Découverte de proportions analogiques dans les bases de données : une première approche

William Correa Beltran,
Hélène Jaudoin,
Olivier Pivert

Université de Rennes 1 – Irisa
Lannion, France
{William.Correa_Beltran@irisa.fr, Helene.Jaudoin@irisa.fr, Olivier.Pivert@irisa.fr}

Résumé. Cet article présente un nouveau cadre pour la découverte de connaissances basé sur la notion de proportion analogique qui exprime l'égalité des rapports entre les attributs de deux paires d'éléments. Cette notion est développée dans le contexte des bases de données pour découvrir des parallèles dans les données. Dans un premier temps, nous donnons une définition formelle des proportions analogiques dans le cadre des bases de données relationnelles, puis nous étudions le problème de l'extraction des proportions analogiques. Nous montrons qu'il est possible de suivre une approche de clustering pour découvrir les classes d'équivalence de paires de n-uplets dans le même rapport de proportion analogique. Ce travail constitue une première étape vers l'extension des langages d'interrogation de base de données avec des requêtes « analogiques ».

1 Introduction

Les travaux de cette dernière décennie dans le domaine de la découverte de connaissances, comme ceux notamment de (Han et al. (2007)) et de (Tiwari et al. (2010)), témoignent du vif intérêt pour le problème de l'extraction des ensembles d'items fréquents, des motifs séquentiels, des motifs structurels (dans les données de type arbres, graphes ou treillis), et la recherche de méthodes efficaces pour extraire ces motifs. Dans cet article, nous nous intéressons à la notion de proportion analogique, essentiellement étudiée dans le domaine de l'intelligence artificielle, pour extraire de nouveaux types de motifs dans les bases de données. Les proportions analogiques relient quatre objets A, B, C, D du même type dans une assertion de la forme « A est à B ce que C est à D ». Ils permettent d'exprimer l'identité (ou la proximité) des rapports existant entre deux paires d'éléments. Des exemples typiques de cette notion en langage naturel sont : « le veau est à la vache ce que le poulain est à la jument », « l'aurochs est au bœuf ce que le mammouth est à l'éléphant ». Ces relations permettent d'exprimer que ce qui distingue A de B est comparable à ce qui distingue C de D. Les exemples ci-dessus montrent la diversité (et la potentielle complexité) des sémantiques possibles

du connecteur « est à » intervenant dans une proportion analogique. Dans le premier exemple, ce connecteur représente une relation de filiation tandis que dans la seconde, il exprime une évolution possible. Le connecteur « ce que » de la proportion représente généralement l'identité ou la similarité. Quand les éléments A, B, C et D sont des valeurs numériques, la relation peut être définie en utilisant les proportions mathématiques classiques, comme la proportion géométrique : $A/B = C/D$ (e.g., $1/3 = 2/6$) ou la proportion arithmétique : $A - B = C - D$ (e.g., $5 - 3 = 9 - 7$). Quand les objets A et B, resp. C et D, représentent les mêmes entités à différents moments ou états de leur vie (par exemple, A et B décrivent le même endroit à deux moments différents), la proportion analogique peut exprimer des évolutions similaires. De manière générale, les proportions analogiques permettent de trouver des parallèles entre quatre événements ou situations.

Nous cherchons à exploiter la notion de proportion analogique dans le contexte des bases de données relationnelles afin d'extraire des combinaisons de quatre n-uplets liés par une telle relation. Notre objectif est de découvrir des *parallèles* entre des paires de n-uplets, i.e., des paires d'éléments qui sont dans les mêmes rapports. Ces parallèles ne reflètent pas forcément une relation de proximité (A est aussi proche de B que C est proche de D), mais plutôt une transformation semblable (On passe de A à B comme on passe de C à D). Ces parallèles sont d'une importance majeure puisqu'il permettent de modéliser des règles d'évolution reproductible dans les systèmes écologiques (les états de deux littoraux qui évoluent dans les mêmes directions : apparition et disparition des mêmes espèces, évolution d'une pollution d'une région à une autre), des mouvements sociétaux (extension d'une crise géopolitique ou comparaison avec des successions d'événements passés), ou des déplacements parallèles d'objets.

Les contributions de cet article sont les suivantes. Nous proposons une méthode pour identifier les proportions analogiques dans les bases de données. À cette fin, nous suivons une approche vectorielle pour définir la notion de proportion analogique adaptée au modèle relationnel. Puis nous montrons qu'il est possible de ramener le problème d'énumération de toutes les combinaisons de quatre n-uplets liés par une relation d'analogie, à un problème de clustering moyennant un prétraitement et l'utilisation d'une métrique. Ceci permet de rassembler des paires d'éléments qui définissent des vecteurs égaux ou presque. Nous analysons ensuite les résultats de notre approche appliquée à un jeu de données réelles.

Notre article est organisé comme suit. En section 2, nous introduisons la notion de proportion analogique et nous proposons une définition graduelle de celle-ci adaptée au contexte des bases de données. La section 3 présente notre approche de découverte des proportions analogiques et l'algorithme qui en découle, tandis que la section 4 détaille les expérimentations effectuées. Enfin, nous présentons les travaux relatifs à notre proposition (Section 5) puis nous concluons (Section 6).

2 Proportions analogiques et modélisation

2.1 Les proportions analogiques

Cette section s'appuie sur les références (Miclet et Prade (2009)) et (Lepage (2012)). Une proportion analogique est une assertion de la forme « A est à B ce que C est à D», notée par la suite $(A : B :: C : D)$ où : dénote un rapport et ::, appelé conformité, exprime la relation entre deux paires d'éléments. La proportion analogique exprime alors la conformité entre deux paires d'objets.

En général, les objets A, B, C, et D correspondent à des descriptions d'items prenant la forme d'ensembles, de multi-ensembles, de vecteurs, de chaînes de caractères et d'arbres. Une interprétation possible de la relation d'analogie est la suivante : A peut être similaire (ou identique) à B d'une certaine façon et différer selon d'autres critères. Cependant, la façon dont C diffère de D doit être la même ou presque la même que la façon dont A diffère de B si on s'en tient à une généralisation des proportions mathématiques.

De plus, il est admis que les proportions analogiques doivent vérifier les postulats suivants :

- (ID) $(A : B :: A : B)$
- (S) $(A : B :: C : D) \Leftrightarrow (C : D :: A : B)$
- (CP) $(A : B :: C : D) \Leftrightarrow (A : C :: B : D)$.

(ID) et (S) expriment la réflexivité et la symétrie selon le connecteur « comme» (::), alors que (CP) exprime la permutation des termes moyens.

Un algorithme naïf pour énumérer les proportions analogiques issues d'un ensemble d'objets de cardinalité n, a une complexité temporelle en n^4. En utilisant un point de vue vectoriel de la notion de proportion analogique, les objets A, B, C et D désignent des points d'un espace à n dimensions. S'ils forment une relation d'analogie alors ces points forment un parallèlogramme. Par exemple, la figure 1 montre la relation de proportion analogique existant entre les points $A(1, 2)$, $B(4, 4)$, $C(3, 1)$, $D(6, 3)$ représentés dans un repère orthonormé.

Ainsi quatre objets A, B, C et D sont en proportion analogique si et seulement si

$$\overrightarrow{AB} = \overrightarrow{CD}$$

et donc si et seulement si

$$||\overrightarrow{AB} - \overrightarrow{CD}|| = 0.$$

La relation de proportion analogique liant A, B, C, et D peut être alors symbolisée par le vecteur $\overrightarrow{AB}$ (ou $\overrightarrow{CD}$).

Dans ce cas particulier (la conformité est la relation d'identité), il est possible de définir un algorithme dont la complexité temporelle est en n^2 : celui-ci calcule tous les vecteurs existant entre les paires de n-uplets et rassemblent toutes les paires de n-uplets définissant des vecteurs égaux en une classe d'équivalence (Lepage (2012)). Une classe d'équivalence, représentée par un vecteur, rassemble ainsi des paires de points qui, prises deux à deux, sont en proportion analogique « selon ce vecteur ». Il est alors aisé de générer l'ensemble de toutes les relations d'analogie à partir de chacune des classes d'équivalence.

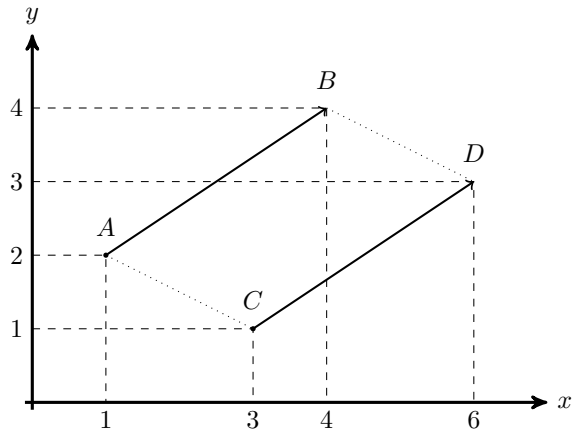

FIG. 1 – *Quatre points en proportion analogique dans un espace à deux dimensions*

Exemple :
Supposons que l'on ait 9 points (A, B, C, D, E, F, G, H et I) définis dans un espace à n dimensions.
On génère tous les vecteurs possibles puis on rassemble ceux qui sont égaux [1].
Supposons que $\overrightarrow{AB} = \overrightarrow{CD}$, et que $\overrightarrow{GH} = \overrightarrow{FD} = \overrightarrow{EI}$.
On en déduit deux classes d'équivalence $\mathcal{C}_{\overrightarrow{AB}} = \{(A, B),\ (C, D)\}$ et
$\mathcal{C}_{\overrightarrow{GH}} = \{(G, H),\ (F, D),\ (E, I)\}$ qui mettent en évidence les proportions analogiques suivantes :

- $A : B :: C : D$ et donc $A : C :: B : D$
- $G : H :: F : D$ et donc $G : F :: H : D$, $G : H :: E : I$ et donc $G : E :: H : I$, $F : D :: E : I$ et donc $F : E :: D : I$.

◇

Dans la suite, nous considérons des proportions analogiques sur des données numériques tout en relaxant la notion d'identité des rapports. En effet, l'égalité des rapports est en générale trop restrictive quand on manipule des données réelles.

2.2 Modéliser les proportions analogiques selon une approche géométrique

La modélisation des proportions analogiques dans le contexte des bases de données relationnelles est influencée par les propriétés du modèle relationnel. Soit un ensemble d'attributs $\{A_1, \ldots, A_m\}$, un schéma de relation est défini comme un sous-ensemble d'attributs $S = \{A_{i_1}, \ldots, A_{i_n}\}$. Une relation définie en termes d'un schéma de relation S est un sous-ensemble fini du produit cartésien des domaines de chacun des attributs de S. Chaque élément d'une relation est appelé n-uplet qui peut être représenté par

1. La propriété de permutation des moyens permet d'éviter de calculer à la fois $\overrightarrow{AB}$ et $\overrightarrow{CD}$ puis $\overrightarrow{AC}$ et $\overrightarrow{BD}$.

un point décrit par n dimensions. Une base de données est un ensemble fini de relations. D'autres contraintes additionnelles comme les dépendences fonctionnelles et les dépendances d'inclusion permettent de restreindre le contenu des relations. Il serait intéressant d'en tenir compte dans la recherche des proportions analogiques mais nous nous limiterons ici au cas de la recherche de proportions existantes entre quatre points.

Ainsi les n-uplets A, B, C, et D d'une relation sont considérés comme des points à n dimensions, et sont dénotés comme suit : $A = (a_1, \ldots, a_n)$, $\ldots$, $D = (d_1, \ldots, d_n)$.

Comme dit précédemment, A, B, C, et D sont liés par une relation de proportion analogique si et seulement si $\overrightarrow{AB} = \overrightarrow{CD}$. L'égalité est difficile à obtenir quand on considère des jeux de données réels. Il convient alors de rendre cette définition plus flexible, notamment en donnant une vision plus graduelle de la relation de proportion analogique. Deux vecteurs ne doivent plus être égaux mais presque égaux ce qui revient à mesurer dans quelle mesure $||\overrightarrow{AB} - \overrightarrow{CD}||$ est proche de 0. On cherche alors à évaluer la « distorsion » entre les deux vecteurs. Pour permettre la commensurabilité des dimensions lorsque les attributs portent sur des domaines différents, les valeurs des vecteurs sont normalisées afin qu'elles appartiennent à l'intervalle $[0, 1]$. Pour cela, chaque valeur v du domaine actif d'un attribut peut être remplacée par la valeur suivante :

$$\frac{v \; - \; min_{att}}{max_{att} \; - \; min_{att}}$$

où min_{att} et max_{att} désignent respectivement la valeur minimale et la valeur maximale du domaine actif de l'attribut.

Plusieurs stratégies peuvent être utilisées pour mesurer à quel point l'expression $||\overrightarrow{AB} \; - \; \overrightarrow{CD}||$ est proche de $||\overrightarrow{0}||$. Différentes normes peuvent être utilisées comme la norme de Minkowsky (norme p), qui donnera la longueur du vecteur correctif permettant de passer de $\overrightarrow{AB}$ à $\overrightarrow{CD}$, ou, la norme infinie qui donnera la coordonnée maximale de ce vecteur correctif.

Définition 1 : Distorsion analogique fondée sur une norme infinie
Soit A, B, C, et D quatre n-uplets de n dimensions.
Posons $\overrightarrow{u} = \overrightarrow{AB}$, $\overrightarrow{v} = \overrightarrow{CD}$.

$$Dist_\infty(\overrightarrow{AB}, \overrightarrow{CD}) \; = \; Dist_\infty(\overrightarrow{u}, \overrightarrow{v}) \; = \; max_{i \in \{1,\ldots,n\}} |u_i - v_i|$$

La norme infinie retourne la plus grande différence de dimension entre les deux vecteurs. Les deux vecteurs sont d'autant plus égaux que le changement maximal sur une dimension est proche de zéro.

Définition 2 : Distorsion analogique fondée sur la norme p
Soit A, B, C, et D quatre n-uplets de n dimensions.
Posons $\overrightarrow{u} = \overrightarrow{AB}$, $\overrightarrow{v} = \overrightarrow{CD}$.

$$Dist_p(\overrightarrow{AB}, \overrightarrow{CD}) \; = \; Dist_p(\overrightarrow{u}, \overrightarrow{v}) \; = \; (\sum_{i \in \{1,\ldots,n\}} |u_i - v_i|^p)^{1/p}$$

Dans ce cas, la définition met l'accent sur la longueur du vecteur correctif permettant de passer de $\overrightarrow{AB}$ à $\overrightarrow{CD}$.

Dans tous les cas (Définition 1 ou Définition 2), la relation de proportion analogique est d'autant plus vraie que la distorsion est proche de 0.

Proposition :
Les deux définitions de distorsion vérifient les propriétés fondamentales des proportions analogiques.
En effet, les propriétés suivantes sont vérifiées :

- l'identité : $Dist(\overrightarrow{AB}, \overrightarrow{AB}) = 0$
- la symétrie : $Dist(\overrightarrow{AB}, \overrightarrow{CD}) = Dist(\overrightarrow{CD}, \overrightarrow{AB})$ puisque les deux définitions reposent sur une valeur absolue des différences entre chaque coordonnée.
- la permutation des moyens : la relation $Dist(\overrightarrow{AB}, \overrightarrow{CD}) = Dist(\overrightarrow{AC}, \overrightarrow{BD})$ est vraie quelle que soit la norme. En effet, on a pour tout i, $((b_i - a_i) - (d_i - c_i)) = ((c_i - a_i) - (d_i - b_i))$. Ainsi, on a toujours $||\overrightarrow{AB} - \overrightarrow{CD}|| = ||\overrightarrow{AC} - \overrightarrow{BD}||$.

Dans la suite, nous utiliserons la norme infinie qui est la plus drastique et possède un meilleur pouvoir de discrimination dans la mesure où elle évite tout effet de compromis entre les composantes des vecteurs.

3 Découvrir les proportions analogiques

Notre objectif est de découvrir toutes les proportions analogiques présentes dans un ensemble de données et si possible de dégager des tendances, i.e., les différents vecteurs représentatifs des proportions analogiques découvertes. Une approche naïve pourrait consister à énumérer tous les vecteurs et à calculer la distorsion entre chaque paire de vecteurs, puis à ne garder que les paires de vecteurs dont la valeur de distorsion ne dépasse pas un certain seuil. Cependant, une telle approche poserait la question du choix du seuil, très dépendant des données. Il nous semble par ailleurs préférable d'utiliser une technique permettant de fournir une vue synthétique des motifs découverts. Une approche de type clustering semble tout à fait appropriée dans ce contexte. Un argument supplémentaire en faveur d'une telle approche est lié à l'objectif final que nous nous sommes fixé, à savoir l'extension des langages d'interrogation de bases de données avec des requêtes analogiques, i.e., des requêtes visant à découvrir des proportions analogiques existant dans un ensemble de données. En effet, l'identification de classes d'équivalence regroupant les paires de points représentant des vecteurs (presque) égaux permettrait la définition d'index, utiles pour optimiser l'évaluation de telles requêtes. Le problème ici étudié, qui consiste à regrouper des vecteurs à n dimensions égaux ou presque, se ramène à un problème de clustering classique dès lors que l'on dispose d'une métrique. Or les définitions des distorsions analogiques ($Dist(\overrightarrow{u}, \overrightarrow{v})$) satisfont les conditions qui caractérisent les métriques, soit :

- l'identité des indiscernables : $Dist(\overrightarrow{u}, \overrightarrow{v}) = 0$ ssi $\overrightarrow{u} = \overrightarrow{v}$,
- la propriété de symétrie et
- l'inégalité triangulaire.

Différentes approches de clustering sont donc utilisables, comme les k-means et l'approche hiérarchique (Xu et al. (2005)). Notre objectif étant de tester l'approche et de la valider sur des jeux de données réels, puis d'identifier les relations découvertes, nous avons choisi de reprendre un algorithme de clustering hiérarchique (Xu et al.

(2005)) dont les étapes sont énoncées dans l'Algorithme 1. Cet algorithme requiert une étape préalable de construction des vecteurs à partir des points de la relation.

Data : Soit un ensemble S de n tuples $S = \{t_1, t_2, ..., t_n\}$
Result : L'ensemble des clusters de proportions analogiques de S
`/* Création des` m `premiers clusters */`
`/* formés d'un vecteur` $\overrightarrow{ij}$ `*/`
Calculer le vecteur $\overrightarrow{ij}$ entre chaque paire de tuples i et j de S;
Calculer la matrice des distorsions entre chaque paire de vecteurs;
while m *!=1* **do**
  Trouver la paire de vecteurs $(\overrightarrow{ij}, \overrightarrow{kl})$ ayant la distorsion la plus petite;
  Fusionner les clusters formés de $\overrightarrow{ij}$ et $\overrightarrow{kl}$ en un cluster dont le représentant est $(\overrightarrow{ij} + \overrightarrow{kl})/2$;
  Actualiser la matrice des distorsions : mise à jour des clusters et des distorsions;
  m :=m-1;
end

Algorithme 1: Calcul des clusters de proportions analogiques

La complexité temporelle de cet algorithme qui est fonction du nombre m de vecteurs générés, est en m^2. Notons néanmoins que la complexité temporelle de l'algorithme est linéaire dans le cas où l'on cherche à lier par une proportion analogique des paires d'objets dans deux états différents de leur vie $(A^1 : A^2 :: B^1 : B^2)$. C'est le cas lorsqu'on cherche à trouver des évolutions communes de situations. Dans ce cas, on ne cherche plus à générer tous les vecteurs possibles à partir des points A et B respectivement. On crée uniquement un vecteur à partir de A reflétant son évolution et un second à partir de B. Un exemple de proportion analogique représentant ce cas d'analogie est le suivant : la description de la "Côte d'Albâtre" en termes de diversité végétale en 2000 est à sa description en 2010 ce que la description de la "Côte d'Argent" en 2008 est à sa description en 2012.

4 Expérimentations

Dans cette section, nous illustrons notre approche avec des données électorales afin de découvrir des parallèles entre les résultats des votes de différentes régions, puis des évolutions des résultats de votes d'une année à l'autre. Pour cela, nous avons exploité les jeux de données ouverts décrivant les résultats des élections présidentielles en France en 2007 [2] et en 2012 [3]. Ces jeux de données contiennent les votes obtenus par chaque candidat, par région, par département et par circonscription législative, et ce, pour les deux tours. Nous nous sommes concentrés sur les pourcentages des votes par région. Les données sont donc déjà normalisées. La mesure de distorsion utilisée repose sur la norme infinie.

2. https ://www.data.gouv.fr/fr/datasets/election-presidentielle-2007-resultats-572120/
3. https ://www.data.gouv.fr/fr/datasets/election-presidentielle-2012-resultats-572124/

À des fins de lisibilité graphique de la qualité des clusters créés, les résultats des votes ont été agrégés selon trois dimensions (à *gauche*, au *centre* et à *droite*) représentant l'orientation politique générale de chaque candidat. La table 1 montre, pour un échantillon des régions de la France, les votes obtenus par la gauche, le centre et la droite, en 2007 et en 2012, au premier tour.

TAB. 1 – *Votes par région, en 2007 et 2012, selon l'orientation politique*

Régions		2007			2012		
		gauche	centre	droite	gauche	centre	droite
11	Ile de France	0.3673	0.2001	0.4327	0.475	0.0946	0.4303
53	Bretagne	0.3934	0.2255	0.3811	0.4769	0.1162	0.407
83	Auvergne	0.4199	0.1961	0.3839	0.4942	0.0979	0.4079
23	Rhône-Alpes	0.3357	0.201	0.4633	0.4109	0.1021	0.4871
91	Languedoc-Roussillon	0.3718	0.1522	0.476	0.4334	0.0702	0.4964
25	Basse-Normandie	0.3403	0.2023	0.4573	0.4156	0.1083	0.4761
52	Pays de la Loire	0.3575	0.2118	0.4305	0.43	0.1189	0.451
54	Poitou-Charentes	0.4003	0.1799	0.4197	0.4605	0.0979	0.4416
72	Aquitaine	0.3912	0.214	0.3948	0.476	0.1094	0.4146

Deux scénarios ont été testés. Dans le premier scénario, seuls les votes de l'année 2012 ont été utilisés. L'objectif était de trouver les quadruplets de régions en proportion analogique. Dans ce cas, chaque couple de tuples de ce tableau constitue un des clusters servant d'entrée à l'algorithme. Par la suite, chaque cluster formé contenant au moins 4 régions r_a, r_b, r_c, et r_d, indique que l'on a observé les mêmes différences de votes par orientation politique, entre les régions r_a et r_b et les régions r_c et r_d respectivement. Lors de ce test, les proportions analogiques suivantes ont notamment été trouvées :

– $53 : 83 :: 23 : 91$ et donc $53 : 23 :: 83 : 91$
– $25 : 52 :: 54 : 72$ et donc $25 : 54 :: 52 : 72$.

La table 1, dans la partie 2012, montre les détails de chacune de ces régions, et la table 2, le vecteur représentatif de chaque paire de régions. À titre d'exemple, la distorsion entre le vecteur identifié par c_1 et celui identifié par c_2 vaut $max_{i\in\{1,\ldots,n\}}|c_{1_i} - c_{2_i}| = |0.0138 - 0.0125| = 0.0013$. Ces proportions reflètent une forme de similarité dans les comportements.

TAB. 2 – *Vecteurs formés entre les régions*

identifiant	région 1	région 2	gauche	centre	droite
c_1	53	83	0.0032	0.0138	-0.0171
c_2	23	91	0.0043	0.0125	-0.0167
c_3	25	52	-0.0144	-0.0106	0.0251
c_4	54	72	-0.0155	-0.0115	0.027

Le second scénario consiste à trouver les couples de régions pour lesquelles on observe une évolution similaire de leur orientation politique de l'année 2007 à l'année 2012. Dans ce cas, le vecteur à trois dimensions (gauche, centre, droit) correspond à l'évolution des votes de 2007 à 2012 pour une même région. Par exemple, pour la région 11 (Ile de France), l'évolution des votes est $\langle(0.475-0.3673), (0.0946-0.2001), (0.4303-0.4327)\rangle = \langle 0.1077, -0.1055, -0.0024\rangle$.

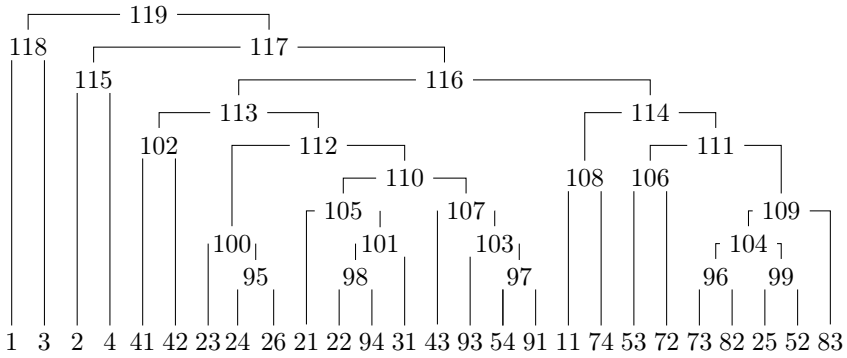

FIG. 2 – *Dendogramme représentatif du clustering hiérarchique des évolutions des votes par région entre les élections présidentielles de 2007 et de 2012*

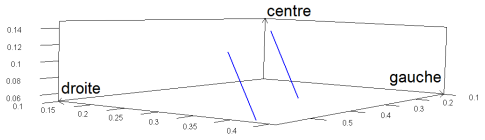

FIG. 3 – *Cluster* 98 *: regroupement des vecteurs* 22 *et* 94

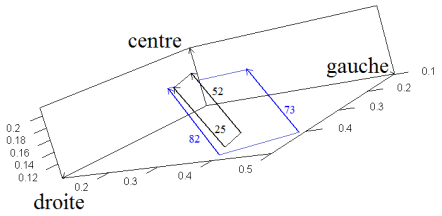

FIG. 4 – *Cluster* 104 *: regroupement des vecteurs* 73*,* 82*,* 25 *et* 52

La Figure 2 montre le dendrogramme qui représente les clusters dans le cas de la recherche d'évolution des votes entre 2007 et 2012 par région. Analysons plus précisément les trois clusters 118, 113 et 114, le premier regroupant uniquement deux points et les deux derniers représentatifs de plusieurs régions. L'étude des vecteurs caractéristiques de chacun de ces clusters montrent des comportements très différents. Le cluster 118 représenté par le vecteur $\langle +17, -4.5, -13\rangle$ montre une nette progression des votes pour les partis de gauche, une baisse des votes au centre et une chute des votes pour les partis de droite pour les régions 1 (Guadeloupe) et 3 (Guyane). En revanche, le cluster 114 représenté par le vecteur $\langle +9.43, -10.1, +0.73\rangle$ met en évidence des régions (comme notamment l'Ile de France, le Limousin, la Bretagne et l'Aquitaine) pour lesquelles on observe une chute des votes au centre et une progression importante des votes à gauche. Les clusters 113 et 114 sont assez semblables au premier abord. Cependant le cluster 113 ($\langle 5.45, -8.79, +3.33\rangle$) regroupe des régions (en gris foncé sur la figure 5) pour lesquelles la progression des votes à gauche est moins importante que celle observée dans les régions du cluster 114 (en gris clair sur la figure 5).

La Figure 3 montre les évolutions parallèles des régions 22 et 94 (cluster 98) tandis que la figure 4 détaille le cluster 104 formé des régions 73, 25, 82 et 52 (cf. Figure 2).

Les expérimentations visaient à montrer que le cadre proposé permet de mettre en

FIG. 5 – *Carte des évolutions des votes entre 2007 et 2012*

évidence des parallèles existant dans les données, ce que montrent nos premiers résultats. Vu la nature de la relation de proportion analogique, l'approche peut évidemment s'appliquer à bien d'autres domaines, comme la recherche de trajectoires parallèles d'objets mobiles, moyennant une adaptation, ou dans les domaines environnemental et sociétal, pour découvrir des évolutions analogues.

5 État de l'art

L'originalité de l'approche présentée ici tient à la nature même du type de régularité que l'on cherche à découvrir dans les données. La majeure partie des travaux en fouille de données visent à découvrir des *caractéristiques* fréquentes (vues comme des valeurs d'attribut ou des séquences de valeurs d'attribut lorsqu'un aspect temporel est pris en compte) dans un ensemble de données. Avec les proportions analogiques, qui sont des relations quaternaires, nous cherchons à vérifier l'existence de "parallèles" entre des couples d'objets d'une collection. Une proportion analogique, lorsqu'elle met en parallèle les situations de deux éléments à deux moments différents, peut être vue comme une sorte de règle d'évolution. Dans un tel contexte, la recherche de proportions analogiques peut constituer une alternative aux approches de la littérature visant à

- extraire des règles d'évolution dans des graphes (Berlingerio et al. (2009)), ou à
- découvrir des trajectoires parallèles d'objets en mouvement (Vlachos et al. (2002); Chen et al. (2005); Lee et al. (2007); Li et al. (2013)), ou encore à
- classer des séquences d'événements (Studer et al. (2010); Guigourès et al. (2014); Lin et al. (2003); Malinowski et al. (2013); Zhou et al. (2013)).

Quoi qu'il en soit, l'approche proposée fournit un cadre plus général que celui dédié spécifiquement à l'extraction de règles d'évolution dans les données. En effet, la notion de proportion analogique n'implique pas l'existence d'une dimension temporelle et peut servir à décrire des parallèles de nature très variée entre deux paires d'objets.

6 Conclusion

Dans cet article, nous nous sommes intéressés à la recherche d'un nouveau type de motifs basé sur la notion de proportion analogique. Celle-ci permet de lier quatre tuples, représentés comme des points dans un espace à n dimensions, pour lesquels il est possible d'identifier des parallèles. Nous avons modélisé cette notion dans le contexte des bases de données en nous appuyant sur une approche vectorielle. Les points sont alors d'autant plus en proportion analogique qu'ils forment un parallèlogramme. Le problème de l'énumération des proportions analogiques existant dans une base de données revient ainsi à rechercher des vecteurs égaux ou presque. Moyennant un prétraitement, il est possible d'exploiter des algorithmes de clustering classiques pour découvrir les classes d'équivalence des paires de points représentant des vecteurs égaux ou presque, et par conséquent les différents motifs de changement, puis d'en déduire tous les quadruplets en proportion analogique. L'approche a été testée avec des jeux de données ouverts portant sur les résultats des élections présidentielles de 2007 et de 2012. À court terme, nous souhaitons utiliser différentes approches de clustering et appliquer différentes mesure de distance.

En terme de perspectives, nous envisageons d'étendre les langages d'interrogation de base de données à l'aide des relations de proportions analogiques, en permettant notamment de rechercher toutes les proportions existantes, toutes les proportions impliquant un point particulier, ou encore, celles impliquant une paire donnée de points. Dans ce cadre, les clusters pourront jouer le rôle d'index qui permettront d'accélérer l'évaluation des requêtes.

Remerciements : Ce travail a été partiellement financé par la région Bretagne et le Conseil Général des Côtes-d'Armor.

Références

Berlingerio, M., F. Bonchi, B. Bringmann, et A. Gionis (2009). Mining graph evolution rules. In *ECML/PKDD (1)*, pp. 115–130.

Chen, L., M. T. Özsu, et V. Oria (2005). Robust and fast similarity search for moving object trajectories. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 491–502. ACM.

Guigourès, R., D. Gay, M. Boullé, et F. Clérot (2014). Clustering de séquences d'évènements temporels. In *EGC'14*, pp. 191–202.

Han, J., H. Cheng, D. Xin, et X. Yan (2007). Frequent pattern mining : current status and future directions. *Data Min. Knowl. Discov. 15*(1), 55–86.

Lee, J.-G., J. Han, et K.-Y. Whang (2007). Trajectory clustering : a partition-and-group framework. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pp. 593–604. ACM.

Lepage, Y. (2012). (Re-)discovering the graphical structure of Chinese characters. In *SAMAI (workshop colacated with ECAI)*, pp. 57–64.

Li, Z., F. Wu, et M. Crofoot (2013). Mining following relationships in movement data. In *ICDM*, pp. 458–467.

Lin, J., E. Keogh, S. Lonardi, et B. Chiu (2003). A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pp. 2–11. ACM.

Malinowski, S., T. Guyet, R. Quiniou, et R. Tavenard (2013). 1d-sax : A novel symbolic representation for time series. In *Advances in Intelligent Data Analysis XII*, pp. 273–284. Springer.

Miclet, L. et H. Prade (2009). Handling analogical proportions in classical logic and fuzzy logics settings. In C. Sossai et G. Chemello (Eds.), *ECSQARU*, Volume 5590 of *Lecture Notes in Computer Science*, pp. 638–650. Springer.

Studer, M., N. S. Müller, G. Ritschard, et A. Gabadinho (2010). Classer, discriminer et visualiser des séquences d'événements. In *EGC*, pp. 37–48.

Tiwari, A., R. Gupta, et D. Agrawal (2010). A survey on frequent pattern mining : Current status and challenging issues. *Information Technology Journal 9*, 1278–1293.

Vlachos, M., G. Kollios, et D. Gunopulos (2002). Discovering similar multidimensional trajectories. In *Data Engineering, 2002. Proceedings. 18th International Conference on*, pp. 673–684. IEEE.

Xu, R., D. Wunsch, et al. (2005). Survey of clustering algorithms. *Neural Networks, IEEE Transactions on 16*(3), 645–678.

Zhou, C., B. Cule, et B. Goethals (2013). Itemset based sequence classification. In *Machine Learning and Knowledge Discovery in Databases*, pp. 353–368. Springer.

Summary

This paper presents an approach aimed at mining a new type of pattern in data, namely analogical proportions. An analogical proportion expresses the equality of the relationships between the attributes of two pairs of structured objects. This notion is investigated in the database context for the discovery of different forms of "parallels" between tuples. First, we give a formal definition of the analogical proportion in the setting of relational databases. Then we focus on the problem of mining analogical proportions. We show that it is possible to use a clustering approach for building equivalence classes made of pairs of tuples that are bound by the same relationship of analogical proportion. This work can be seen as a first step to the extension of database query languages that could be completed with "analogical queries".