Gilbert Babin and Erik Audet. "Sélection des indicateurs et des dimensions pour le forage d'information." Expertise informatique. 3(1). 1997. pp. 26-30.

Sélection des indicateurs et des dimensions pour le forage d'information

Gilbert Babin[†], ift.a. et Erik Audet[‡]

Article soumis au journal *Expertise informatique*.

[†]Département d'informatique, Université Laval
Ste-Foy, Québec, Canada
babin@ift.ulaval.ca

[‡]OLAP Products Division, Oracle Corporation
Boston, MA, Etats-Unis
eaudet@us.oracle.com

Février 1997

Gilbert Babin and Erik Audet. "Sélection des indicateurs et des dimensions pour le forage d'information." Expertise informatique. 3(1). 1997. pp. 26-30.

Résumé

Les indicateurs sont des valeurs numériques permettant aux entreprises d'effectuer le contrôle des résultats en fonction des objectifs. Un exemple d'indicateur serait le total des ventes pour une période donnée. Habituellement, les indicateurs sont obtenus en agrégeant un très grand ensemble de données. Afin d'obtenir plus d'information que la simple valeur numérique finale, ces données sont organisées sous une forme matricielle, où les cellules de la matrice correspondent aux données brutes et les dimensions représentent les façons de composer ces données pour obtenir la valeur de l'indicateur. Afin de réduire le travail associé à la configuration de telles matrices, nous avons développé une métrique qui permet de classer les différents attributs d'une base de données comme bons indicateurs (données pouvant servir au calcul d'un bon indicateur) ou bonnes dimensions (données servant de valeur dans les dimensions de la matrice). Cet article présente la métrique et l'illustre à l'aide d'une base de données utilisée dans une entreprise de distribution.

1- Introduction

Dans une entreprise, l'atteinte des objectifs est vérifiée en comparant les résultats attendus aux résultats obtenus. Il arrive fréquemment qu'on formule les résultats en fonction des ventes effectuées, des dépenses encourues, ou de toute autre valeur numérique servant de base de comparaison entre les objectifs et les résultats. Un *indicateur* est une valeur agrégée représentant une situation donne et permettant de comparer cette situation avec une autre situation. La valeur de l'indicateur n'est pas significative en elle-même; elle doit être comparée avec une autre valeur du même indicateur à un autre moment pour déterminer l'évolution d'une situation. Ainsi, la somme de toutes les ventes pour toutes les divisions d'une entreprise peut constituer un indicateur, de même que la somme de toutes les dépenses pour ces mêmes divisions.

Le forage d'information permet d'obtenir les valeurs intermédiaires qui ont servi au calcul d'un indicateur. Par exemple, on pourrait «forer» l'indicateur vente pour obtenir le total des ventes pour chaque division. Dans ce contexte, même si les ventes totales sont raisonnables selon les objectifs établis, on pourrait découvrir qu'une division est nettement en deça de ses objectifs et qu'une autre est nettement au delà des siens.

Un système de forage est un système d'information qui permet la définition d'indicateurs et offre les outils pour les manipuler (calcul, forage, etc.). Bon nombre de systèmes de forage d'information utilisent des structures de données matricielles dans lesquelles sont emmagasinées les données servant au calcul d'un indicateur. Les *dimensions* de la matrice correspondent aux différentes façons de forer l'indicateur, tandis que les valeurs contenus dans les différentes cellules de la matrice correspondent aux données brutes servant au calcul de l'indicateur.

Ces structures matricielles ne sont cependant pas appropriées pour les systèmes transactionnels de l'entreprise, qui utilisent habituellement des bases de données relationnelles. Des efforts importants doivent donc être consacrés pour arrimer le système de forage avec les systèmes transactionnels. On devra notamment:

1. définir les indicateurs sur lesquels portera le forage,
2. identifier les dimensions dans lesquelles s'effectuera le forage,
3. spécifier comment rafraîchir le contenu du système de forage à partir des systèmes transactionnels.

Dans cet article, nous présentons une métrique permettant de classer les colonnes (attributs) d'une base de données selon qu'ils constitueront de bons indicateurs ou de bonnes dimensions. Cette méthode est systématique: on classe d'abord les tables de la base de données, puis les colonnes contenus dans ces tables. La prochaine section présente globalement la méthode de calcul de la

métrique. Le classement des tables est présenté à la section 3, tandis que le classement des colonnes fait l'objet de la section 4. Nous concluons l'article à la section 5.

2- Métrique de classification des attributs

La métrique que nous avons développée permet de classer toutes les colonnes d'une base de données relationnelle. Pour ce faire, nous représentons le schéma de la base de données (ou un sous-ensemble) par un graphe, où les nœuds correspondent aux relations et les arcs représentent les liens fonctionnels. Un lien fonctionnel existe de la relation A vers la relation B lorsque la clé de la relation A est une colonne de la relation B (clé étrangère ou partie de la clé primaire). Ces arcs représentent en fait la direction de plus grande multiplicité et vont dans le sens opposé des dépendances fonctionnelles; en effet, pour une valeur de B, on aura au plus une valeur de A correspondante, alors que pour une valeur de A, on peut avoir plusieurs valeurs de B.

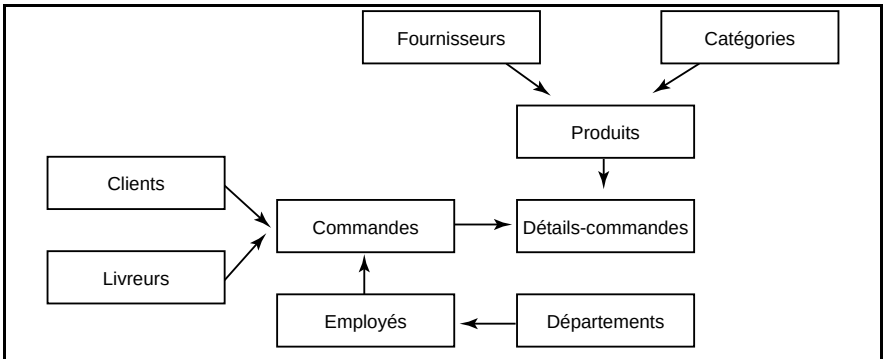

Figure 1- Graphe de description d'une base de données

Ce type de graphe est illustré à la figure 1; on y retrouve un graphe décrivant une base de données dans le domaine de la distribution. Dans cet exemple, la table *Détails_commandes* contient l'information sur un produit dans une commande; les tuples de cette table correspondent aux lignes des commandes. Une commande (tuple dans la table *Commandes*) peut donc être en relation avec plusieurs tuples de *Details-commandes*. De plus, un client peut passer plusieurs commandes, qui peuvent contenir plusieurs lignes.

2.1- Caractéristiques discriminantes

Le classement des colonnes s'effectue à partir des caractéristiques d'un bon indicateur et d'une bonne dimension. Une fois ces caractéristiques identifiées, on doit déterminer comment les utiliser pour classer les attributs et comment combiner les classements obtenus pour chaque caractéristique. Intuitivement, un indicateur est associé à une colonne dite «agrégée»; plus cette colonne contient de données, plus elle est susceptible d'être un bon indicateur. De plus, les données qu'elle contient varient constamment et ont un caractère dynamique[1], ce qui rend leur surveillance importance.

À la figure 1, la table *Détails_commandes* contient des colonnes comme le montant de la vente pour une ligne et la quantité commandée. Ces colonnes représentent de très bons indicateurs, car elles sont très dynamiques. Par ailleurs, la table *Details-commandes* a la plus grande multiplicité[2] de la base de données, ce qui rend ces colonnes encore plus attrayantes comme indicateur.

Qu'en est-il des colonnes représentants les dimensions? Ces colonnes doivent avoir un caractère statique. Les mises à jour y sont donc moins importantes et le nombre de valeurs distinctes varie

---

[1] Les données ont un caractère dynamique lorsqu'elles sont souvent mises à jour. À l'opposé, nous avons les données à caractère statique qui ne sont pas souvent mises à jour.

[2] La multiplicité est la mesure du notre de tuples dans une table.

très peu. La plus part du temps, ces colonnes sont utilisées pour ventiler des données à caractère dynamique. Par exemple, la table *Départements* contient des colonnes comme le nom du département. Il est peu probable que les données contenues dans cette table varient souvent. De plus, on aura tendance à vouloir connaître les ventes par département.

En récapitulant, on peut dire qu'une colonne indicateur[3] est caractérisée par:

- une grande multiplicité,
- un grand niveau d'agrégation,
- un caractère dynamique,
- des numériques;

alors qu'une colonne dimension est caractérisée par:

- une faible multiplicité,
- un faible niveau d'agrégation,
- un caractère statique,
- des valeurs non-numériques.

2.2- Méthode de classement

Le classement d'effectue en quatre temps. On détermine d'abord le *classement local d'une table* (notée α), qui indique si une table a tendance à être indicateur ou dimension en ne tenant compte que des arcs entrant et sortant de la table. On peut alors déterminer le *classement global de la table* (noté γ), en utilisant les différents classements locaux α. Par la suite, on détermine le *classement local des colonnes* (noté β), en tenant compte des caractéristiques de la colonne indépendamment de la table où elle se trouve. Finalement, le *classement global des colonnes* (noté λ) est calculé à partir de β et γ. La section 3 présente la méthode de calcul de α et γ, alors que la section 4 donne les critères pour le calcul de β et λ.

3- Classement des tables

Comme une colonne indicateur est caractérisée par une grande multiplicité et un grand niveau d'agrégation, on voudra classer les tables en tenant compte des relations de multiplicité qui entrent et sortent de la table. Dans le graphe de description de la base de données (cf. fig. 1), on regarde donc les liens entre les tables et la direction des ces liens. Ainsi, pour une table t_i, on notera par $L^e_{t_i}$et $L^s_{t_i}$ le nombre de liens entrant et sortant de la table t_i, respectivement. Le classement local α_{t_i} est calculé par la formule

$$\alpha_{t_i} = \frac{L^e_{t_i} - L^s_{t_i}}{L^e_{t_i} + L^s_{t_i}}$$

Le classement global d'une table t_i est calculé à partir du classement local des tables qui lui sont adjacentes. On note par A_{t_i} l'ensemble de ces tables adjacentes. L'idée du classement global est de déterminer si une table est un meilleur indicateur que ses tables voisines. Le calcul est donc une moyenne pondérée du classement local de la table et de celles qui lui sont adjacentes. On calcule γ_{t_i} comme suit

[3] Dans ce qui suit, indicateur et dimension seront utilisés comme adjectifs.

$$\gamma_{t_t} = \frac{\alpha_{t_i} + \sum_{t_j \in A_{t_i}} \left(\alpha_{t_i} + \alpha_{t_j}\right)}{2 \cdot \left\|A_{t_i}\right\| + 1} = \frac{\alpha_{t_i} \cdot (\left\|A_{t_i}\right\| + 1) + \sum_{t_j \in A_{t_i}} \alpha_{t_j}}{2 \cdot \left\|A_{t_i}\right\| + 1}$$

Les valeurs de α_{t_i} et γ_{t_i} pour notre exemple sont données au tableau 1.

Table	α	γ
Départements	-1,00	-0,667
Livreurs	-1,00	-0,500
Employés	0,00	-0,100
Produits	0,33	0,048
Commandes	0,50	0,167
Fournisseurs	-1,00	-0,556
Détails_commandes	1,00	0,767
Catégories	-1,00	-0,556
Clients	-1,00	-0,500

Tableau 1- Classement local et global des tables

On voit que le classement global reflète mieux la notion de multiplicité et d'agrégation. En effet, on s'attend à ce qu'une table qui n'a pas de liens entrants ait une valeur négative. Dans notre exemple, les tables, *Clients*, *Livreurs*, *Fournisseurs*, *Catégories* et *Départements* ont des valeurs plus petites que -0,500. On s'attend aussi à ce qu'une table qui n'a que des liens entrants ait une valeur proche de 1, ce qui est le cas de la table *Détails_commandes*. Finalement, le classement devrait croître lorsqu'on suit les liens, ce qui est vérifié par l'exemple.

On conclut que les tables ayant un classement global près de 1 contiennent les colonnes qui seront de bons indicateurs, alors que les tables ayant un classement près de -1 contiennent les colonnes que seront de bonnes dimensions.

4- Classement des colonnes

Pour chaque colonne, on détermine son classement, sans tenir compte du classement global de la table qui la contient. Ce classement ne tient compte que des caractéristiques propres à la colonne. Ce classement local (β) est ensuite combiné au classement global de la table (γ) pour obtenir le classement global de la colonne (λ). Dans ce qui suit, nous présentons en détails le calcul du classement local et du classement global des colonnes.

4.1- Classement local des colonnes

Le classement local des colonnes se fait à partir de plusieurs critères: la variabilité des valeurs, le type de données, la participation à la clé de la table. Le classement est obtenu en faisant la moyenne pondérée de ces critères. Nous présentons comment chacun de ces critères est évalué et comment ils sont combinés pour obtenir le classement local de la colonne. On voudra que chaque critère soit évalué dans l'intervalle [-1,1], où une valeur de -1 dénote une bonne dimension et une valeur de 1 dénote un bon indicateur.

4.1.1- Variabilité des valeurs

La variabilité des valeurs (critère χ_1) permet de déterminer la proportion de valeurs distinctes d'une colonne dans une table. On notera que plus la variabilité est grande, plus la colonne est un bon indicateur, alors qu'une faible variabilité implique une bonne dimension. Par exemple, une colonne ne contenant que les valeurs 0 et 1 sera un très mauvais indicateur.

Le calcul de χ_1 doit donc donner une valeur de 1 lorsque la variabilité est grande et de -1 lorsqu'elle est faible. Ce calcul est effectué avec la formule

$$\chi_1 = 2 \cdot \frac{\text{nombre valeurs distinctes}}{\text{nombre valeurs}} - 1$$

Un élément important à remarquer est que le nombre de valeurs distinctes ainsi que le nombre total de valeurs peuvent être obtenus à l'aide de simples requêtes SQL.

4.1.2- Type des données

Le type de données de la colonne peut influencer le classement. Le critère χ_2 servira à indiquer l'influence du type sur le classement local. Ainsi, les types de données numériques (virgule flottante, réel, entier) tendent à être de bons indicateurs; dans ce cas, χ_2 prend la valeur 1. Par contre, les types textuels (chaîne de caractères, date) sont habituellement utilisés pour stocker de l'information statique, et donneront de bonnes dimensions; dans ce cas, χ_2 prend la valeur -1 [Gluck et Corter 85].

4.1.3- Participation à la clé

Lorsqu'une colonne fait partie de la clé, elle est nécessairement une bonne dimension et un mauvais indicateur. Le critère χ_3 permet de forcer le classement comme dimension lorsque la colonne fait partie de la clé, en lui donnant la valeur -1. Pour les autres colonnes, χ_3 prend la valeur 1.

4.1.4- Classement arbitraire de l'usager

Un usager peut influencer le classement d'une colonne en donnant des valeurs au critère χ_4. Ce critère donne la possiblité à l'usager de biaser le calcul des classements.

4.1.5- Calcul du classement local

À chaque critère χ_i on associe un poids k_i, qui sera utilisé pour pondérer les valeurs des différents critères dans le calcul du classement local β. Les différents poids doivent refléter les relations entre les différents critères. Par exemple, une colonne numérique avec faible variabilité sera un mauvais indicateur et une mauvaise dimension; il en sera de même pour une colonne textuelle avec grande variabilité. Il faut donc que les critères de variabilité et du type de données puissent s'annuler. Leurs poids relatifs doivent donc être égaux ($k_1 = k_2$). On peut faire le même raisonnement pour le critère de participation à la clé; malgré le fait qu'une colonne numérique avec grande variabilité soit un bon indicateur, si elle fait partie de la clé, on ne voudra pas utiliser cette colonne comme indicateur. Il faut donc que le poids associé au critère de participation à la clé puisse annuler les critères de variabilité et de type de données ($k_3 \geq k_1+k_2$).

Une fois les poids déterminés (dans notre cas $k_1=k_2= 1$, $k_3=2$, $k_4=0$), le classement local de la colonne c_j est effectué par la formule suivante:

$$\beta_{c_j} = \frac{\sum_{i=1}^{4} k_i \chi_i}{\sum_{i=1}^{4} k_i}$$

Le tableau 2 donne les caractéristiques de quelques colonnes choisies dans notre exemple. On remarquera que les colonnes choisies peuvent être le résultat d'une opération arithmétique, comme c'est le cas de la colonne *Date_requise - Date_livraison*. Le tableau indique la table contenant la colonne, le nombre de valeurs distinctes, le nombre de tuples dans la table, inclusion de la colonne dans la clé, ainsi que le type de la colonne. Ces valeurs serviront au calcul des valeurs associées aux différents critères χ_i et à β, qui sont données au tableau 3. On peut voir que les colonnes numériques non-clés sont de bons indicateurs, avant classement global.

Colonne	Table	Valeurs distinctes	Tuples	Clé ?	Type
Date_livraison	Commandes	546	1078	non	date
Date_requise - Date_livraison	Commandes	57	1078	non	entier
No_commande	Commandes	1078	1078	oui	entier
Nom_département	Départements	3	3	non	texte
No_département	Départements	3	3	oui	entier
No_produit	Détails_commandes	77	2820	oui	entier
Quantité	Détails_commandes	55	2820	non	entier
Total	Détails_commandes	1030	2820	non	réel

Tableau 2- Données sur quelques colonnes

Colonne	Table	χ_1	χ_2	χ_3	β
Date_livraison	Commandes	0,013	-1	1	0,253
Date_requise - Date_livraison	Commandes	-0,894	1	1	0,526
No_commande	Commandes	1,000	1	-1	0,000
Nom_département	Départements	1,000	-1	1	0,500
No_département	Départements	1,000	1	-1	0,000
No_produit	Détails_commandes	-0,945	1	-1	-0,486
Quantité	Détails_commandes	-0,961	1	1	0,510
Total	Détails_commandes	-0,270	1	1	0,683

Tableau 3- Classement local des colonnes

4.2- Classement global des colonnes

Le classement global d'une colonne c_j est fonction du classement local de la colonne (β_{c_j}) et du classement global de la table qui contient cette colonne ($\lambda_{T(c_j)}$); notons que $T(c_j)$ est une fonction qui retourne la table contenant la colonne c_j. Pour calculer cette fonction, nous utilisons la formule suivante:

$$\lambda_{c_j} = 2 \cdot \left(\frac{\beta_{c_j} + 1}{2} \right) \cdot \left(\frac{\gamma_{T(c_j)} + 1}{2} \right) - 1 = \frac{(\beta_{c_j} + 1) \cdot (\gamma_{T(c_j)} + 1)}{2} - 1$$

Une colonne ayant un classement global près de 1 sera un bon indicateur alors qu'une colonne avec un classement près de -1 sera une bonne dimension. Le tableau 4 donne les valeurs de β_{cj}, $\gamma_{T(cj)}$ et λ_{cj} pour les colonnes sélectionnées. On remarque que les colonnes *Total* et *Quantité* sont les meilleurs indicateurs, avec des valeurs de λ de 0,486 et 0,334, respectivement. Cela correspond à ce qu'on aurait pu s'attendre intuitivement. On remarque aussi *No_commande* (λ=-0,417), *No_produit* (λ=-0,546), *Nom_département* (λ=-0,750) et *No_département* (λ=-0,833) sont de bonnes dimensions; on note que ces deux colonnes forment la clé de *Details_commandes*.

Colonne	Table	β	γ	λ
Date_livraison	Commandes	0,253	0,167	-0,269
Date_requise - Date_livraison	Commandes	0,526	0,167	-0,110
No_commande	Commandes	0,000	0,167	-0,417
Nom_département	Départements	0,500	-0,667	-0,750
No_département	Départements	0,000	-0,667	-0,833
No_produit	Détails_commandes	-0,486	0,767	-0,546
Quantité	Détails_commandes	0,510	0,767	0,334
Total	Détails_commandes	0,683	0,767	0,486

Tableau 4- Classement global des colonnes

5- Conclusion

Nous avons présenté une métrique qui permet de classer les différentes colonnes d'une base de données selon qu'elles seront de bons indicateurs ou de bonnes dimensions. Les formules utilisées résultent d'un effort à trouver une métrique offrant le classement le plus réaliste possible, compte tenu des facteurs caractérisant un indicateur et une dimension.

Comme le démontrent les résultats obtenus dans l'exemple de l'entreprise de distribution, le classement global des colonnes reflète la réalité. On n'a qu'à penser au cas de la colonne *Total* qui représente la vente totale dans une ligne de commande. Cette colonne se trouve dans la table ayant la plus grand multiplicité et le meilleur classement local (*Détails_commandes*). De plus, cette colonne a toutes les caractéristiques d'un bon indicateur: grande variabilité, type numérique, de faisant pas partie de la clé. Intuitivement, on utiliserait certainement cette colonne comme indicateur (ventes de l'entreprise).

On peut alors déterminer les différentes dimensions qui pourraient être utilisées pour cet indicateur. Par exemple, on peut avoir les ventes par commande (dimension = *No_commande*), puis par département (dimension = *No_département*). On peut aussi avoir les ventes par produit (dimension = *No_produit*). En fait, pour chaque dimension, on pourrait créer autant de matrices que de dimensions.

En réalité, on n'a qu'à créer une matrice par chemin fonctionnel. Par exemple,un des chemins fonctionnels partant de la table *Détails_commandes* passe successivement par les tables *Commandes*, *Employés* et *Départements*. Ainsi, une seule matrice peut contenir l'information sur les ventes par département et par commande, les deux dimensions étant fonctionnellement liées. Le long du chemin fonctionnel, on n'a qu'à choisir la meilleure dimension dans chacune des tables rencontrées.

L'avantage de la métrique proposée est que sont calcul peut être complètement automatisé. Par exemple, dans un environnement hétérogène, on peut utiliser les outils ODBC disponibles pour générer le graphe des liens fonctionnels, et de ce fait, générer les classements local et global des tables. De même, quelques requêtes SQL permettent de déterminer le degré de variabilité d'une colonne, son type et l'inclusion de cette colonne dans la clé.

Remerciements

Ce travail a été subventionné par le Conseil de Recherche en Sciences Naturelles et Génie du Canada (CRSNG #OGP0155899). Les auteurs tiennent à remercier André Gamache pour les commentaires constructifs qu'il a apporté.

Bibliographie

[Gluck et Corter 85] M. A. Gluck and J. E.Corter. Information, uncertainty, and the unity of categories. *In Proceedings of the 7th Annuel Conference of the Cognitive Science Society*, pages 283-287, Irvine, CA, 1985.