

Il est pour le moins inhabituel de faire suivre un article publié dans une revue d'une discussion des problèmes posés par certains des aspects de cet article. Il nous a paru intéressant de le faire dans le cas de l'article d'Anne Laurent. En effet, l'article présente manifestement une contribution intéressante sur une problématique en plein développement. C'est ce que les trois relecteurs, qui dans la tradition de la revue I3 reflètent des sensibilités et des intérêts différents, ont considéré au vu de la version révisée de l'article. L'un d'entre eux, cependant plus réservé sur un des aspects de l'article, a accepté, à notre initiative et en accord avec l'auteur, d'écrire quelques pages offrant un complément de discussion prolongeant l'article sur certains aspects. Nous espérons que cette initiative, que nous croyons propre à nourrir un débat scientifique de qualité, sera bien accueillie par nos lecteurs.

Le Comité Editorial

**Quelques commentaires sur l'article d'A. Laurent intitulé
"FUB et FUB Miner : deux systèmes pour la représentation,
la manipulation et la fouille de données multidimensionnelles
floues"**

Patrick BOSC

Cet article se situe au confluent de trois domaines : les bases de données, les ensembles flous et l'apprentissage. La ligne directrice du travail présenté est de stocker des données (éventuellement imprécises ou incertaines) dans un système de bases de données multidimensionnelles qui permettra par des requêtes flexibles de fournir des données permettant de construire des arbres de décision flous.

À ce titre, l'article illustre une problématique de recherche novatrice et susceptible d'avoir des développements importants à l'avenir. Son contenu met surtout l'accent sur le cadre informatique général, sans s'intéresser particulièrement à certains aspects touchant plus à la représentation des

informations. C'est ce dernier point que je souhaite éclairer ici de quelques remarques.

L'utilisation des ensembles flous dans les bases de données a fait l'objet de travaux assez nombreux depuis une quinzaine d'années dans le cadre de ce qu'il est convenu d'appeler "les bases de données floues". Ce domaine recouvre essentiellement deux problématiques : l'interrogation de données imparfaites et l'expression de requêtes flexibles. La première vise la prise en compte d'attributs à valeur mal connue, i.e., dont la valeur pour un attribut mono-valué ne se ramène pas à un singleton. Un tel attribut est alors représenté par un ensemble pondéré (ensemble flou) de candidats appelé distribution de possibilité, qui est au cadre possibiliste ce qu'une distribution de probabilité est au cadre probabiliste. Il est possible d'établir un lien entre ce type de donnée et des données usuelles en considérant qu'une distribution décrit des états (ou mondes) plus ou moins préférés associés à chacun des candidats. La seconde, quant à elle, s'intéresse à des requêtes dont les critères ne sont pas de nature booléenne, mais au contraire graduelle. Les conditions intervenant dans une requête font alors appel à la notion de préférence et leur résultat rend compte du niveau de satisfaction (préférence) atteint par une donnée dont la valeur est précise. De par leur sémantique, les prédicats flous, dérivés des ensembles flous, offrent un cadre général puissant pour formuler de telles requêtes.

Bien qu'a priori orthogonales, ces deux problématiques peuvent être associées. En effet, la notion d'ensemble flou, commune aux distributions de possibilité et aux requêtes flexibles permet d'envisager l'interrogation flexible de données imprécises. Dans les années 80, H. Prade et C. Testemale ont proposé une approche permettant d'effectuer une sélection flexible sur une relation dont certaines valeurs d'attribut sont imprécises. De ce point de vue, l'article d'A. Laurent innove sur deux points : l'introduction d'un degré de fiabilité associé aux données et l'utilisation d'un modèle de données multidimensionnel.

Dès lors qu'une donnée n'est pas précise, son interrogation conduit la plupart du temps à un résultat incertain, associé à la réalisation de l'événement considéré. Par exemple, si on sait que la voiture de Jean a entre 7 et 10 ans, il est certain qu'elle a moins de 12 ans ou plus de 6 ans, mais il est possible qu'elle ait entre 8 et 9 ans sans que cela soit certain. L'utilisation de

distributions de possibilité, c'est-à-dire l'affectation d'un niveau de préférence aux valeurs admissibles, ne change pas cet état de fait. Elle conduit à raffiner le caractère possible et/ou certain de la satisfaction (de type tout ou rien) en un couple degré de possibilité Π et de certitude N . En présence d'une condition booléenne, le degré de possibilité Π correspond au degré du monde le plus possible dans lequel cette condition est vérifiée. Le degré de certitude correspond au complément (à 1) du degré du monde le plus possible dans lequel la condition n'est pas vérifiée. Le couple (Π, N) permet d'ordonner de façon totale les événements considérés (par exemple des n-uplets de relation) depuis la certitude de fausseté ($\Pi = N = 0$) jusqu'à la certitude de vérité ($\Pi = N = 1$) en passant par l'indécision ($\Pi = 1, N = 0$), traduisant que l'élément peut être complètement satisfaisant tout comme il peut ne pas l'être du tout. Ce mécanisme s'étend au cas où des conditions flexibles sont en jeu, en perdant certaines propriétés du couple (Π, N) puisqu'on n'a plus d'ordre total par exemple. On dispose donc d'un mécanisme d'appariement flou fondé sur deux mesures (une seule suffit dans le cadre probabiliste compte tenu de la complémentarité entre la probabilité d'un événement et celle de l'événement contraire) rendant compte de l'incertitude relative à la satisfaction de la condition considérée.

Dans son article, A. Laurent choisit une mesure dite de satisfiabilité (voir 3.3.1) pour effectuer l'appariement entre une donnée imprécise et une condition graduelle. Un de ses représentants est précisé en sous section 4.2 et n'est autre qu'un degré d'inclusion du sous-ensemble flou représentant la donnée dans celui représentant la condition. La seule justification donnée par l'auteur concerne le fait que lorsque la donnée est précise, on retrouve par cette formule le degré de satisfaction obtenu pour cette donnée par la fonction d'appartenance. Si cette observation est juste, elle peut apparaître néanmoins insuffisante puisque rien n'est dit sur la capacité de cette mesure à rendre compte de l'incertitude relative à la satisfaction de la condition par la donnée concernée. Il semble en particulier difficile d'établir un lien clair entre la valeur délivrée et la satisfaction de la condition dans les états quelque peu possibles. Soit la donnée $D = \{0.2/x_1 + \dots + 0.2/x_{20} + 1/x_{21}\}$ et la condition $C = \{0.3/x_1 + \dots + 0.3/x_{20} + 0/x_{21} + 1/x_{22}\}$, le degré d'inclusion de D dans C vaut $(20 * 0.2) / (20 * 0.2 + 1) = 0.8$. Cependant, dans le monde totalement possible x_{21} de D la satisfaction est nulle pour C alors que dans tout autre état possible au degré 0.2, le degré de satisfaction est 0.3. Cet exemple illustre une situation où le niveau de satisfaction de la condition est

quoiqu'il arrive faible alors que le degré rendu est lui assez élevé. Il semble donc peu adéquat de s'en remettre seulement à un tel indicateur pour rendre compte de la notion d'incertitude.

Un second problème sémantique survient au niveau du calcul du degré associé à une cellule au moyen de la formule donnée en sous section 3.3.1. En effet, le calcul proposé agrège par une norme trois grandeurs : le résultat de l'appariement entre valeur et condition (dont il a été question dans le paragraphe précédent) dénoté $\alpha(\mu_o, v(\vec{x}))$, le degré de fiabilité associé à la cellule ($d(\vec{x})$) et le degré précédemment associé à la cellule x , qui, d'après ce qui est dit en sous section 3.2.3, peut avoir deux significations bien différentes : un degré de fiabilité ou le résultat d'une opération précédente sur la base de données. Si ces trois valeurs sont définies sur l'intervalle unité ($[0, 1]$), cela ne signifie pas que leur combinaison (par exemple au moyen de l'opération minimum), ait un sens précis, quand bien même le résultat sera garanti (au plan syntaxique) appartenir à l'intervalle unité, et donc se prêter à des opérations ultérieures. Tout comme en programmation, la notion de type abstrait a été introduite afin (entre autres) d'associer des opérations où les paramètres ont des signatures en terme de type (indépendamment de toute mise en œuvre/codage), il faut définir le sens véhiculé par le mécanisme proposé. Il ne me semble pas que combiner des degrés de satisfaction et des degrés de fiabilité ait une signification très évidente. Quand bien même le terme relatif à l'appariement aurait une sémantique d'incertitude, cela aurait-il du sens de le combiner avec un degré de fiabilité ?

Pour conclure, je voudrais juste attirer l'attention sur l'importance que revêtent les aspects liés à la sémantique véhiculée par les calculs qu'on est amené à effectuer. La question me semble se poser dans cet article à deux reprises. La prise en compte simultanée de données imprécises et de conditions flexibles pose un certain nombre de questions. Celles-ci ne peuvent vraisemblablement trouver des réponses satisfaisantes qu'en se référant à un cadre théorique de l'incertain bien identifié et en adoptant une démarche sémantiquement solide.