Equivalence entre mesures de similarité floues : Application à la recherche d'images par le contenu

Jean-François Omhover, Bernadette Bouchon-Meunier
LIP6 – Pôle IA, Université Pierre et Marie Curie – Paris VI
contact : jean-francois.omhover@lip6.fr

Résumé : Pour retrouver des documents dans une base de données, les systèmes de recherche effectuent des comparaisons entre une requête et les descripteurs extraits des documents. Comme résultat, l'utilisateur obtient une liste des documents ordonnés par leur degré de ressemblance avec la requête. Le choix d'une mesure de similarité pour cette tâche est vaste. Dans cet article, nous montrons que les mesures de similarité peuvent être regroupées en classes de mesures équivalentes. Dans l'optique d'une recherche par similarité, nous montrons que le choix d'une mesure ou d'une autre peut être réduit au choix d'une famille de mesures. Enfin, nous explorons les conséquences de cette équivalence pour la recherche d'informations.

Abstract : To retrieve documents from a database, retrieval systems use similarity measures to compare a given request to the descriptors extracted from documents. As a result, documents are ordered in a list by decreasing similarity to the request. Several comparison measures are used in the field, and it is difficult to choose one or another. In this paper, we show that they can be grouped into classes of equivalent behaviour. The choice of these measures can then be reduced to the choice of a family of them.

Mots-clés : mesures de similarité floues , recherche d'images, segmentation

Keywords: fuzzy similarity measures, image retrieval, segmentation

1. Introduction

Les systèmes de recherche d'images par le contenu sont actuellement basés sur un même paradigme : la recherche par similarité à un exemple. Pour retrouver des images qui l'intéressent dans une base de données d'images, l'utilisateur fournit une image comme requête. Cette requête est ensuite comparée aux images de la base en termes visuels. Enfin on retourne à l'utilisateur la liste des images de la base qui ressemblent le plus à sa requête. Cette liste est ordonnée selon la similarité décroissante des résultats avec la requête.
Pour calculer la similarité entre l'image de requête et une image de la base, les systèmes utilisent diverses représentations visuelles extraites préalablement des images par un procédé automatique. Ces représentations, aussi appelées signatures, résument chaque image à une série de caractéristiques visuelles (couleur, texture, forme). Pour comparer deux signatures, et permettre la recherche par similarité, les systèmes utilisent diverses mesures de comparaison, les plus courantes étant les mesures de distance et les mesures de similarité.
Si l'utilisation d'une mesure de similarité peut permettre de couvrir des comportements différents en terme de comparaison (ressemblance, inclusion) répondant à différents besoins de l'utilisateur, le choix de ces mesures reste large. De plus, il n'est pas évident de déterminer, au sein de nombreuses mesures, quelle mesure permettra d'obtenir les meilleurs résultats.
Dans cet article, nous exposons une théorie permettant de répondre à ces questions. Nous montrons tout d'abord que dans le cadre de la recherche par comparaison à un (ou plusieurs) exemples, les mesures de similarité peuvent être regroupées en familles de mesures équivalentes dans le sens où elles aboutissent aux mêmes résultats. Nous montrons aussi que l'équivalence entre deux mesures conduit nécessairement les requêtes basées sur ces mesures à obtenir les mêmes taux de satisfaction. Enfin, nous montrons que dans le cadre connu et largement utilisé des mesures de Tversky, il existe une correspondance entre l'équivalence telle que nous la définissons et l'équivalence en terme de comportement de mesure.

Enfin, nous explorons les conséquences de cette équivalence pour la recherche d'informations visuelles.

2. Recherche d'images par similarité

Pour effectuer une recherche automatique d'images, les systèmes actuels se basent sur la comparaison à un ou plusieurs exemples. L'utilisateur fournit une image ou des régions de référence. Ces éléments sont ensuite comparés aux images de la base au moyen de différentes mesures de similarité. Dans cette partie, nous exposerons tout d'abord la définition des mesures de similarité. Nous verrons ensuite comment ces mesures sont utilisées pour effectuer des recherches automatiques, notamment dans le cadre de la recherche par des régions visuelles.

2.1. Les mesures de similarité floues

Les mesures de similarité sont divisées en deux catégories : les mesures géométriques et les mesures entre ensembles. Les modèles de distances (géométriques) sont les plus couramment utilisés. Les objets comparés sont considérés comme des points dans un espace métrique. Ces modèles sont contraints par quatre propriétés : la positivité, la symétrie, la minimalité et l'inégalité triangulaire.
Ces axiomes ont en particulier été étudiés par Tversky[1]. A partir d'une argumentation basée sur des considérations psychologique, il a proposé une approche basée sur la comparaison entre ensemble de caractéristiques. Dans le cadre proposé, les objets comparés sont décrits par leurs ensembles de caractéristiques binaires. Pour comparer deux objets *a* et *b*, décrits par leurs ensembles de caractéristiques *A* et *B*, les mesures de Tversky mettent en rapport trois composantes : les caractéristiques communes à *a* et *b* ($A \cap B$), et les caractéristiques propres à *a* ($A - B$) et propres à *b* ($B - A$). Une mesure de similarité *s* est ainsi définie par une fonction réelle F de trois variables $s(a,b) = F(A \cap B, A - B, B - A)$.
Les travaux de Tversky aboutissent à la proposition d'un modèle de mesure appelé le *Ratio Model*. Avec une mesure f permettant de mesurer les trois ensembles de caractéristiques communes et distinctives, on aboutit à la formulation suivante :

$$S(a,b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha . f(B - A) + \beta . f(A - B)}$$

où α et β, deux paramètres réels positifs, permettent de pondérer l'importance accordée aux deux ensembles distinctifs de caractéristiques (ceux qui appartiennent à *A* mais pas à *B*, et l'inverse).

A cause de la restriction des mesures de Tversky aux caractéristiques binaires (les ensembles *A* et *B* sont stricts, c'est-à-dire classiques), une extension de ces mesures a été proposée[2] pour permettre la comparaison entre ensembles graduels (ou flous). Dans ce cadre, pour Ω un ensemble d'éléments, on a $P_f(\Omega)$ l'ensemble des sous-ensembles flous de Ω, et *M* une mesure d'ensembles flous donnée telle que : *M* est une fonction de $P_f(\Omega)$ dans $\Re$, $M(\varnothing) = 0$, et *M* monotone au sens de l'inclusion (par exemple *M* mesure de l'aire d'un ensemble flou *A* défini par sa fonction caractéristique f_A: $M(A) = \sum_{x \in \Omega} f_A(x)$).

[1] A. TVERSKY, (1977), *Features of similarity*, p. 327-352, Psychological Review, vol 84
[2] BOUCHON-MEUNIER B., RIFQI M. And BOTHOREL S., (1996), *Towards general measures of comparison of objects*, p. 143-153, Fuzzy sets and systems, vol. 84(2).

Définition 1 : Une *M-mesure de comparaison* S sur Ω est une fonction $S : P_f(\Omega) \times P_f(\Omega) \to [0,1]$ telle que

$$S(A,B) = F_S(M(A \cap B), M(B - A), M(A - B))$$

où F_S est une fonction $F_S : \Re^{+3} \to [0,1]$ et M une mesure d'ensembles flous sur $P_f(\Omega)$.

Comme les mesures de Tversky, dont elles sont une extension, les mesures de similarité floues couvrent différentes familles de mesures. Dans le cadre de la recherche d'images, nous nous intéressons principalement aux mesures dites *de similitude* qui évaluent la ressemblance entre deux objets. Ces mesures sont définies comme suit :

Définition 2 : une M-mesure de similitude S sur Ω est une M-mesure de comparaison S telle que $F_S(X,Y,Z)$ est :

- monotone non décroissante selon X (les composantes communes)
- monotone non croissante selon Y et Z (les composantes propres)

2.2. Recherche d'images par ressemblance floue

Les moteurs de recherche d'images par le contenu sont basés sur des calculs de similarité. La plupart du temps, on utilise des mesures d'intersection d'histogrammes[3] ou des mesures de distances[4]. A cause de leur caractère général couvrant différents comportements de mesures (telles que les intersections d'histogrammes), nous avons construit un moteur de recherche d'images à partir des mesures de similarité floues. De plus, ce formalisme facilite l'agrégation des différentes valeurs issues des mesures de similarité.
Ce moteur s'appuie sur une représentation régionale des images. Chaque image est segmentée en régions au moyen d'un algorithme de segmentation automatique. Ces régions répondant à certains critères d'homogénéité en couleur et en texture sont grosso modo représentatives des objets présents dans l'image. L'utilisateur peut ensuite choisir parmi les régions disponibles dans la base les objets (régions) qu'il souhaite retrouver.
Notre démarche se décompose de la façon suivante : une mesure de similarité est utilisé comme brique de base pour permettra la comparaison région à région, cette mesure est ensuite utilisée pour construire une requête simple permettant de retrouver un objet dans une image, enfin nous agrégeons des scores issus de multiples mesures de similarité pour effectuer différents types de requêtes basées sur de multiples objets.

La première étape consiste à construire une mesure de similarité région à région. Cette mesure est une simple application de la mesure de similarité floue sur les descripteurs visuels régionaux. Ces descripteurs sont, dans notre approche, au nombre de trois : un histogramme de couleur et un masque de région. Ces trois descripteurs respectifs pouvant être assimilés à des sous-ensembles flous de la palette de couleur et du domaine de l'image[5]. La similarité entre une région requête R_i et une région d'une image de la base I_j est donc calculée en moyennant les similarités entre les descripteurs de R_i et de I_j pour chaque descripteur couleur et masque. Pour H_{Ri} et H_{Ij} les histogrammes de R_i et de I_j, M_{Ri} et M_{Ij} les masques de R_i et de I_j, la similarité $S_{reg}(R_i, I_j)$ est calculée par :

[3] M. SWAIN, D. BALLARD, (1991) *Color indexing*, p.11–32 ,International Journal of Computer Vision, v. 7(1)
[4] M.D. FLICKNER, H. SAWHNEY, W. NIBLACK, et al. (1995), *Query by image and video content: The qbic system*, p. 23-32, Computer, vol 28(9), Septembre.
[5] J.F. OMHOVER, M. DETYNIECKI and B. BOUCHON-MEUNIER, (2004), *A Region-Similarity-Based Image Retrieval System*, Proc. Of IPMU'2004, pp. 1461-1468

$$S_{reg}(R_i, I_j) = \lambda_H .S(H_{Ri}, H_{Ij}) + \lambda_M .S(M_{Ri}, M_{Ij})$$

Où λ_H et λ_M sont deux paramètres qui nous permettent de pondérer l'influence des deux descripteurs au sein de la mesure (habituellement, $\lambda_H = 0.7$ et $\lambda_M = 0.3$) et *S* est une mesure de similarité classique parmi les mesures suivantes :

$$S_{jaccard}(A,B) = \frac{X}{X+Y+Z}$$

$$S_{dice}(A,B) = \frac{2X}{2X+Y+Z}$$

$$S_{dice}(A,B) = \frac{X}{\sqrt{X+Y}\sqrt{X+Z}}$$

avec $X = M(A \cap B), Y = M(B-A)$ et $Z = M(A-B)$.

Nous obtenons doncune mesure $S_{reg}(R_i, I_j)$ qui évalue la ressemblance entre une région R_i de requête et une région I_j d'une image *I* de la base.

Si nous maximisons cette similarité sur chaque région R_i de l'image *I*, nous obtenons la mesure suivante :

$$S_{reg}(R_i, I) = \max_{I_j \in I}(S_{reg}(R_i, I_j))$$

Et nous pouvons remarquer que cette mesure de similarité « région à image » indique la valeur de vérité de la proposition : « il y a une région similaire à R_i dans *I* ». En effet si l'une des régions de I est fortement similaire à R_i, la similarité $S_{reg}(R_i, I)$ sera élevée. A l'inverse, si aucune des régions de I n'est similaire à R_i, la valeur de $S_{reg}(R_i, I)$ sera faible. Cet outil, évalué sur chacune des images de la base, nous permet de retourner à l'utilisateur les images contenant la région R_i de son choix (celles qui maximisent $S_{reg}(R_i, I)$).

Maintenant, nous pouvons agréger différents scores de similarité pour construire des requêtes basées sur de multiples exemples régionaux. En effet, les mesures de similarité apparentées à des valeur de vérité s'agrègent comme telles. Si bien que l'on peut simplement construire les requêtes suivantes en agrégeant par l'opérateur adéquat. Selon que l'on souhaite retrouver plusieurs objets ensembles, ou retrouver l'un ou l'autre de différents aspects d'un même objets, on agrégera les mesures de similarité simples par un opérateur de fusion conjonctif ou disjonctif. (resp. les opérateurs minimum et maximum de logique floue). Nous proposons aussi de filtrer certains éléments jugés négatifs par l'utilisateur en inversant (négation floue) la mesure de similarité associée à cet élément dans l'agrégation. Plus de détails sur ces opérations peuvent être trouvées dans d'autres publications[5].

3. Equivalence entre mesures de similarité

Les systèmes de recherche automatique mettent en œuvre les mesures de similarité pour retrouver les documents qui s'apparentent à une requête. Une liste de résultats est renvoyée à l'utilisateur. Ces résultats sont ordonnés par la valeur de leur ressemblance à la requête : de la plus forte à la plus faible. Mais comme le signale Santini[6], l'information pertinente pour

[6] S. SANTINI, R. JAIN, (1999), *Similarity measures*, IEEE Trans on Pattern Analysis and Machine Intelligence, vol 21(9), Sept.

l'utilisateur est moins la valeur de la similarité d'un document que le rang auquel le document a été placé dans la liste. Bien souvent, l'utilisateur remarque à peine la valeur de similarité, il ne se préoccupe que de consulter les réponses dans leur ordre. Sur la base de cette constatation, le choix entre l'une ou l'autre de deux mesures de similarité pour effectuer cette recherche perd son sens si ces deux mesures aboutissent au même ordre de résultats.
Dans cette partie, nous explorons l'équivalence des mesures de similarité. Nous établissons cette équivalence au moyen de trois définitions : l'équivalence en ordre, l'équivalence par une fonction croissante, et l'équivalence par courbes de niveaux. Nous exploitons ces définitions pour étendre l'équivalence entre mesures simples (utilisées pour des recherches par comparaison à un seul exemple) à l'équivalence entre mesures agrégées (recherches par de multiples exemples).

3.1. Equivalence entre mesures de similarité

Trois définitions peuvent être proposées pour la relation d'équivalence entre mesures de similarité et basées sur la conservation de l'ordre. Nous montrons que ces trois définitions aboutissent aux mêmes classes d'équivalence dans le cadre des mesures de similitude continues.
La première relation est une simple écriture de la conversation de l'ordre entre deux mesures de similitude :

Définition 3 : deux mesures S_a et S_b sont dites *équivalentes en ordre* si et seulement si

$$\forall (X,Y,Z) \in \Re^{+3}, \forall (X',Y',Z') \in \Re^{+3}$$

$$S_a(X,Y,Z) \leq S_a(X',Y',Z') \Leftrightarrow S_b(X,Y,Z) \leq S_b(X,Y,Z)$$

Cette relation est réflexive, symétrique et transitive. Il s'agit donc bien d'une relation d'équivalence et elle aboutit à l'existence de classes d'équivalence entre mesures de similitude.

La deuxième définition établit l'équivalence entre deux mesures si l'une peut être écrite comme une fonction croissante de l'autre :

Définition 4 : deux mesures S_a et S_b sont dites *équivalentes par une fonction* si et seulement si il existe une fonction *f* strictement croissante :

$$f : \begin{cases} \mathrm{Im}(S_a) \rightarrow \mathrm{Im}(S_b) \\ x \mapsto f(x) \end{cases}$$

Avec $\mathrm{Im}(S_{a/b}) = \{\alpha / \exists (X,Y,Z) \in \Re^{+3}, \alpha = S_{a/b}(X,Y,Z)\}$, telle que $S_b = f \circ S_a$.

Par définition, une fonction *f* satisfaisant ces conditions est surjective. Etant strictement croissante, elle est aussi injective. Une telle fonction *f* établit donc une bijection entre $\mathrm{Im}(S_a)$ et $\mathrm{Im}(S_b)$. La relation définie en 4 est donc réflexive, symétrique et transitive. Il s'agit bien d'une relation d'équivalence entre mesures de similarité.
Si les définitions 3 et 4 semblent différentes, il est cependant trivial de montrer qu'elles aboutissent à la même relation d'équivalence. Si deux mesures sont équivalentes en ordre, on peut en effet construire simplement une fonction croissante qui les lie par composition, et inversement.
Il faut ici remarquer que l'équivalence entre deux mesures est liée à l'existence d'une fonction bijective entre leurs ensembles de valeurs. Si les deux ensembles de valeurs prises par deux

mesures S_a et S_b ne peuvent mis en bijection ces deux mesures ne peuvent être équivalentes. Cela peut être le cas si l'une des deux mesures a un ensemble de valeur discret (par exemple $\{0,0.2,0.4,0.6,0.8,1\}$) et l'autre continu ($[0,1]$).

A partir de cette définition, nous pouvons montrer simplement que les mesures de Jaccard et de Dice appartiennent à la même classe d'équivalence, tandis qu'elles ne sont pas équivalentes à la mesure d'Ochiai (qui appartient à une classe différente). La figure 1 illustre le fait que les mesures de Jaccard et de Dice peuvent s'écrire comme une fonction l'une de l'autre, tandis qu'il est impossible d'écrire la mesure d'Ochiai comme une fonction à une seule valeur de la mesure de Jaccard.

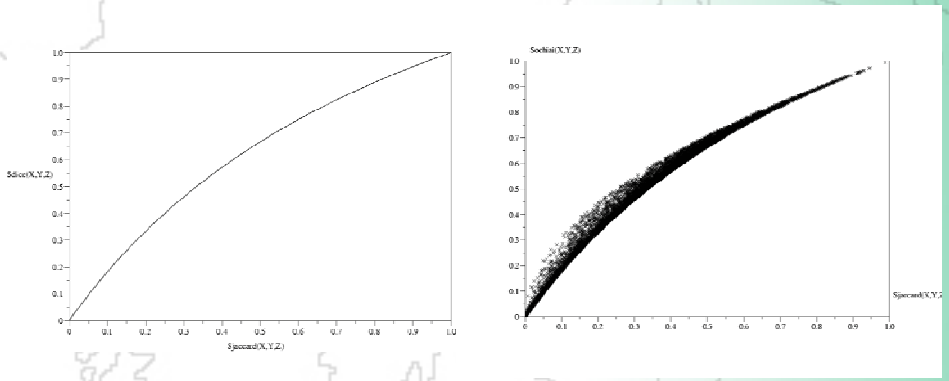

Figure 1 : tracé des couples de valeurs prises par les mesures (a) de Jaccard et de Dice, (b) de Jaccard et d'Ochiai.

En effet, comme nous pouvons le constater en figure 1, à une valeur de la mesure de Dice correspond une seule valeur de la mesure de Jaccard, et inversement de façon univoque. Cependant, une même valeur prise par la mesure de Jaccard correspond à plusieurs valeurs prises par la mesure d'Ochiai.

La troisième définition concerne l'équivalence en courbes de niveaux. Nous écrirons S_l^λ la courbe d'une mesure S_l au niveau λ :

$$S_l^\lambda = \{(X,Y,Z)/S_l(X,Y,Z) = \lambda\}$$

Définition 5 : deux mesures S_a et S_b sont dites *équivalentes en courbes de niveaux* si et seulement si :

$$\forall \beta \in \mathrm{Im}(S_b), \exists! \alpha \in \mathrm{Im}(S_a)\, t.q.\, S_a^\alpha = S_b^\beta$$

Cette relation est évidemment réflexive, et il peut être montré facilement qu'elle est aussi symétrique et transitive. L'équivalence de cette troisième définition ne peut cependant être établie que dans le cadre des mesures de similarité continues, où nous pouvons montrer le théorème suivant :

Théorème 1 : quelque soient deux mesures de similitude continues S_a et S_b, s'il existe une fonction :

$$f : \begin{cases} \mathrm{Im}(S_a) \to \mathrm{Im}(S_b) \\ x \mapsto f(x) \end{cases}$$

telle que $S_b = f \circ S_a$, alors f est monotone non décroissante.

Si deux mesures de similitude continues S_a et S_b sont équivalentes en courbes de niveaux, une fonction qui lie S_a à S_b par composition peut-être construire. D'après le théorème ci-

dessus, cette fonction est monotone non décroissante. Or grâce à la symétrie de la relation 5, il nous montrons que cette fonction est en fait strictement croissante. A l'inverse, si deux mesures S_a et S_b sont équivalentes par une fonction, nous montrons simplement que ces deux mesures sont équivalentes en courbes de niveaux. Nous obtenons ainsi l'équivalence entre les relations 3, 4 et 5 dans le cas des mesures continues.

3.2. Equivalence entre mesures de similarité agrégées

La triple définition de l'équivalence entre mesures de similarité proposée dans la section précédente établit l'équivalence entre une simple paire de mesures (et par extension, l'existence de classes d'équivalence entre mesures de similarité). Mais les systèmes de recherche d'images n'utilisent que rarement les mesures pour une seule comparaison. On effectue plutôt une agrégation de différents résultats de similarité : soit que l'on compare deux objets dans de multiples espaces de descriptions (plusieurs ensembles de caractéristiques sont attachés à un même objet), soit que l'on compare un même objet à plusieurs autres (dans le cas des requêtes par de multiples exemples). La question est ici de déterminer si l'équivalence entre deux mesures se conserve quand chacune de ces mesures est utilisée dans un même schéma d'agrégation.
Pour généraliser les cas de figure liés à l'agrégation de mesures de similarité, nous écrirons l'agrégation par un opérateur *Agg* de chacune des mesures S_a et S_b de la façon suivante :

$$S_a^{Agg,n}(O_1,...,O_n) = Agg(S_a(O_1),...,S_a(O_n))$$
$$S_b^{Agg,n}(O_1,...,O_n) = Agg(S_b(O_1),...,S_b(O_n))$$

Il s'agit du cas courant, et que nous utilisons nous-mêmes, où une mesure de similarité est utilisée pour différentes comparaisons (dans de multiples dimensions, ou entre différents objets) et agrégée au moyen d'un opérateur, chacun des O_i correspondant à un triplet :

$$(X_i, Y_i, Z_i) = (M(A_i \cap B_i), M(A_i - B_i), M(B_i - A_i))$$

Quand une mesure est utilisée pour comparer une paire d'objets dans différents espaces de représentation, les A_i et B_i correspondent aux différents ensembles de caractéristiques pris par les deux objets *a* et *b* à comparer dans chacun des espaces. Quand une mesure est utilisée pour comparer un même objet *O* à différents exemples E_i (comme pour comparer une image à de multiples images de requêtes), les A_i sont en fait une seule et même description de l'objet *O*, et les B_i sont les différents descripteurs de chacun des exemples E_i.
En restant dans le cadre général proposé ci-dessus, nous définissons l'équivalence entre mesures agrégées de la façon suivante :

Définition 6 : pour *n* entier donné, et un opérateur d'agrégation *Agg* donné, pour n'importe quel couple de mesures S_a et S_b, $S_a^{Agg,n}$ et $S_b^{Agg,n}$ sont dites *équivalentes en ordre* si et seulement si :

$$\forall(O_1,...,O_n), \forall(O'_1,...,O'_n) \; avec \; \forall i, O_i = (X_i, Y_i, Z_i), O'_i = (X'_i, Y'_i, Z'_i)$$

$$S_a^{Agg,n}(O_1,...,O_n) \le S_a^{Agg,n}(O'_1,...,O'_n) \Leftrightarrow S_b^{Agg,n}(O_1,...,O_n) \le S_b^{Agg,n}(O'_1,...,O'_n)$$

Aussi simplement que la relation définit en 3, la relation définie ici est réflexive, symétrique et transitive. Elle définit donc une relation d'équivalence entre mesures agrégées *n* fois.

De plus, nous avons montre le théorème suivant établissant un rapport nécessaire de conservation des courbes de niveaux de l'opérateur d'agrégation pour préserver l'équivalence des mesures S_a et S_b au niveau des mesures agrégées $S_a^{Agg,n}$ et $S_b^{Agg,n}$.

Théorème 2 : Etant données deux mesures S_a et S_b équivalentes, avec f une fonction strictement croissante telle que $S_b = f \circ S_a$, si $S_a^{Agg,n}$ et $S_b^{Agg,n}$ sont équivalentes en ordre, alors la fonction :

$$F_n : \begin{cases} \mathrm{Im}(S_a)^n \to \mathrm{Im}(S_b)^n \\ (x_1,...,x_n) \mapsto (f(x_1),...,f(x_n)) \end{cases}$$

est telle que :

$$\forall \alpha \in \mathrm{Im}(S_a^{Agg,n}), \exists! \beta \in \mathrm{Im}(S_b^{Agg,n}) t.q.\ F_n\left\langle Agg_n^{\alpha} \cap \mathrm{Im}(S_a)^n \right\rangle = Agg_n^{\beta} \cap \mathrm{Im}(S_b)^n$$

Cela signifie que la transformation F_n fait correspondre chacune des courbes de niveaux de l'opérateur *Agg* à une autre courbe de niveau de ce même opérateur, tout en restant dans les espaces de variations $\mathrm{Im}(S_a^{Agg,n})$ et $\mathrm{Im}(S_b^{Agg,n})$ des mesures agrégées. Ce théorème montre une relation entre la forme des courbes de niveaux de l'opérateur *Agg* et la fonction *f* utilisée pour transformer S_a en S_b. Il signifie aussi qu'étant donné un opérateur d'agrégation, certaines seulement des équivalentes constatées entre des mesures de similarité S_a et S_b pourront être observées entre les mesures agrégées $S_a^{Agg,n}$ et $S_b^{Agg,n}$.

En pratique, ceci nous permet de montrer que l'agrégation par l'opérateur maximum (ou minimum) préserve l'équivalence de n'importe quel couple de mesures S_a et S_b au niveau agrégé (et ce quelque soit la fonction *f* qui lie les deux par composition). Nous avons aussi montré que l'agrégation par une moyenne brise toutes les équivalences qui n'ont pas la forme triviale d'une transformation linéaire. Sur la figure 2, nous pouvons voir qu'une simple transformation de la forme $f(x) = \sqrt{x}$ préserve la forme des courbes de niveaux de l'opérateur maximum, de sorte qu'il est possible d'identifier les courbes de l'opérateur à celles résultant de la transformation. Il est cependant impossible d'effectuer une telle identification pour l'opérateur moyenne : la transformation brise la forme de ses courbes de niveaux.

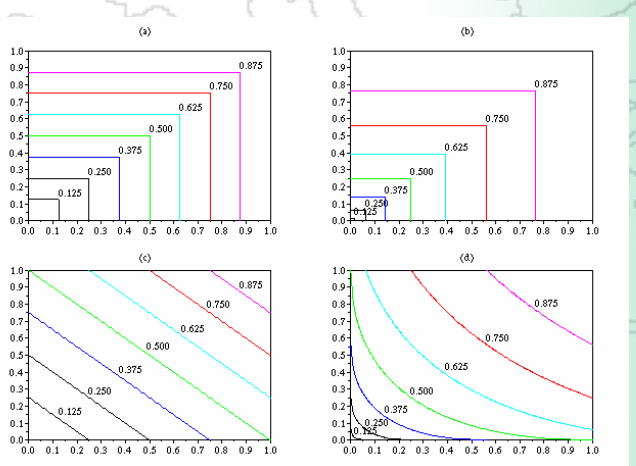

Figure 2 : courbes de niveaux de (a) $Max(x_1, x_2)$, **(b)** $Max(\sqrt{x_1}, \sqrt{x_2})$, **(c)** $Avg(x_1, x_2)$, **(d)** $Avg(\sqrt{x_1}, \sqrt{x_2})$

4. Equivalence des mesures du *Ratio Model* de Tversky

En guise d'application de la théorie exposée précédemment, nous caractérisons ici les classes de mesures équivalentes dans la famille des mesures du *Ratio Model* de Tversky.
Comme introduit en section 2.1, Tversky a proposé une expression générale des mesures de similarité, le *ratio model* formulé comme suit :

$$S_{(\alpha,\beta)}(X,Y,Z) = \frac{X}{X + \alpha.Y + \beta.Z}$$

Cette formulation fournit deux paramètres a et b. Le choix de ces paramètres conduit la mesure à suivre un certain comportement, par exemple :

- si $\alpha = \beta$, la mesure est symétrique. Les mesures de Jaccard et de Dice sont deux exemples de cette sorte ($S_{jaccard} = S_{(1,1)}, S_{dice} = S_{(1/2,1/2)}$).
- pour α quelconque, si $\beta = 0$, la mesure est appelée mesure d'inclusion, elle évalue le degré d'inclusion d'un ensemble *A* dans un ensemble *B*.
- pour β quelconque, si $\alpha = 0$ la mesure et appelée mesure de satisfiabilité, elle évalue par exemple la satisfaction d'une observation *B* à la prémisse *A* d'une règle (utilisé en raisonnement flou).

Comme nous l'avons montré en section 3.3, l'équivalence entre deux mesures continues (comme le sont les mesures du *ratio model*) peut être montrée en étudiant leurs courbes de niveaux. Pour montrer l'équivalence entre deux mesures $S_{(\alpha,\beta)}$ et $S_{(\alpha',\beta')}$, nous devons montrer que pour n'importe quel niveau *h'* pris par $S_{(\alpha',\beta')}$, il existe un unique niveau *h* pris par $S_{(\alpha,\beta)}$ tel que les courbes de niveaux correspondantes aux niveaux *h'* et *h* soient égales. Nous avons montré que cela ne peut arriver que lorsque nous avons l'égalité : $\alpha.\beta' - \alpha'.\beta = 0$.
Ainsi, deux mesures de Tversky $S_{(\alpha,\beta)}$ et $S_{(\alpha',\beta')}$ sont équivalentes si et seulement si $\alpha.\beta' = \alpha'.\beta$, autrement dit si leurs paramètres suivent le même rapport $\frac{\alpha}{\beta}$. En conséquence, si nous ne nous intéressons qu'à l'ordre des résultats induits par une mesure de similarité du *ratio model*, le choix des paramètres α et β se réduit au choix d'un seul paramètre $k = \frac{\alpha}{\beta}$:

- $k = 0$ pour une mesure d'inclusion
- $k = 1$ pour une mesure de ressemblance (symétrique)
- $k = +\infty$ pour une mesure de satisfiabilité.

5. Conséquences et conclusions

Nous étudions ici les conséquences pour la recherche de documents de l'utilisation de mesures prises dans une même classe d'équivalence, autrement dit, les implications de l'invariance en ordre pour certaines applications de recherche d'informations.

La première conséquence concerne le domaine de la recherche de documents où des mesures de similarité sont utilisées pour comparer des caractéristiques extraites des documents. Utiliser des mesures issues d'une même classe d'équivalence conduira le système à fournir, pour chacune de ces mesures, les mêmes résultats à l'utilisateur. Il n'y aura pour lui aucune différence.
Une conséquence directe de cette constatation est plus profonde et concerne l'évaluation des résultats. Pour deux mesures équivalentes, les taux de rappel et de précision servant habituellement à comparer l'efficacité de deux mesures seront exactement identiques. Ces

taux sont en effet calculés à partir de l'ordre des résultats pertinents. Si cet ordre est invariant entre deux mesures, elles seront aussi efficaces l'une que l'autre.
De même, dans une procédure d'appariement entre paires d'objets basée sur la ressemblance maximale (où un objet d'un ensemble A est apparié avec celui qui lui ressemble le plus au sein d'un ensemble d'objets B, comme par exemple entre deux graphes), les correspondances seront établies de la même façon selon que l'on utilise l'une ou l'autre de deux mesures équivalentes.

Dans la plupart des applications, comme dans notre moteur de recherche d'images, les valeurs de similarité calculées entre paires de descripteurs ne sont pas utilisées seules mais agrégées. Dans notre système, il s'agit de permettre à l'utilisateur d'effectuer des recherches basées sur de multiples exemples. Il peut s'agir aussi de comparer des documents selon de multiples échelles ou dans de multiples espaces de représentation. Comme nous l'avons montré dans cet article, l'équivalence entre mesures peut, dans certains cas, être conservée après fusion par un opérateur d'agrégation. Ce n'est cependant pas toujours le cas, comme nous l'avons montré pour l'opérateur moyenne qui est couramment utilisé dans le domaine.

De plus, le pouvoir de discrimination[7] des mesures présentées en section 2 peut varier. Ainsi, dans des applications où l'ordre des valeurs issues de mesures de similarité est combiné avec ces valeurs elles-mêmes, deux mesures équivalentes peuvent obtenir des résultats différents. Le résultat d'invariance n'est en effet observable que dans le cas où seul l'ordre détermine le résultat. Dans le domaine de la recherche automatique de documents, l'ordre est la plupart du temps le seul facteur déterminant, mais il peut arriver par exemple qu'un système tronque la liste de résultat selon la force souhaitée de la ressemblance avec la requête. Dans ce cas, les listes peuvent être plus ou moins longue selon que l'on utilisera l'une ou l'autre de deux mesures équivalentes.

Pour ce qui est de la recherche par similarité, et tant que l'ordre seul compte comme résultat d'une requête, le choix d'un mesure de similarité peut être réduit au choix d'une classe d'équivalence de mesures, les mesures équivalentes étant nécessairement conduites à obtenir les mêmes résultats. Dans cet article nous avons de plus montré que les mesures agrégées pouvaient elles aussi être équivalentes sous certaines conditions liant l'opérateur de fusion à la fonction permettant de lier deux mesures par composition. Comme application de notre étude, nous avons entièrement caractérisé les classes d'équivalence parmi les mesures du *Ratio Model* de Tverky.

[7] M. RIFQI, V. BERGER, B. BOUCHON-MEUNIER, (2000), *Discrimination power of measures of comparison*, p. 189–196, Fuzzy Sets and Systems, vol. 110(2), Mars