Project 1999/IPP/1006 – Data Mining et Développement d'outils spécifiques

La synthèse opérationnelle

- Nous sommes partis du fait que l'ancien système EHLASS de recueil d'information a permis de recueillir des données sur plusieurs millions d'accidents domestiques et de loisirs (ADL) en Europe (près de 5 500 000 de cas fin 1999). Ces données ont été exploitées statistiquement au niveau national. Mais il apparaît que le potentiel d'informations contenues dans les bases nationales et a fortiori dans les bases européennes récemment développées est largement sous-exploitée et qu'il importe de mieux les valoriser.

- A partir de ce constat, nous avons mené une enquête préalable auprès de la DG SANCO et des équipes nationales en charge de ce système d'information (SI). Elle a permis de connaître les outils logiciels utilisés (pour les logiciels statistiques : SAS - 5 citations, SPSS - 3 citations) et de constater que les outils spécifiques déjà développés sont rares, tandis que les besoins en outils de recueil, de contrôle et d'exploitation des données sont très nombreux et divers.

- De plus, nous exposions dans cette enquête quatre propositions de procédures pour une meilleure exploitation des données EHLASS :

> - la Procédure SSRD (Score Synthétique de Dangerosité Relative) pour hiérarchiser la dangerosité potentielle des produits;
> - la Procédure SAA (Système d'Alerte Automatisée) pour de mettre en place une alerte automatique à partir des données recueillies par le système;
> - la Procédure NGA (Note de Gravité de l'Accident) pour définir la sévérité d'un accident;
> - la Méthode SCENAR (Méthode des scénarios) pour définir des scénarios types d'accidents.

- Ces quatre procédures possèdent en moyenne, au regard des notes obtenues concernant l'utilité décisionnelle, la validité logique et l'utilisabilité, un niveau d'acceptabilité relativement bon et comparable. Cependant, la variance des notes est grande, ce qui indique que certaines équipes n'approuvent pas ces méthodes, tandis qu'un plus grand nombre ont manifesté leur intérêt pour celles-ci.

- Nous avons ensuite exposé l'apport du Data Mining qui peut être vu comme un processus d'aide à la décision où les utilisateurs cherchent eux-mêmes des modèles d'interprétation dans les données. On s'accorde à définir le Data Mining comme un ensemble de procédures de découverte de connaissances dans les bases de données de gros volume (Knowledge Discovery in Database - KDD). Ces procédures englobent des outils statistiques mais, les méthodes statistiques classiques sont plus descriptives et confirmatives, tandis que les méthodes du Data Mining sont plus exploratoires et décisionnelles.

- Nous avons ensuite montré que le SI relatif aux ADL peut être vu comme un DataWarehouse (entrepôt de données) où les méthodes du Data Mining sont applicables. Nous avons aussi souligné l'importance du cercle vertueux du Data Mining :

> 1- Identifier les données d'intervention
> 2- Utiliser les techniques du Data Mining pour transformer les données en informations utiles
> 3- Transformer les informations en actions concrètes
> 4- Evaluer les résultats

- Nous avons voulu ensuite rendre compte de l'expérience que nous avons acquise dans l'utilisation de certaines méthodes relativement sophistiquées de Data Mining, pour mettre en évidence les qualités et les défauts de ce type d'outils. Nous rapportons ainsi les expériences faites avec les méthodes des prédicteurs neuronaux et de la segmentation.

- Nous avons constaté que la mise en oeuvre de ces outils est relativement lourde (recodage des données, détermination des axes factoriels, paramétrage fin de l'outil, interprétation délicate) et nécessite une certaine expérience statistique. D'autre part, ces méthodes standards ne peuvent répondre qu'à des questions générales. C'est pourquoi, les questions spécifiques issues du SI, nous ont amené à vouloir mettre au point des procédures décisionnelles mieux adaptées.

- A chaque niveau possible d'exploitation des informations correspond une plage d'outils efficaces de caractéristiques différentes. Ainsi pour l'approche macro-accidentologique, on utilisera de préférence des outils de type indicateurs à visée épidémiologique, où la qualité statistique de l'outil est primordiale. Pour l'approche micro-accidentologique, outre les outils classiques de description statistique et de Data Mining, nous avons proposé de développer des outils spécifiques à caractère pragmatique où l'aspect validité statistique est moins essentiel. Ces outils concernent l'exploitation de données non agrégées, plus particulièrement celles issues du recueil dans les services d'urgence.

- Nous avons volontairement construit des outils qui n'utilisent que des variables endogènes au système sans procédures mathématiques complexes. Nous n'utilisons donc que les variables du système EHLASS dans des modèles additifs à partir de construction de scores et de croisements de variables. Nous avons voulu développer des outils simples d'emploi et facilement compréhensibles, fonctionnant sur les données telles qu'elles sont.

- Nous avons exposé dans le détail les quatre procédures et méthodes développées qui tiennent compte des remarques et des suggestions faites dans l'enquête et par le groupe d'experts Data Mining mis en place dans le cadre de ce projet avec nos partenaires danois du NIPH et autrichien de l'Institut Sicher Liben. Ainsi, plutôt que de développer la procédure NGA, comme exposée initialement, nous avons décidé de développer la procédure SSC (Severity Scale) portant sur le même sujet, telle qu'elle a été proposé par nos partenaires danois.

- Nous avons donc développé les procédures choisies en utilisant le logiciel de statistique le plus répandu dans les équipes : le logiciel SAS (SAS Institute). C'est ainsi que nous avons mis à disposition des équipes :

- **dmtools.c** : un écran d'interface permettant de choisir et de lancer les différentes procédures
- **saa.sas** : procédure SAA (System of Automated Alert)
- **ssrdd1.sas** : procédure SSRD utilisant la distribution des variables pour la variable « Produit impliqué dans l'accident »
- **ssrdd2.sas** : procédure SSRD utilisant la distribution des variables pour la variable « Produit ayant causé l'accident »
- **ssrdp1.sas** : procédure SSRD utilisant les percentiles pour « Produit impliqué dans l'accident »
- **ssrdp2.sas** : procédure SSRD utilisant les percentiles pour « Produit ayant causé l'accident »
- **ssc.sas** : procédure SSC (Severity SCale)

- **Intropro.doc** : présentation en anglais du projet et des procédures
- **saa.doc** : documentation en anglais relative à la procédure SAA
- **ssrd.doc** : documentation en anglais relative aux procédures SSRDxx
- **ssc.doc** : documentation en anglais relative à la procédure SSC

- **fr99v96.zip** : fichier d'essai compacté des données françaises anonymisées de l'année 1999.

- La période de test en grandeur réelle devra être sans doute plus longue que le mois initialement prévu. De l'avis de certaines équipes, elle devrait s'étendre sur un an. Ce n'est qu'à l'issue de cette période que les équipes pourront vraiment juger de la pertinence et de l'utilisabilité des méthodes proposées.

- Quoiqu'il en soit, et quel que soit le degré d'intérêt des procédures développées, nous montrons plus largement dans ce rapport qu'il importe :

- de mieux valoriser les données existantes en renforçant la coopération transnationale;

- de partager les expériences dans l'utilisation d'outils et de logiciels;
- d'utiliser les méthodes éprouvées de Data Mining dans une perspective décisionnelle;
- de développer, par ailleurs, d'autres approches et d'autres outils plus spécifiques;
- de tester les outils simples proposés en tenant compte du contexte de leur développement.

- Il est clair, toutefois, qu'aucune technique d'analyse de données ou de Data Mining ne remplacera l'expertise humaine. Mais, l'expertise humaine peut être enrichie par la confrontation aux données réelles, par l'utilisation de logiciels et de méthodes standards de Data Mining et par les résultats issus d'outils nouveaux qui viendront à leur tour guider les bons choix méthodologiques et techniques. Il y a donc fertilisation commune entre l'expertise humaine du domaine et la maîtrise d'un ensemble d'outils d'analyse.