dans les bandes. Par exemple, une bande étiquetée 40 signifie que les températures pour les zones comprises dans la plage de 40 à 49 degrés

Pression : lorsque les lignes relient les points de même pression, les cartes est parfois appelé une carte isoligne ou isobare. Les valeurs sont indiquées sur les lignes et sont généralement exprimées en termes de millibars (mb), en utilisant uniquement les deux ou trois derniers chiffres. Par exemple, 20 sur la carte correspond à une valeur de 1020 mb 207 mb équivaut 1020,7 et 996 est égale à 999,6 mb. Les lettres H (High)et L (Low) désignent les points de pression le plus élevée et le plus bas dans la région.

Front[3] chaud et froid : Symboles pour transmettre des informations sur le type des fronts météorologiques, où il se trouve, et les directions de son mouvement. six des symboles les plus couramment utilisés pour les fronts météorologiques sont présentés dans la figure 2.2.

Informations météorologiques locales : certaines cartes fournissent des données très détaillées pour un emplacement spécifique en utilisant une combinaison de symboles météorologiques et des données numériques. L'exemple de la figure 2.3 montre comment les symboles sont utilisés pour désigner la direction du vent. Plusieurs autres symboles sont représentés dans la figure 2.4

[1] Une ligne reliant les points ayant les mêmes précipitations à la surface du globe.

[2] Se dit d'une ligne reliant sur une carte des points où la température est identique à un moment donné.

[3] Un front météorologique est une surface de discontinuité étendue, qui sépare deux masses d'air ayant des propriétés physiques différentes (source : Wikipédia)

Figure 2.2 – Les fronts météorologiques

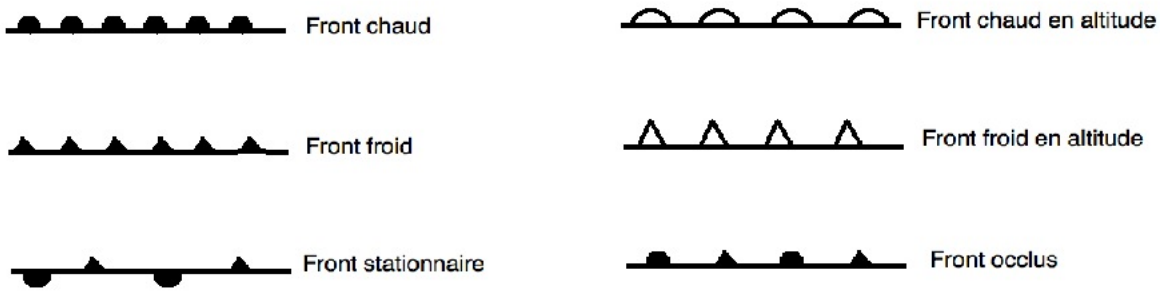


Figure 2.3 – Les symboles des directions du vent

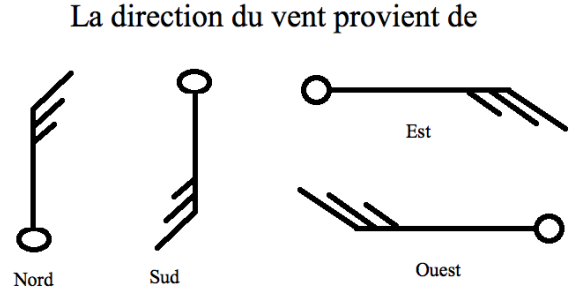
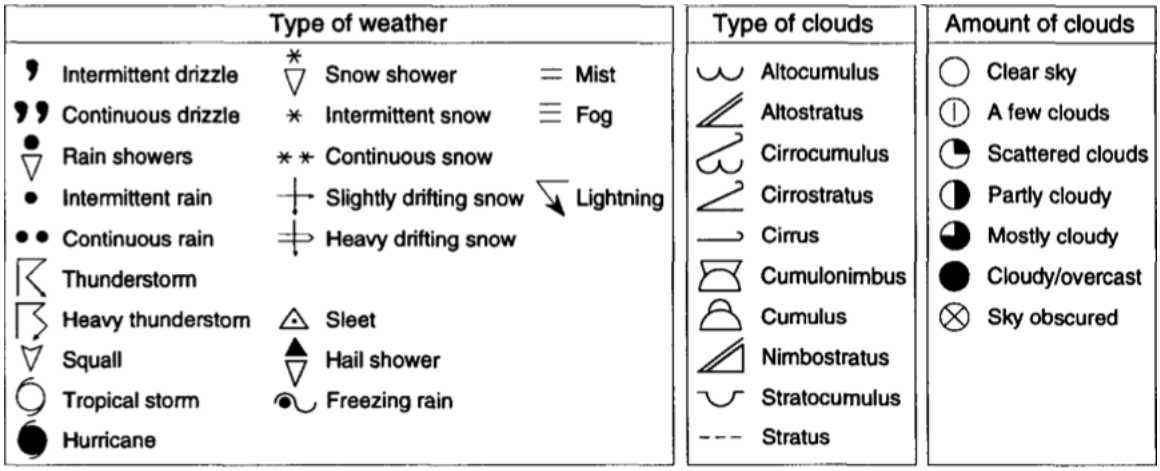


Figure 2.4 – Les symboles météo



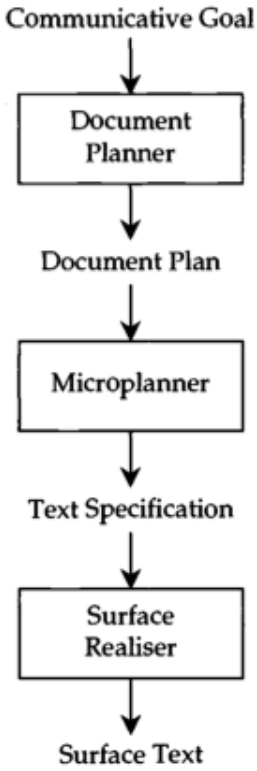
L'utilisation des graphes est sans aucune doute essentielle pour résumer et communiquer l'information. La tache de génération des graphes ne se résume aucunement en l'encodage d'une masse d'information sous forme graphique. Il faut tenir compte du décodage qui sera fait par l'utilisateur. Si l'utilisateur ne réussit pas à décoder l'information contenu dans le graphe, nous considérons que la génération du graphe à échoué. C'est pourquoi, la génération de graphiques doit tenir compte de la perception humaine. Robbins (2005) explique comment nous pouvons créer la meilleure présentation en tenant compte de tous les paramètres (choix de type de graphique, la quantité d'information, choix d'attributs de style...) et de la perception vis à vis des graphes.

2.1.2 Génération de texte

La génération du langage naturel (NLG) est un sous-champ de l'intelligence artificielle et de la linguistique computationnelle qui traite des systèmes de construction de logiciels informatiques qui peuvent produire des textes significatifs en langue naturelle à partir d'une représentation sous-jacente non linguistique de l'information. Les systèmes NLG utilisent des connaissances sur la langue et le domaine d'application pour produire automatiquement les documents, rapports, messages d'aide, et d'autres types de textes.

Dans cette thèse, pour la génération de texte, nous nous baserons sur les résultats du projet SumTime (Sripada et al., 2001) qui est un modèle générique de calcul pour la production de résumés textuels à partir d'une série chronologique de données. SumTime est constitué d'un modèle en deux étapes pour la détermination du contenu. La première étape consiste à construire une vision qualitative de l'ensemble des données, et la seconde consiste à utiliser cet aperçu, avec les données réelles, afin de produire des résumés. SumTime a été testé sur les prévisions météorologiques (Sripada et al., 2003) et le diagnostic des turbines à gaz (Yu, 2004). Cette approche est inspirée de la méthodologie décrite dans Reiter et Dale (2000) qui décrit le processus de génération du langage naturel. Cette méthodologie (voir figure 2.5) est basée sur une décomposition architecturale particulière du processus de la génération du langage naturel en trois modules : la planification du document, la microplanification, et la réalisation de la surface. La planification du document est ce qui est souvent appelé « la planification du

Figure 2.5 – Architecture d'un système de génération de langue naturel (Reiter et Dale, 2000)



texte ≫, et comprend deux sous-tâches : la détermination du contenu et la structuration du document. La microplanification comprend l'agrégation, la génération d'expression référentielle, et certains aspects de lexicalisation. La réalisation de la surface comprend la réalisation linguistique et la réalisation de structure.

En termes de représentations intermédiaires, la sortie du planificateur de document dans ce modèle est une spécification de document. Ceci est un arbre constitué d'informations portant des unités appelées messages, souvent avec des relations de discours spécifié entre les parties de l'arbre. La sortie de la microplanificateur est une spécification du texte. C'est un arbre dont les nœuds externes spécifient les caractéristiques de la phrase et dont les nœuds internes spécifient la structure logique du document en termes de paragraphes, sections, etc.

2.1.3 Génération combinée des textes et des graphiques

Un graphe simplifie beaucoup la tâche de l'usager pour analyser une grande quantité d'information. Toutefois, à lui seul, il peut être difficile de décoder tout le contenu. Le texte et les graphiques jouent des rôles complémentaires dans la transmission de l'information à l'usager. Fasciano (1996) décrit le système PostGraphe qui génère des rapports statistiques contenant du texte et des graphiques en se servant d'une description annotée des données à présenter. Les annotations utilisées correspondent aux critères établis dans le modèle théorique. Ainsi, l'utilisateur peut spécifier au système ses intensions (comparaison, évolution, réparation, corrélation...), les types des données à présenter (temporelles, numériques, ordonnées, . . .) et les relations entre les données. SelTex (Corio et Lapalme, 1999), est la suite des travaux de Fasciano (1996). C'est un système de génération qui produit des textes courts et des légendes pour accompagner les graphes qui sont générés selon les intentions de l'auteur. SelTex utilise des règles qui ont été extraites d'une étude de corpus de plus de 400 extraits de texte. Ces règles sont à la base de la génération d'un texte simple qui décrit les tendances générales des données des graphes (évolution, comparaison, corrélation...). D'une manière similaire, mais moins simple, Mittal et al. (1998) décrit un système pour produire des légendes pour les tableaux complexes. Ce système détermine le contenu et la structure du sous-titrage en analysant la structure des représentations graphiques et de la complexité de ses éléments perceptifs et en utilisant des transformations linguistiques telles que la commande, l'agrégation et le centrage.

2.2 Techniques de visualisation

La sélection et la création de la conception la plus efficace parmi toutes les alternatives pour une situation donnée exigent habituellement beaucoup de connaissances et de créativité de la part du concepteur. Alors que la compréhension des caractéristiques des données ainsi que les propriétés graphiques pertinentes sont importantes dans la construction de techniques de visualisation, étant conscient que la compréhensibilité de toute image ou graphique est essentielle pour la présentation efficace de l'information

inhérente dans les données.

2.3 Perception

La visualisation des données est efficace lorsqu'elle établit un équilibre entre la perception et la cognition pour mieux tirer parti des capacités du cerveau. La perception visuelle, qui est gérée par le cortex visuel situé à l'arrière du cerveau, est extrêmement rapide et efficace. Nous voyons tout de suite, avec peu d'effort. La perception cognitive, qui est gérée principalement par le cortex cérébral à l'avant du cerveau, est beaucoup plus lente et moins efficace. La construction du sens de données et des méthodes de présentation nécessite une réflexion consciente pour la presque totalité des travaux. La visualisation des données déplace l'équilibre vers une plus grande utilisation de la perception visuelle, en profitant de la puissance de nos yeux autant que possible.

Pineo et Ware (2011) présentent un modèle de traitement de l'information de la perception visuelle humaine. Ce modèle est en deux étapes. Dans la première, l'information est traitée en parallèle pour extraire les caractéristiques de base de l'environnement. Dans la deuxième, l'attention visuelle joue un rôle beaucoup plus actif que dans la première étape et les éléments de l'environnement ont tendance à être examinés dans l'ordre.

Étape 1 : Processus parallèle pour extraire les propriétés de bas niveau de la scène visuelle

L'information visuelle est d'abord traitée par de grands réseaux de neurones dans les yeux et dans le cortex visuel primaire à l'arrière du cerveau. Les neurones individuels sont sélectivement stimulés à certains types de renseignements, tels que l'orientation des bords ou la couleur d'une tache de lumière. Dans chaque sous-zone, de grands réseaux de neurones fonctionnent en parallèle, extrayant des caractéristiques particulières de l'environnement. Au début, ce traitement parallèle est efficace, et il est largement indépendant de ce que nous choisissons de suivre. Il est également rapide. Si nous voulons que les gens comprennent vite l'information, nous devrions la présenter de telle manière qu'elle peut être facilement détectée par ces grands systèmes de calcul rapide dans le cerveau.

Etape 2 : traitement séquentiel dirigé par les buts

À la deuxième étape, il y a une bifurcation en un sous-système spécialisé pour la reconnaissance d'objets et un sous-système spécialisé pour interagir avec l'environnement. Dans le cas de la reconnaissance d'objets, des facteurs tels que l'attention visuelle et la mémoire deviennent importants. De toute évidence, pour identifier un objet, les spectateurs doivent, en quelque sorte, faire correspondre les caractéristiques visuelles avec des propriétés de l'objet stocké dans la mémoire. En général, les tâches que l'observateur effectue influent sur ce qui est perçu. L'un des principaux mécanismes concernant ce qui est perçu à une partie précise de la visualisation est l'attention visuelle. Nous savons que certains aspects de cette deuxième transformation se produisent de manière séquentielle, un seul objet visuel est traité à la fois.

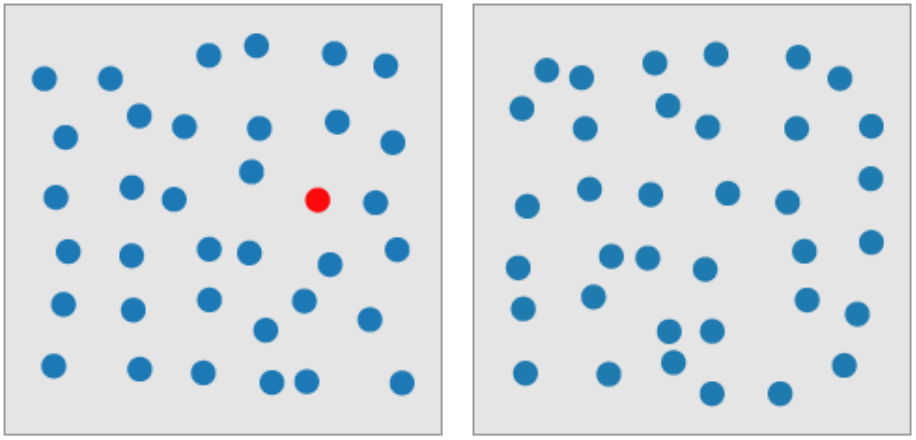
2.3.1 Traitement Préattentif

Pendant de nombreuses années, les chercheurs ont étudié comment le système visuel humain analyse les images. Un résultat important a été la découverte initiale d'un ensemble limité de propriétés visuelles qui sont détectées très rapidement et avec précision par le système de bas niveau visuel. Ces propriétés ont été initialement appelées préattentives, car leur détection semblait précéder une attention particulière. Nous savons maintenant que l'attention joue un rôle essentiel dans ce que nous voyons, même à ce stade précoce de la vision. Cependant, le terme préattentive continue d'être utilisé, car il exprime une notion intuitive de la vitesse et la facilité avec laquelle ces propriétés sont identifiées.

Un exemple simple d'une tâche préattentive est la détection d'un cercle rouge dans un groupe de cercles bleus (figure 2.6). L'objet cible a un visuel de propriété "rouge" et les objets distracteur sont bleus (tous les objets non-cibles sont considérés comme distracteurs). Un spectateur peut dire à un coup d'œil si la cible est présente ou absente.

Dans notre domaine d'application, ce concept peut nous aider à mettre en évidence les avertissements concernant les intempéries. Cela permettrait aux usagers de détecter immédiatement les mauvaises conditions météorologiques. Les avertissements dans le domaine météorologique sont d'une grande importance.

Figure 2.6 – Un exemple de recherche d'un cercle cible rouge basée sur une différence de couleur : (a) la cible est présente dans une mer de cercle distracteurs bleu (b) la cible est absent



(a) (b)

2.3.2 Les principes de Gestalt

Gestalt est un terme de psychologie qui signifie « tout unifié ». Il se réfère aux théories de la perception visuelle développée par des psychologues allemands dans les années 1920. Ces théories tentent d'expliquer comment les gens ont tendance à organiser les éléments visuels en groupes ou ensembles unifiés où certains principes sont appliqués. Les lois de Gestalt se traduisent facilement en un ensemble de principes de conception pour les visualisations d'information. Les lois de Gestalt et des principes de conception connexes sont présentés ci-dessous (Ware, 2004).

La proximité : La proximité spatiale est un principe d'organisation et de perception. Les objets proches les uns des autres sont perceptivement regroupés. La loi de proximité suppose que lorsque nous percevons une collection d'objets, nous allons voir les objets proches comme formant un groupe.

La similitude : Les formes des éléments peuvent aussi déterminer la manière dont ils sont regroupés. La loi de similitude saisit l'idée que les éléments seront regroupés perceptuellement s'ils sont semblables.

La continuité : Le principe de continuité indique que nous sommes plus enclins à construire des entités visuelles à partir d'éléments visuels qui sont lisses et continues, plutôt que ceux qui contiennent de brusques changements de direction. Il devrait être plus facile d'identifier les sources et les destinations des lignes de connexion si elles sont lisses et continus.

La symétrie : La loi de symétrie adopte l'idée que lorsque nous percevons des objets, nous avons tendance à les percevoir comme des formes symétriques établies autour de leur centre. La plupart des objets peuvent être divisés en deux moitiés plus ou moins symétriques et quand, par exemple, nous voyons deux éléments sans lien qui sont symétriques, inconsciemment nous les intégrons dans un objet cohérent (ou percept). Plus les objets sont semblables, plus ils ont tendance à être regroupés.

Le contour : Un contour fermé tend à être considéré comme un objet. Partout où un contour fermé est vu, il y a une tendance très forte de la perception à diviser les régions de l'espace en « intérieur » et « extérieur » du contour. Beaucoup de contours fermés sont utilisés pour délimiter les relations entre différents ensembles qui se chevauchent.

Dans notre analyse du site actuel d'environnement Canada et des sites de ses concurrents réalisée dans la section 1.2 du chapitre précédent, nous ne pouvons pas parler de ces principes proprement dits. La principale raison de cela est que les graphiques utilisés se limitent à des icônes génériques montrant le résumé de l'état météorologique (soleil, nuages, pluie, neige...) d'une période de temps bien défini (une journée, une période de la journée ou une heure précise). Le seul principe utilisé est le contour qui est utilisé pour limiter une zone bien définie contenant des conditions météorologiques identiques ou similaires.

2.4 Interactivité

En plus des techniques de visualisation, pour une exploration de données efficace, il est nécessaire d'utiliser une certaine interaction. Les techniques d'interaction permettent

à l'analyste de données d'interagir directement avec les visualisations et de changer dynamiquement les visualisations en fonction des objectifs d'exploration, et ils permettent aussi de relier et de combiner des visualisations indépendantes. Keim (2002) propose une classification des techniques de visualisation de l'information et de data mining basée sur le type de données à visualiser, la technique de la visualisation et la technique d'interaction et de déformation.

La motivation pour l'interaction est claire, mais nous devons tenir compte de ce qui motive l'usager à interagir. Yi et al. (2007) ont réalisé une étude pour répondre à cette question. Ils se sont basés sur différentes intentions des usagers et ont introduit une liste de catégories qui décrivent pourquoi les utilisateurs souhaiteraient interagir. Dans la suite, nous utilisons les catégories deYi et al. (2007) :

Selectionner - Marquer quelque chose d'intéressant

Lorsque l'usager repère une partie intéressante dans la représentation visuelle, il veut la marquer et la mettre en évidence en tant que telle, que ce soit temporairement pour des résultats intrigants ou de façon permanente pour mémoriser des résultats importants.

Explorer - Montrez-moi autre chose

Pour que la visualisation d'une grande masse de données complexe qui varie dans le temps soit pratiquement utilisable, l'usager doit se concentrer sur seulement une sous-plage de temps et sur une partie des variables de données. En conséquence, les usagers doivent être capables de consulter de manière interactive les différentes parties du domaine du temps et pouvoir considérer les variables alternatives pour l'inclusion dans le codage visuel pour arriver à une vue globale des données.

Reorganiser - Montrez-moi un arrangement différent

Différentes organisations possibles de temps et des données associées peuvent communiquer des aspects complètement différents, un fait qui devient évident se rappelant la distinction entre les représentations linéaires et cycliques du temps. Comme les usagers veulent regarder le temps sous des angles différents, ils doivent

être pourvus d'installations qui leur permettent de générer interactivement différentes dispositions spatio-temporelles orientées données.

Encoder - Montrez-moi une représentation différente

De même pour ce qui a été dit à propos de la disposition spatiale, le codage visuel de valeurs des données a un impact majeur sur ce qui peut être dérivé d'une représentation visuelle. Parce que les données et les tâches sont variées, les utilisateurs doivent être capables d'adapter le codage visuel pour répondre à leurs besoins, que ce soit pour effectuer des tâches de localisation ou de comparaison, ou pour confirmer une hypothèse générée à partir d'un encodage visuel en le vérifiant avec une autre alternative.

Résumer / Elaborer - Montrez-moi plus ou moins de détails

Lors de l'analyse visuelle, les usagers ont besoin de regarder certaines choses en détail, tandis que pour d'autres des représentations schématiques sont suffisantes. Les niveaux de granularité structurés hiérarchiquement, où les abstractions de haut niveau fournissent des aperçus agrégés, et les niveaux inférieurs, les détails correspondants.

Filtrer - Montrez-moi quelque chose de conditionnel

Lorsque les usagers recherchent des informations particulières dans les données ou évaluent une certaine hypothèse sur les données, il est logique de restreindre la visualisation pour n'afficher que les éléments qui respectent les conditions imposées par les critères de recherche ou les contraintes de l'hypothèse. En filtrant interactivement ou en atténuant les éléments de données non pertinentes, nous éclaircissons la visualisation pour les usagers et leur permettant de se concentrer sur leur tâche en cours.

Connecter - Montrez-moi les éléments liés

Lorsque les utilisateurs font une découverte potentiellement intéressante dans les données, ils se demandent généralement si des découvertes similaires ou connexes

peuvent être faites dans d'autres parties des données. Ainsi, les utilisateurs ont l'intention, de façon interactive, de trouver, comparer et évaluer de telles similitudes ou des relations, par exemple, pour voir si une tendance qu'ils ont découverte dans une saison d'une année est présente pour les variables d'autres données ou se répète dans le même temps dans les années subséquentes.

Annuler / Refaire - Laissez-moi aller où j'ai déjà été

Les utilisateurs ont à naviguer dans le temps et regarder à différents niveaux de granularité, ils doivent essayer différentes modalités et codages visuels, et ils ont à expérimenter avec des conditions de filtrage et de seuils de similarité. Pour tenir compte de la nature exploratoire et interactive du raisonnement analytique, un mécanisme d'historique avec les annuler et refaire des opérations sont nécessaires. Annuler / refaire permet aux utilisateurs d'essayer de nouvelles vues sur les données et de retourner sans effort à la représentation visuelle précédente si la nouvelle n'a pas fonctionné comme prévu.

Changer la configuration - Permettez-moi d'ajuster l'interface

En plus d'adapter la représentation visuelle des données et des tâches à accomplir, les utilisateurs veulent également adapter le système d'ensemble qui assure la visualisation. Cela inclut l'adaptation de l'interface utilisateur (par exemple, l'arrangement des fenêtres ou les éléments dans les barres d'outils), mais aussi la gestion générale des ressources du système (par exemple, la quantité de mémoire utilisée).

Dans leur ensemble, ces intentions constituent ce qu'un système de visualisation doit soutenir en termes d'interaction, afin de profiter pleinement de la synergie des humains et des capacités de la machine. Dans ce qui suit, nous essayons de déterminer si l'usager des sites cités dans notre analyse de la section 1.2 a la possibilité de mettre en pratique ses intentions.

Alors que le marquage (ou la sélection) des éléments intéressants et la navigation dans le temps sont quasi obligatoires, les installations pour d'autres intentions ne sont

pas souvent parvenues à un état de développement suffisant ou ne sont pas encore prises en considération. Ceci est probablement dû à l'effort supplémentaire que nous devons dépenser pour la mise en œuvre de méthodes d'interaction efficaces. Mais en fait, toutes ces intentions des usagers sont toutes aussi importantes et les techniques correspondantes devraient être fournies.

2.5 La théorie confrontée à la pratique

Dans cette section, nous analysons les visualisations des conditions météorologiques existantes des sites (annoncées dans la section 1.3) par rapport aux techniques vues dans ce chapitre pour dégager les qualités et les insuffisances.

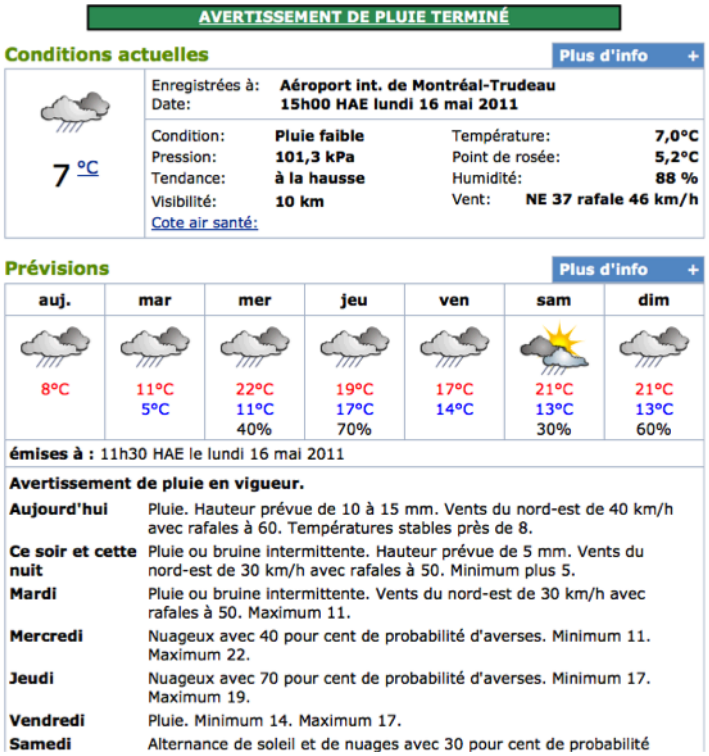
2.5.1 Environnement canada

La page d'accueil du site d'environnement Canada (figure 2.8) présente une "carte de prévision" cliquable qui résume les informations génériques sur les conditions météorologiques dans le pays. En cliquant sur la carte, l'utilisateur est redirigé vers la page de prévisions qui montre des prévisions textuelles et iconiques des prochains 7 jours pour cette région (2.7).

La description des prévisions est à un trop gros niveau de granularité pour un utilisateur ayant des plans à court terme et qui voudrait vérifier les informations météorologiques pour une heure précise de la journée. Les textes de prévisions de la semaine sont tous affichés ensemble, sans donner la possibilité à l'utilisateur de sélectionner et d'afficher uniquement les informations pertinentes. L'interface utilisateur Web n'est pas suffisamment interactive, trop d'informations non essentielles sont affichées et aucune possibilité de personnalisation n'est offerte.

Les techniques de visualisation, qui visent la perception de l'utilisateur, étudiées dans la section 2.3 sont à peine utilisées dans les présentations générées par EC. Parmi toutes ces techniques, EC en utilise quelques-unes pour aider l'utilisateur à mieux percevoir la visualisation. Le principe de contour, utilisé dans la page d'accueil (voir figure 2.8), pour délimiter les provinces et les territoires. L'interaction, à laquelle nous accordons une

Figure 2.7 – Le site officiel d'Environnement Canada : http ://www.meteo.gc.ca. Pour arriver à cette page (une ville précise), l'usager doit sélection la langue (français ou anglais) puis selectionner la ville de son choix sur une carte ou dans une liste.



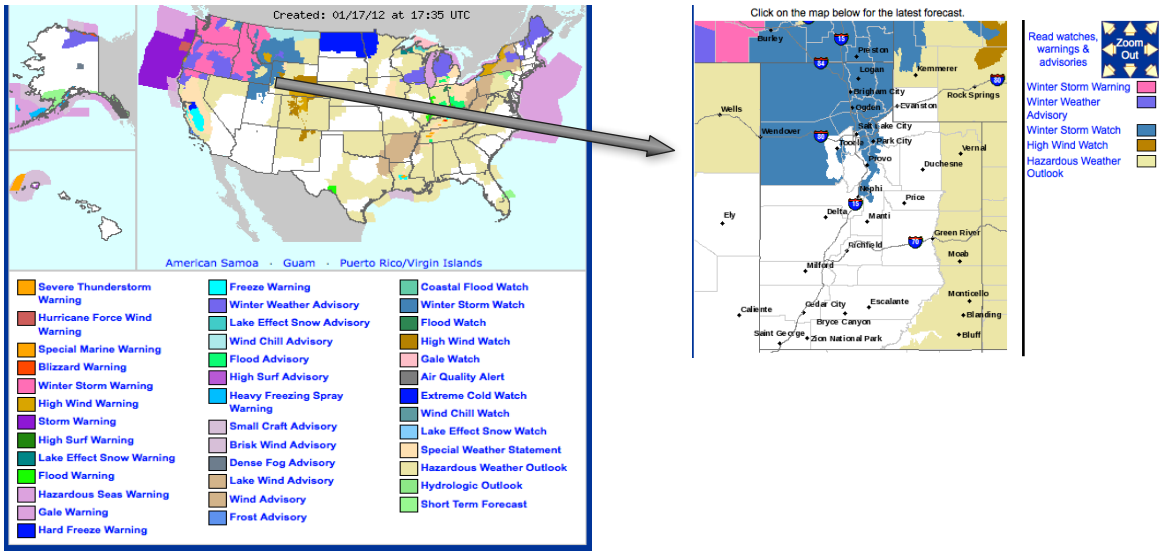
grande importance dans notre étude, est quasi absente dans les visualisations générées par EC. L'exploration des données est offerte. L'utilisateur, de la page d'accueil, peut cliquer sur une région de la carte pour explorer les informations météorologiques de cette région. Aucune autre technique d'interaction est offerte.

Figure 2.8 – La page d'accueil du site web d'EC

2.5.2 NOAA

Le site de NOAA affiche des avertissements à la page d'accueil. NOAA a mis au point des techniques graphiques permettant à l'utilisateur d'avoir facilement un aperçu de la situation nationale. Les visualisations de NOAA sont automatiquement mises à jour avec de nouveaux bulletins. Ils sont interactifs. Nous pouvons zoomer sur les États (voir figure 2.9). Un autre clic nous amènera à la page des prévisions détaillées. Les avertissements sont regroupés par catégories identifiées par des couleurs. Malgré l'exhaustivité des informations affichées, nous avons constaté que le grand nombre de catégories (39 catégories) est ergonomiquement désavantageux. Pour savoir que signifie une couleur, nous devons consulter la légende. Mais le grand nombre de catégories ne permet pas un choix diversifié des nuances des couleurs, ainsi plusieurs catégories auront des nuances de couleurs proches. L'utilisateur peut donc avoir des difficultés pour distinguer les significations de chaque couleur.

Figure 2.9 – NOAA : visualisation et zoom des avertissements



Le système de prévision détaillée NOAA (voir figure2.10) est basé sur la bibliothèque de Google Maps et génère une prévision précise pour des coordonnées spécifiques sur la carte. Un simple clic suffit pour régénérer les prévisions pour un nouvel emplacement. Les prévisions sont indiquées dans les deux modes textuels et graphiques utilisant

Figure 2.10 – NOAA : Prévisions détaillées

des icônes sémantiquement riches. En bas de page de la prévision détaillée, une carte géographique indique une zone rouge, c'est cette zone qui est concernée par la prévision. La méthode utilisée dans NOAA pour indiquer les points de prévision détaillés est l'interpolation sur une grille.

NOAA fait un usage intensif du système de prévision graphique interactive. Une grande quantité de cartes sont générées automatiquement par interpolation des valeurs environnementales comme la température, la vitesse du vent, du ciel couvert, la quantité de neige,. . . .

Un grand nombre de visualisations sont générées par NOAA. Ces graphiques couvrent la quasi-totalité des informations météorologiques collectées et archivées. NOAA utilise plusieurs principes de perception et d'interactivité. Nous pouvons constater l'utilisation du principe de contour pour délimiter les états. Dans la page d'accueil, l'utilisation du

principe de la similitude permet aux utilisateurs d'avoir une idée générale en un coup d'oeil (couleur similaire = condition similaire). L'interactivité est une technique bien présente. Elle est offerte aux utilisateurs de NOAA en leur permettant d'explorer plus de détails pour une région choisie en cliquant sur cette région. Aussi, l'usager peut choisir un arrangement des données qui lui plait en générant d'autres visualisations ou même encoder les mêmes informations d'une façon complètement différente. La technique de résumer/élaborer les données est partiellement utilisée. L'utilisateur peut consulter, dans des visualisations différentes, une visualisation pour chaque détail et une visualisation pour tous les détails. Néanmoins, nous ne pouvons pas paramétrer des détails au choix pour les afficher dans la même présentation.

Des dizaines de présentations graphiques des informations météorologiques sont générées sur le site web de NOAA. L'inconvénient majeur est que ces présentations ne sont pas accompagnées d'une description textuelle.

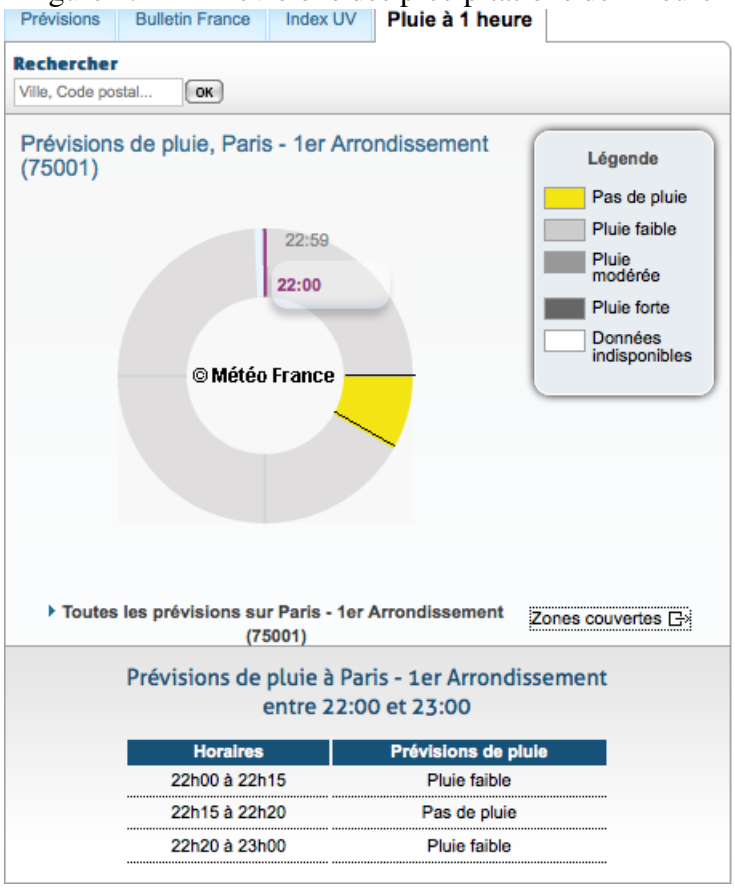
2.5.3 MétéoFrance

Meteo France distingue 4 catégories de mises en garde (rouge, orange, jaune et vert), sans préciser la nature exacte de ces avertissements qui sont signalés dans une boite dans la page d'accueil (voir figure 2.11). En suivant ce lien, nous trouvons une carte géographique de couleur signalant des niveaux d'alertes faciles à comprendre. Nous pouvons alors cliquer sur chaque région pour obtenir un bulletin d'alerte plus précis.

Figure 2.11 – Page d'accueil de Météo France

Nous avons trouvé intéressants certains systèmes d'affichage innovants déployés par les cas étudiés. Météo France a construit un outil visuel pour indiquer les prévisions de précipitations de 1 heure (voir figure 2.12).

Figure 2.12 – Prévisions des précipitations de 1 heure



La France est un pays beaucoup moins étendu que le Canada et les É-U. La masse d'information météorologique est alors moins importante. Ce qui laisse penser que les visualisations générées par météo France comportent plus de détails et de clarté. Ce qui n'est pas du tout le cas. Météo France se contente d'afficher les informations qu'elle juge essentielles sans donner aux utilisateurs la possibilité d'interagir avec la visualisation pour demander plus de détails.

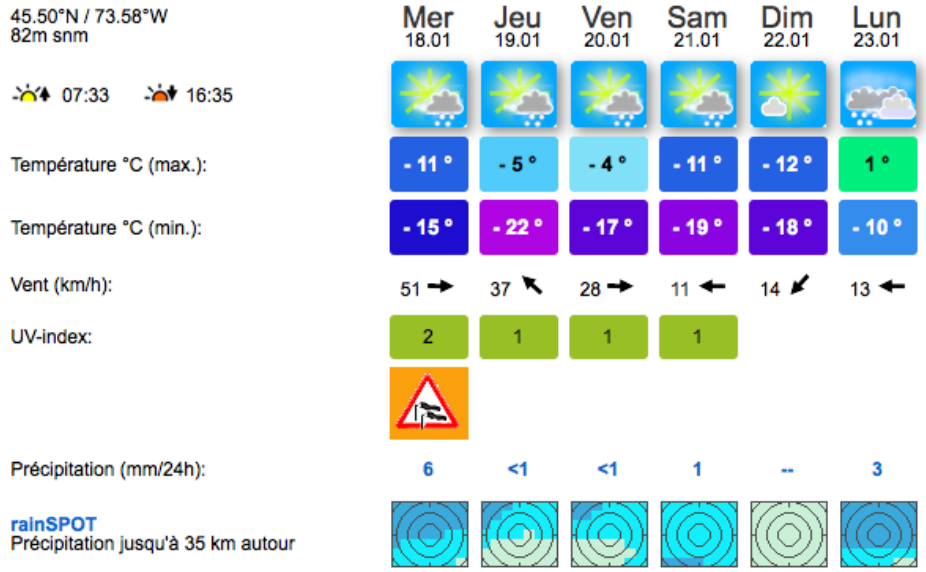
Le principe de contour est aussi présent dans météo France pour délimiter les villes. Nous pouvons explorer les détails d'une ville en positionnant le curseur sur cette région (température minimum et maximum et couverture nuageuse) ou bien, pour plus de détails, cliquer sur la région. Une très bonne description textuelle est générée dans l'onglet bulletin France. Elle décrit, textuellement, la condition météorologique globale. Météo France ne génère pas d'autre description textuelle du moins pour accompagner les graphiques.

2.5.4 Meteoblue

Meteoblue est un site privé présentant les conditions météorologiques partout dans le monde. Il détecte automatiquement l'emplacement de l'utilisateur à l'aide de son adresse IP et affiche par conséquent les conditions météo dans sa région. Meteoblue offre à l'utilisateur la possibilité de choisir un autre emplacement. Dans la page d'accueil, meteoblue génère deux présentations :

- La première (figure 2.13) d'une apparence simple utilise des caractéristiques innovantes comme : (a) les pictogrammes pour indiquer la température, la vitesse et la direction du vent et l'indice UV. (b) Le rainSPOT pour donner un aperçu des précipitations dans un rayon de 35 km par un gradient de couleur. Cette visualisation présente les informations des six prochains jours. Pour explorer des détails plus fins pour une journée précise, l'utilisateur, en cliquant sur la journée, est redirigé vers une visualisation présentant les conditions pour chaque plage de trois heures de la journée.

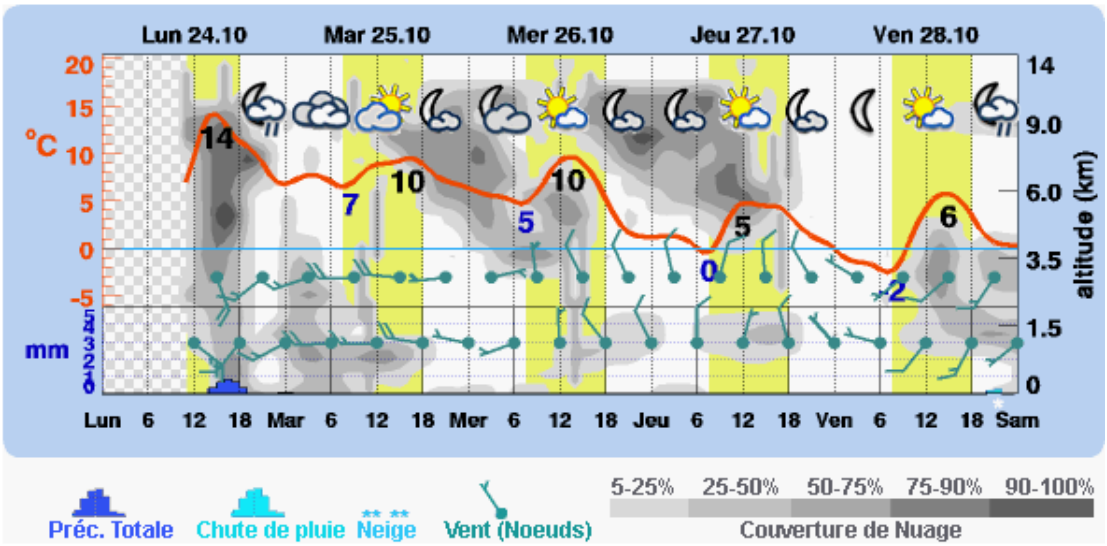
Figure 2.13 – Visualisation générée par Meteoblue utilisant des pictogrammes pour indiquer la couverture nuageuse, les températures max et min, la vitesse et la direction du vent, l'indice UV et le rainSPOT pour indiquer la précipitation dans un rayon de 35 km.



- La deuxième (figure 2.14) d'une apparence moins simple que la première, mais indiquant plus de détails d'une manière plus technique en utilisant les symboles météorologiques vus dans les figures 2.3 et 2.4.

Dans cette visualisation plusieurs techniques sont utilisées : la continuité de la ligne indiquant la température montre l'évolution de celle-ci au fil du temps. La nuance du gris utilisée montre la similitude de la couverture nuageuse. Cependant, l'interactivité n'est pas le point fort de Meteoblue. Elle n'est présente que dans la première visualisation dans le fait d'explorer plus de détails et avoir une présentation (en bloc de trois heures) d'une journée.

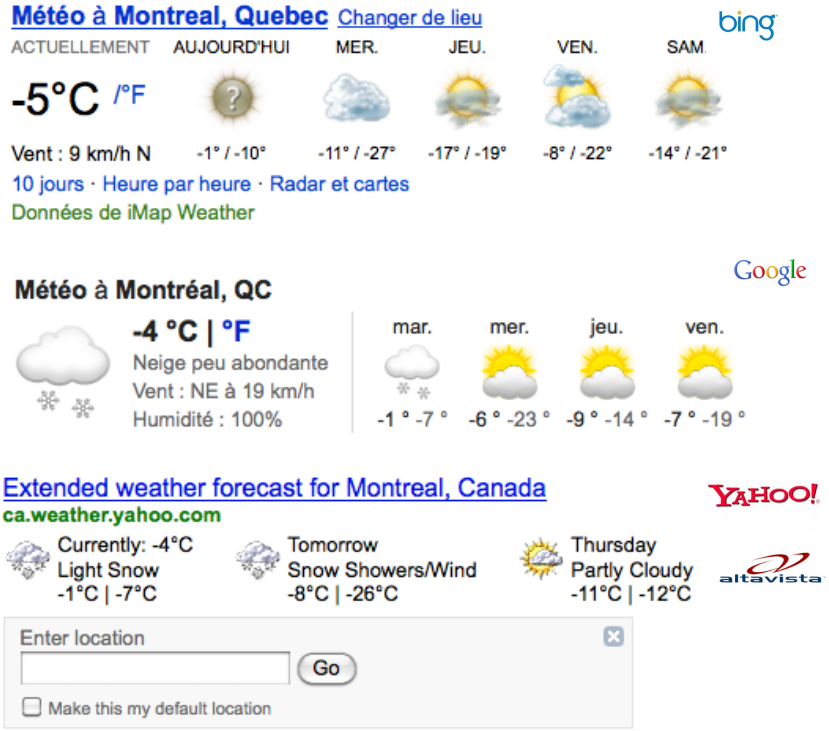
Figure 2.14 – Visualisation générée par Meteoblue indiquant la couverture de nuage (par rapport à l'altitude), la température, la précipitation, la vitesse et la direction du vent. La légende au dessous de la figure indique la signification des symboles utilisés.



2.5.5 Moteur de recherche

De nos jours, les moteurs de recherches jouent un rôle très important. Ils essayent toujours de faciliter, le plus possible, l'accès à l'information pour les utilisateurs. Une vue d'ensemble des conditions météorologiques est affichée comme premier résultat de la recherche du mot "météo" dans la majorité des moteurs de recherches. Cette vue d'ensemble contient les informations de base de la condition météorologique (température, vent, couverture nuageuse et précipitation). Ces informations sont présentées sous formes iconiques et sous forme de valeurs numériques. La figure 2.15 est un aperçu des résultats affichés par les moteurs de recherche les plus utilisés : bing, google, yahoo et altavista. Les moteurs de recherche détectent, à l'aide de la géolocalisation de l'adresse IP, l'emplacement de l'utilisateur et lui affiche directement l'aperçu des conditions météorologiques de son emplacement mais il peut aussi demander "météo Québec" pour obtenir la météo d'autres villes importantes.

Figure 2.15 – Résultats affichés par différents moteurs de recherche (bing, google, yahoo et altavista) suite à la requête "météo Montral"



Les visualisations générées par les moteurs de recherche sont affichées aux utilisateurs dans un but informatif et ne sont pas exhaustives. Une quantité très minime d'information est communiquée. Aucun texte descriptif n'accompagne ces présentations. Aucune interaction dans le but exploratif n'est offerte.

2.5.6 Récapitulation

Dans l'analyse faite dans la section précédente, nous avons étudié des sites présentant de l'information météorologique pour ensuite comparer ces présentations aux méthodes et principes de construction d'une visualisation en général. Nous sommes convaincu que l'interactivité joue un rôle important quand l'information ne doit pas être présentée de la même façon pour tous les utilisateurs. D'autre part, un minimum de principes de base de la construction des visualisations doit être pris en considération.

Dans les visualisations que nous avons analysées, aucune n'est parfaite. Loin de là, parfois nous nous apercevons que même le minimum n'est pas atteint. Selon notre analyse, nous considérons que les présentations générées par NOAA permettent le mieux aux utilisateurs la perception des informations visualisées. En ce qui concerne des principes visant la perception des utilisateurs (principes de Gesalt), NOAA emploie le principe de la similitude pour les catégories d'avertissement et le principe de contour délimiter les États. Le principe du contour et aussi utilisé, dans le même but par EC et Météo France.

Du point de vue interactivité, tous les sites excepté les moteurs de recherche offrent à leur utilisateur la possibilité de demander plus de détails. EC, Météo France et Meteoblue ne proposent pas d'autre moyen d'interaction. Les visualisations générées par NOAA peuvent être réorganisées en permettant aux utilisateurs d'arranger les données comme leur plait ou encoder les mêmes données d'une autre manière.

Concernant les moteurs de recherches, ils affichent un aperçu des conditions météorologiques juste à titre indicatif et ces aperçus ne peuvent être en aucun cas considérés comme étant une visualisation à part entière. Dans le tableau 2.I, nous récapitulons les résultats de notre analyse.

Techniques	Principes	EC	NOAA	Météo France	Météo blue	Mot. Rech.
Principes de Gesalt	Proximité					
	Similitude		√		√	
	Continuité				√	
	Symétrie					
	Contour	√	√	√		
L'interactivité	Selectionner					
	Explorer	√	√	√	√	
	Réorganiser		√			
	Encoder		√			
	Résumer/Élaborer		√			
	Filtrer					
	Connecter					
	Annuler/refaire					
	Changer la config.					
combinaison texte et graphique		√		√		

Tableau 2.I – Tableau récapitulatif des techniques de visualisation utilisés dans les sites web d'EC, NOAA, Météo France, Meteoblue et des moteurs de recherche.

2.6 Profil de l'utilisateur

Le but de notre travail est de personnaliser la présentation de l'information selon l'utilisateur. Pour pouvoir y arriver, nous avons besoin de connaitre les préférences de cet utilisateur. Ces préférences sont le résultat d'une mise en équation de son profil. Les recherches sur les profils des utilisateurs et plus précisément la personnalisation de l'information ont été abordées principalement dans le domaine de recherche d'informations. Kostadinov (2003) résume les principaux travaux qui ont été faits sur le sujet du profil de l'utilisateur.

Nous pouvons observer deux types de profils d'utilisateurs. Un profil déterminé par l'utilisateur lui-même en indiquant ses préférences et ses besoins. Et un profil détecté

automatiquement à l'aide de plusieurs techniques : historique, comportement, règles d'associations, techniques de classification et algorithmes de clustering. Mobasher et al. (2002) présente deux techniques de clustering de profil d'utilisateurs et de pages web consultées dans le but de proposer des recommandations personnalisées en temps réel.

La première technique PACT (Profile Aggregations based on Clustering Transactions) regroupe les transactions similaires de pages consultées d'un usager dans des clusters. Cette technique peut considérer un certain nombre d'autres facteurs pour déterminer le poids des éléments au sein de chaque profil. Ces facteurs additionnels peuvent inclure la distance de liaison de pages consultées à l'emplacement actuel de l'utilisateur sur le site ou le rang du profil en fonction de son importance.

La deuxième méthode de génération de profil est de calculer directement des groupes de références de pages consultées basées sur le nombre de fois qu'ils se produisent ensemble à travers des transactions d'utilisateurs (plutôt que regrouper les transactions elles-mêmes). Cette technique est appelée ARHP (Association Rule Hypergraph Partitionning).

2.7 Conclusion

Dans ce chapitre, nous avons fait le tour d'horizon de tout ce qui touche à notre sujet. Nous avons épluché les travaux dans le domaine de la visualisation, la perception et l'interactivité. Les visualisations existantes que nous avons analysées dans ce chapitre nous ont permis de dégager des règles de conception d'une visualisation dans le domaine de la météo. Dans le prochain chapitre, nous utilisons ces règles et principes pour créer une visualisation qui sera notre point de départ. Nous proposons des méthodes et des approches pour : profiler l'utilisateur, dégager ses préférences, générer une visualisation personnalisée qui combine du texte et du graphique.

CHAPITRE 3

CONTRIBUTION

3.1 Introduction

Plusieurs recherches ont été faites dans les domaines de la visualisation de la présentation personnalisée de l'information et de la génération des textes et des graphiques. Néanmoins, il reste du chemin à parcourir pour présenter, sélectivement, une grande masse d'informations en tenant compte des préférences de chaque usager. Dans ce chapitre, nous détaillons les méthodes que nous proposons pour résoudre ce problème.

3.2 Analyse et extraction des données

EC génère des bulletins météorologiques dans des fichiers XML appelés `citypage`. Ces fichiers servent pour la génération des bulletins graphiques et textuels que nous retrouvons sur le site web d'EC [1] (figure 2.7). Les informations contenues dans chacun de ces fichiers sont le résumé d'une masse plus grande d'information contenue dans deux autres fichiers appelé meteocode02 (contenant les détails des conditions actuelles et des prévisions des deux jours suivants) et meteocode37 (contenant les prévisions des quatre jours d'après). Un fichier XML crée par EC appelé `sitelist.xml` contient 832 entrées correspondant aux emplacements répertoriés de `citypage`. Chaque fichier citypage contient les prévisions d'un des 832 emplacements. La figure 3.1 est un aperçu du fichier `sitelist.xml`.

Pour plus de flexibilité et pour pouvoir donner tous les détails dont l'utilisateur pourra avoir besoin, nous allons utiliser, dans ce travail, les deux fichiers meteocode (02 et 37). Notre point de départ sera les travaux réalisés par Alessandro Sordoni[2].

Dans ses travaux, Alessandro a utilisé les deux fichiers meteocode. Pour cela, il avait besoin de créer un fichier semblable à `sitelist.xml` pour savoir quel fichier meteo-

[1] http ://www.meteo.gc.ca

[2] http ://www-etud.iro.umontreal.ca/ sordonia/deploy/prototypeV2/

Figure 3.1 – Fichier `sitelist.xml`. `<site code='s000...'>` est le nom du fichier XML correspondant à une ville. `<nameEn>` et `<nameFR>` sont respectivement le nom de la ville en anglais et en français. `<provinceCode>` est le code de la province.

```
- <siteList>
  - <site code="s0000001">
      <nameEn>Athabasca</nameEn>
      <nameFr>Athabasca</nameFr>
      <provinceCode>AB</provinceCode>
    </site>
  - <site code="s0000002">
      <nameEn>Clearwater</nameEn>
      <nameFr>Clearwater</nameFr>
      <provinceCode>BC</provinceCode>
    </site>
  - <site code="s0000003">
      <nameEn>Valemount</nameEn>
      <nameFr>Valemount</nameFr>
      <provinceCode>BC</provinceCode>
    </site>
  - <site code="s0000004">
      <nameEn>Grand Forks</nameEn>
      <nameFr>Grand Forks</nameFr>
      <provinceCode>BC</provinceCode>
    </site>
  - <site code="s0000005">
      <nameEn>McBride</nameEn>
      <nameFr>McBride</nameFr>
      <provinceCode>BC</provinceCode>
    </site>
```

code utiliser pour un emplacement précis. Il a modifié le fichier sitelist.xml pour l'adapter à ces nouveaux besoins. Ce nouveau fichier (figure 3.2) est basé sur `siteliste.xml` pour lequel il a rajouté les informations suivantes :

`<region>` : le nom en anglais et en français de la région ainsi que le code correspondant dans le fichier meteocode.

`<province>` : le nom en anglais et en français correspondant au code de la province.

`<station>` : les coordonnées géographiques (longitude et latitude) et le nom en français et en anglais de la ville.

`<meteocode02>` : le dossier et l'identifiant du fichier `meteocode02` correspondant à cette ville.

`<meteocode37>` : le dossier et l'identifiant du fichier `meteocode37` correspon-

Figure 3.2 – Nouveau fichier `sitelist.xml`. Le résultat de la modification du fichier `sitelist.xml`. Les informations `<station>` , `<meteocode02>` et `<meteocode37>` ont été ajoutées pour répondre aux nouveaux besoins.

```
- <siteList>
  - <site>
      <city en="Athabasca" fr="Athabasca" code="s0000001"/>
      <region en="Westlock - Barrhead - Athabasca" fr="Westlock - Barrhead - Athabasca" code="r16.6"/>
      <province en="Alberta" fr="Alberta" code="AB"/>
      <station lat="+54.63" lon="-113.38" en="Athabasca" fr="Athabasca"/>
      <meteocode02 dir="pnr" id="FPWG16"/>
      <meteocode37 dir="pnr" id="FPWG54"/>
    </site>
  - <site>
      <city en="Clearwater" fr="Clearwater" code="s0000002"/>
      <region en="North Thompson" fr="Thompson nord" code="r2"/>
      <province en="British Columbia" fr="Colombie-Britannique" code="BC"/>
      <station lat="0" lon="0" en="" fr=""/>
      <meteocode02 dir="pyr" id="FPVR14"/>
      <meteocode37 dir="pyr" id="FPVR54"/>
    </site>
  - <site>
      <city en="Valemount" fr="Valemount" code="s0000003"/>
      <region en="YellowHead" fr="YellowHead" code="r1"/>
      <province en="British Columbia" fr="Colombie-Britannique" code="BC"/>
      <station lat="0" lon="0" en="" fr=""/>
      <meteocode02 dir="pyr" id="FPVR16"/>
      <meteocode37 dir="pyr" id="FPVR52"/>
    </site>
  - <site>
```

dant à cette ville.

Dans les fichiers qui contiennent l'information des conditions et des prévisions, l'information météorologique est organisée par les codes des régions. Notre nouveau fichier `sitelist.xml` spécifie un code de région pour chaque ville. Dans les fichiers `meteocode02` et `37`, chaque arbre `<meteocode-forecast>` contient les informations météorologiques spécifiques à un emplacement précis (voir figure 3.3). Par conséquent, nous pouvons extraire des informations pertinentes en recherchant dans l'arbre de `<meteocode-forecast>` identifié par le code correspondant de la région de la ville.

Figure 3.3 – Aperçu du fichier météocode. Chaque emplacement à un `<msc-zone-code>` unique.

3.3 Interactivité

Généralement, l'interactivité est utilisée pour donner plus de souplesse à l'utilisateur en lui permettant de rapprocher la visualisation le plus possible de ses préférences. Dans notre cas, il y aura deux usages de l'interactivité.

3.4 Approche

Les informations sauvegardées suite à l'interactivité seront utilisées pour entrainer notre système pour qu'il puisse prévoir les préférences de nos utilisateurs et générer des visualisations personnalisées. L'approche que nous proposons consiste à créer des groupes d'utilisateurs (cluster) selon la similitude de leur profil. Chaque cluster contiendra des utilisateurs avec un profil similaire. Le nouvel utilisateur sera affecté à un de ces

clusters. Nous considérons que les utilisateurs qui sont dans le même cluster sont les utilisateurs qui sont le plus similaires. Pour cela nous nous basons sur leurs préférences pour prévoir les préférences du nouvel utilisateur. Les utilisateurs les plus proches de notre utilisateur sont les utilisateurs qui lui sont plus similaires et doivent influer plus le résultat. Pour cela nous affectons un poids $w = 1/d$ à chaque utilisateur du même cluster. d est la distance entre notre utilisateur et les autres éléments du même cluster

3.4.1 Répondre aux besoins de l'usager

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 3.4.3 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. Nous mettons à sa disponibilité tous les outils permettant au moins les 9 points détaillés dans la section 2.4.

3.4.2 Mieux connaitre l'usager

Les recherches sur le profilage automatique sont en vogue. Plusieurs grandes firmes, surtout dans le domaine du commerce électronique et de la publicité, payent cher pour connaitre à quoi pense celui qui se trouve derrière l'écran afin de lui offrir les produits qui l'intéressent le plus. Pour y arriver, ils considèrent toutes les méthodes (légales) permises y compris, et essentiellement, les témoins de connexion (cookies) et l'historique des résultats des moteurs de recherche.

Dans le domaine de la visualisation, nous nous intéressons généralement à bien présenter l'information que nous voulons transmettre à l'utilisateur plutôt que s'intéresser à présenter l'information que l'utilisateur veut avoir. Une des contributions de notre étude est le profilage automatique des utilisateurs (détaillé dans la section 3.4.3). Notre système archive l'interaction entre l'utilisateur et la présentation pour jauger les préférences des usagers.

3.4.3 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 présentation s'approche toujours le plus des préférences de chaque utilisateur, nous pouvons commencer par une agrégation des "types d'utilisateurs" et donner à chaque utilisateur une présentation qui s'approche le plus de ce qu'il veut avoir. Les informations que nous avons de notre utilisateur (sans toucher à sa vie privée) sont :

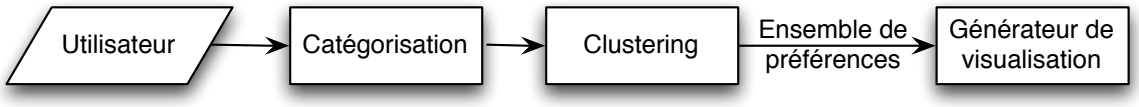
- Son emplacement (à l'aide de la géolocalisation de son adresse IP). Deux variableseront générées de cette information à savoir la longitude et la latitude.
- La langue qu'il préfère (français ou anglais) à l'aide des paramètres de son navigateur. Nous affecterons 0 pour la langue française et 1 pour l'anglais.
- L'heure actuelle (lors de sa connexion) selon son emplacement. Cette variable sera de type numérique et variera de 1 à 24.
- La saison pendant laquelle nous sommes (automne, hiver, printemps ou été). Nous échangerons cette information par un valeur qui appartient à l'ensemble $E : [1,2,3,4]$ et qui correspond respectivement à chaque saison.

Deux autres informations, que nous avons de l'utilisateur, qui ne seront pas utilisées pour le clustering, mais qui seront utilisées dans la phase de catégorisation.

- La nature de son système d'exploitation ainsi que son navigateur.
- Le type du périphérique qu'il utilise.

Ces informations ne nous permettent pas de prévoir les préférences de l'utilisateur. La première étape de la méthode (voir figure 3.4) que nous proposons consiste à archiver (anonymement) les interactions appliquées par les utilisateurs sur nos visualisations et

Figure 3.4 – Aperçu du fichier météocode



surtout les résultats finals. Nous considérons que, s'il y a interaction, le résultat final répond aux préférences de l'utilisateur. En seconde étape, nous catégorisons nos utilisateurs selon des types d'utilisateur ensuite nous regroupons les utilisateurs de chaque catégorie selon la similarité de leurs informations en nous intéressant maintenant au cluster qui contient 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 pourrons pondérer le taux de similarité en se basant sur la distance entre les vecteurs caractéristiques.

3.4.4 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 suit un moyen simple et facile de classer un ensemble de données parmi un certain nombre 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 sachant la vaste surface du Canada. La prochaine étape est d'associer chaque utilisateur archivé dans notre base de données (à 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.

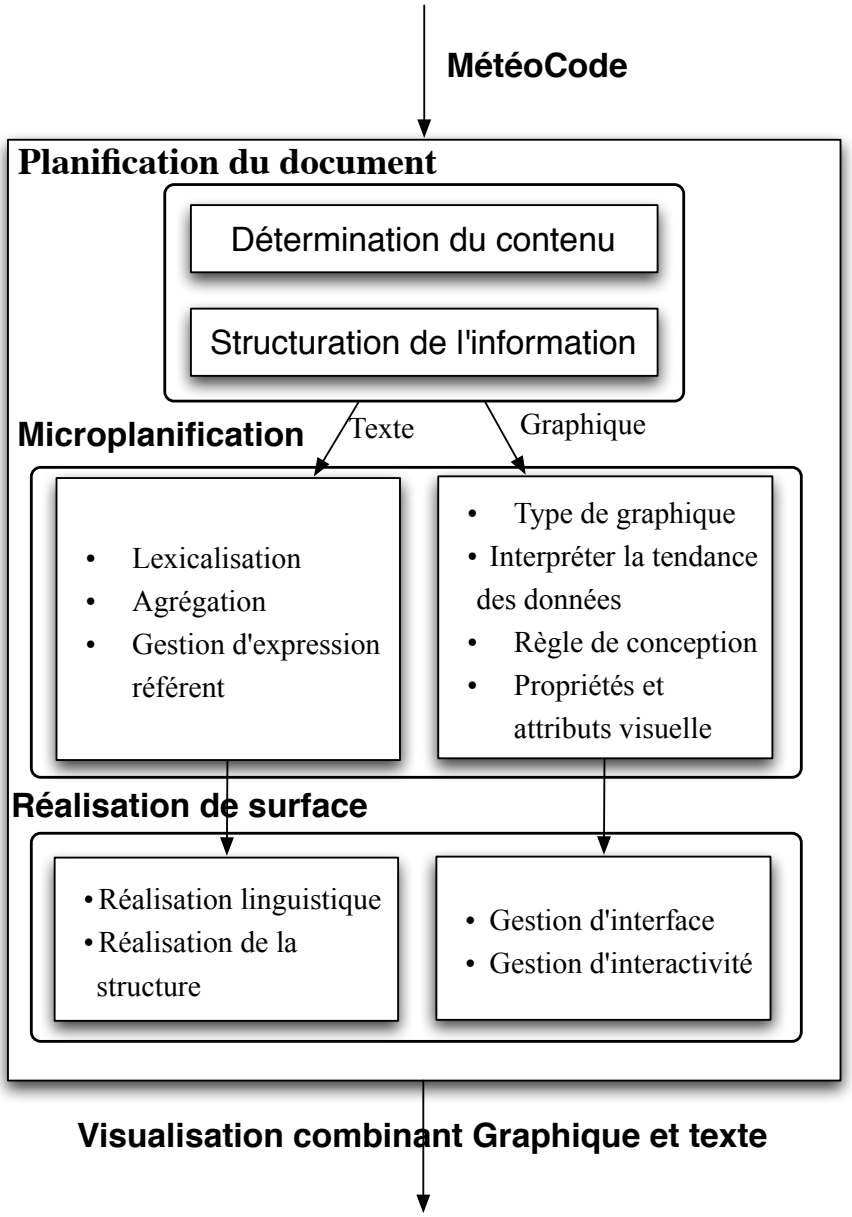
Cela produit une séparation des utilisateurs en groupes à partir de laquelle la métrique à minimiser peut être calculée.

Le problème du clustering par les K-means est NP-difficile (Dasgupta, 2008) dans le cas général. Pour cela, nous avons décidé de faire usage de cet algorithme avec un corpus de départ (de petite taille) pour fixer les classes, puis faire usage de l'apprentissage supervisé en utilisant la méthode de classement.

3.5 Génération de graphique et de texte

Le graphique et le texte générés doivent être synchronisés de telle sorte qu'ils soient complémentaires et descriptifs. Pour cela, nous proposons d'étendre la méthode de Reiter et Dale (2000) dans laquelle il décrit un processus pour la génération de texte à partir d'un ensemble de données. Nous étendons cette méthode pour l'utiliser dans la génération combinée de texte et du graphique (voir figure 3.5).

Figure 3.5 – Méthode de génération combinée de texte et de graphique



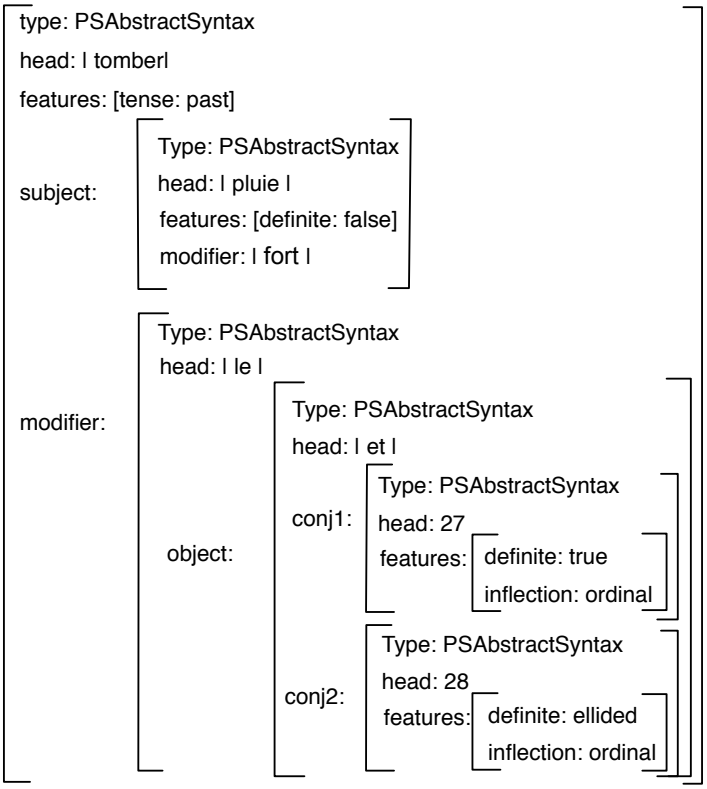
Nous ne modifions pas les étapes vues dans la section 2.1.2 à savoir : la planification du document, la microplanification et la réalisation du surface. L'étape de la planification du document traite la détermination du contenu de l'ensemble du document (graphique et texte). Ensuite, nous devons structurer cette information pour décider quelle forme aura l'information (textuelle, graphique ou bien une combinaison des deux).

Le processus à réaliser dans chaque étape concernant la génération de texte est décrit dans la méthode originale de Reiter et Dale (2000) vue dans la section 2.1.2. Nous décrivons dans la section 3.5.2 l'application des mêmes étapes pour la génération combinée des textes et des graphiques.

3.5.1 Génération de texte

Nous devons générer le texte dans les deux langues officielles du Canada. Pour cela nous utilisons la bibliothèque Java SimpleNLG-ENFR. C'est une version adaptée par Pierre-Luc Vaudry à partir de SimpleNLG v4.2 (Gatt et Reiter, 2009) qui permet de réaliser du texte en français et en anglais. La figure 3.6 est un exemple de la génération de la phrase : "Une forte pluie est tombée le 27 et le 28" qui suit la méthode décrite dans Reiter et Dale (2000) et qui est générée à l'aide de SimpleNLG-ENFR.

Figure 3.6 – Exemple d'application de la méthode de Reiter et Dale (2000). La phrase générée est "Une forte pluie est tombée le 27 et le 28"



3.5.2 Génération de graphique

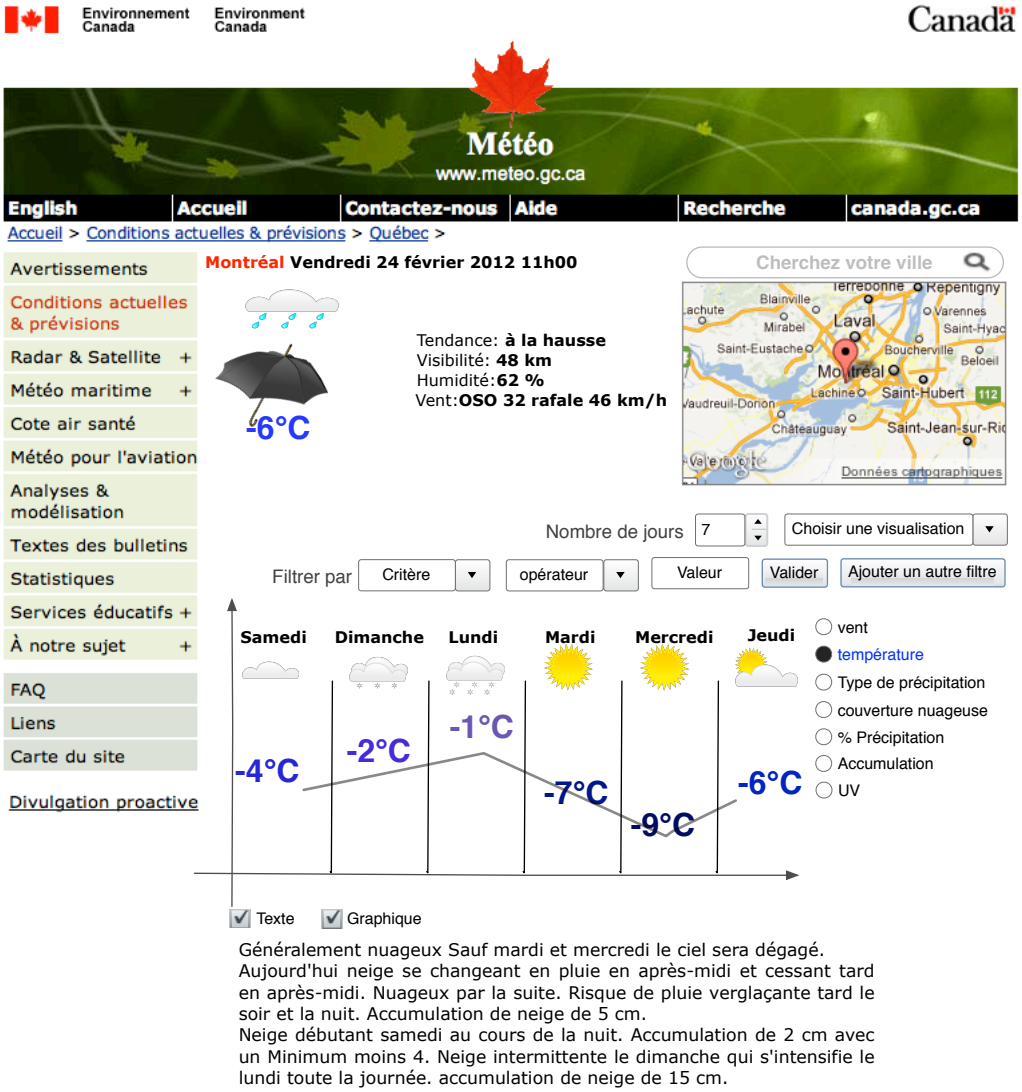
Dans l'étape de microplanification, le traitement concernant la partie graphique se fait séparément du traitement concernant le texte. Tout d'abord, nous devons décider du type du graphique à générer. Cela dépend de plusieurs paramètres : les préférences de l'utilisateur déduites suite au clustering, le choix de l'utilisateur, le type de périphérique utilisé... Ensuite, nous devons interpréter la tendance des données pour pouvoir mettre en évidence les avertissements concernant les intempéries. Cela permettrait aux usagers de détecter immédiatement les mauvaises conditions météorologiques. Nous devons aussi préciser les règles de conception et les propriétés et attributs visuels dégagés suite à la deuxième étape de l'approche de Agrawala et al. (2011).

La dernière étape est la réalisation de surface. Nous préparons une visualisation de départ avec laquelle nos utilisateurs pourront interagir. Cette visualisation répond le plus aux normes standards de création de visualisation et elle a été créée en se basant sur l'approche Agrawala et al. (2011) qui décrit en trois étapes comment crée une visualisation dans un domaine précis. La première étape est d'identifier les principes de conception en analysant des visualisations existantes. La deuxième est de créer une visualisation en se basant sur les règles et les principes dégagés dans la première étape. Enfin, dans une troisième étape, nous évaluons la visualisation que nous avons conçue. Dans la première étape, notre analyse s'est basée sur les principes de Gestalt Ware (2004) et les catégories d'interactivité décris dans **?**. La figure 3.5.2 est la visualisation conçue suite à ce processus. Cette visualisation permet à l'utilisateur d'interagir avec toutes ses composantes. Cette interaction est d'une grande importance pour la méthode que nous décrivons pour la personnalisation de la présentation.

Dans les prototypes réalisés par l'équipe du RALI, ce sont des représentations iconiques qui ont été réalisées. D'autres réalisations génèrent des graphes, mais jusque-là aucun travail n'a abouti à une génération d'un graphique avec lequel l'utilisateur peut interagir directement sur la visualisation. Dans ce travail, nous comptons utiliser la technologie SVG (Scalable Vector Graphics).

SVG

Figure 3.7 – Un exemple de visualisation généré selon les principes utilisé dans l'analyse des visualisations existantes. Cette exemple est encore une maquette qui sera connectée sous peut aux données d'EC.



La spécification SVG est un standard ouvert élaboré par le World Wide Web Consortium (W3C) depuis 1999. Les images SVG et leurs comportements sont définis dans des fichiers XML. Cela signifie qu'ils peuvent être recherchés, indexés, scénarisés et, si nécessaire, comprimés. Comme ce sont des fichiers XML, des images SVG peuvent être créées et éditées avec n'importe quel éditeur de texte.

Tous les grands navigateurs Web ont au moins un certain degré de soutien et d'interprétation du balisage SVG directement, y compris Mozilla Firefox , Internet Explorer 9 , Google Chrome , Opera et Safari. Toutefois, les versions antérieures de Microsoft Internet Explorer (IE) ne prennent pas en charge nativement SVG.

Les applications Web utilisant SVG permettent aux utilisateurs d'entrer leurs propres données, modifier des données, ou même produire de nouveaux graphiques. Puisque les données sont résidentes sur l'ordinateur client, l'interactivité est presque instantanée.

Le code source des images SVG est défini dans un fichier XML. Tous les fichiers de données générés par EC sont aussi en XML. La famille XML offre une possibilité pour transformer des documents XML en d'autres documents XML avec des éléments, des attributs, un contenu et une structure différente. La recommandation de W3C pour faire ainsi s'appelle XSLT (eXtensible Stylesheet Language Transformation). XSLT est un sous-langage de XSL [3]. C'est un avantage nous permettant de générer des visualisations directement à partir des données à l'aide d'un processus simple.

3.6 Conclusion

Le but de notre travail est de proposer des modèles ou 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 puisse analyser toutes les informations dont il a besoin. Pour cela, nous voulons personnaliser cette visualisation pour chaque usager en fonction de son profil que nous devrions détecter automatiquement. Nous voulons également utiliser 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. Pour cela, nous proposons de nouvelles méthodes combinant les textes et les graphiques dans une visualisation. Nous devons aussi tenir compte de l'évolution de l'information

[3] http ://www.w3.org/Style/XSL

au fil du temps. À la fin de ce projet, nous aurons fait des progrès dans le domaine de la présentation synthétique des informations objectives. Les nouvelles approches et méthodes utilisées dans ce travail pourraient également trouver leur application dans d'autres domaines où nous retrouvons également des modifications des informations au fil du temps en grande quantité. La présentation visuelle et la génération automatique de rapports peuvent également être appliquées dans la négociation financière, la bio-informatique, la médecine…

BIBLIOGRAPHIE

M. Agrawala, W. Li et F. Berthouzoz. Design principles for visual communication. *Communications of the ACM*, 54(4):60–69, 2011.

I. Cadez, D. Heckerman, C. Meek, P. Smyth et S. White. Visualization of navigation patterns on a web site using model-based clustering. Dans *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 280–284. ACM, 2000.

W.S. Cleveland. *The elements of graphing data*. AT&T Bell Laboratories, 1994. ISBN 9780963488411.

M. Corio et G. Lapalme. Generation of texts for information graphics. Dans *Proceedings of the 7th European Workshop on Natural Language Generation EWNLG'99*, pages 49–58. Citeseer, 1999.

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

M. Fasciano et G. Lapalme. Intentions in the coordinated generation of graphics and text from tabular data. *Knowledge and Information Systems*, 2(3):310–339, 2000. ISSN 0219-1377.

Massimo Fasciano. *Génération intégrée de textes et des graphiques statistiques*. Thèse de doctorat, Université de Montréal, 1996.

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

R.L. Harris. *Information graphics : A comprehensive illustrated reference*. 978-0-19-513532-9. Oxford University Press, USA, 1999.

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

D A Keim. Information visualization and visual data mining. *IEEE transactions on visualization and computer graphics*, 8(1):1, 2002. ISSN 1077-2626.

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

U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007. ISSN 0960-3174.

V.O. Mittal, G. Carenini, J.D. Moore et S. Roth. Describing complex charts in natural language : A caption generation system. *Computational Linguistics*, 24(3):431–467, 1998.

B. Mobasher, H. Dai, T. Luo et M. Nakagawa. Discovery and evaluation of aggregate usage profiles for web personalization. *Data Mining and Knowledge Discovery*, 6(1): 61–82, 2002.

M. Mouine. Textual and graphical presentation of environmental information. *Advances in Artificial Intelligence*, pages 319–322, 2011.

D. Pineo et C. Ware. Data visualization optimization via. computational modeling of perception. *IEEE Transactions on Visualization and Computer Graphics*, 2011.

E. Reiter et R. Dale. *Building natural language generation systems*. Cambridge Univ Pr, 2000.

N.B. Robbins. *Creating more effective graphs*. Wiley-Interscience, 2005.

S. Sripada, E. Reiter et I. Davy. Sumtime-mousam : Configurable marine weather forecast generator. *Expert Update*, 6(3):4–10, 2003.

S.G. Sripada, E. Reiter, J. Hunter et J. Yu. Sumtime : Observations from ka for weather domain. Rapport technique, Citeseer, 2001.

C. Ware. *Information visualization : perception for design*. Morgan Kaufmann, 2004. ISBN 1558608192.

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

J. Yu. *SumTime-Turbine : a knowledge-based system to generate English textual summaries of gas turbine time series data*. Thèse de doctorat, University of Aberdeen., 2004.

J. Yu, E. Reiter, J. Hunter et C. Mellish. Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering*, 13(01):25–49, 2007. ISSN 1351-3249.

G. Zelazny. *Say it with charts : the executive's guide to visual communication*. 0-7863-0894-X. McGraw-Hill Companies, 2001.