Ordre, agrégation et répétition : des paramètres fondamentaux dans les comparaisons d'objets informationnels

Christine Michel

ORDRE, AGREGATION ET REPETITION : DES PARAMETRES FONDAMENTAUX DANS LES COMPARAISONS D'OBJETS INFORMATIONNELS

Michel Christine

CEM-GRESIC
MSHA - Université Bx3
10, Esplanade des Antilles
33607 PESSAC Cedex
Tel : 05 56 84 68 13
Christine.Michel@montaigne.u-bordeaux.fr

RESUME :
On désigne généralement les objets informationnels comme les supports, les médiateurs permettant d'améliorer la perception dans un contexte de travail avec des ordinateurs (Auziol, 2001). Concrètement ces objets peuvent être de simples textes, des images fixes ou animées produites manuellement par un ou plusieurs auteurs; ou bien des constructions dynamiques, réponses de systèmes, générées par le besoin informationnel ou la définition du profil particulier d'une personne. Ces objets informationnels sont généralement des agrégats d'éléments indexés et stockés dans des bases de données. Les stratégies de construction varient en fonction des objectifs spécifiques du système, nous ne les détaillerons pas car notre propos n'est pas ici de faire une typologie exhaustive de ces objets. Nous nous contenterons de les regarder au travers de trois caractéristiques : *l'agrégation, l'ordonnancement et la répétition (ou l'unicité) des éléments.* Ces trois caractéristiques sont très souvent prises en compte par les ergonomes et développeurs de systèmes pour en améliorer l'utilisation, la pertinence et l'efficacité. Paradoxalement, parce qu'aucun formalisme n'a été défini, ils sont rarement pris en compte pour évaluer ou comparer les objets informationnels une fois construit. Nous proposons dans un premier temps de présenter les concepts de semi ordre (de classe et d'éléments) et de disjonction, nécessaires pour définir les différents types d'objets informationnels. Dans un second temps nous présentons une démarche utilisée pour résoudre complètement le problème de la comparaison d'objets de type semi-ordonné disjoint de classe. Pour conclure nous ouvrons sur les perspectives de travaux à mener dans les autres contextes et pour les autres objets informationnels.

ABSTRACT :
Informational objects are support or mediator used to improve perception of information when people are working with computer. They must be simple texts, combined or not with fixed or animated picture. They must be produced manually by one or many authors, or dynamically by information retrieval systems like answers are. These objects must be viewed as aggregates of indexed elements, generally called fragments, stocked in databases. Strategies of constructions are varying from systems to systems so objects are very different from each others but we can describe formerly them by three characteristics; there degree of aggregation, scheduling and repetition of fragments. Developers and ergonomicists very often use these three characteristics to improve use, relevance and efficacy of systems. Paradoxically, because any formalism as been defined, they are not taken into account to evaluate or compare results constructed. In this paper we propose to present a mathematical formalism based on the concept of semi-order and possible to use in the case of IR results. Then we present how it can be write with the fuzzy set theory and used to construct similarity measure taken into account the presence and decrease rank of documents. To finish we present other contexts of use and research perspectives.

MOTS CLES : mesure de similarité, ordre, ordonnancement, agrégation, recherche d'information, évaluation,

KEY WORDS : similarity measures, order, scheduling, aggregation, information retrieval, evaluation

1- SEMI-ORDRE ET DISJONCTION

Les réponses proposées par les systèmes de recherche d'information peuvent se présenter a l'extrême sous la forme d'éléments non ordonnés (figure 1) ou d'éléments totalement ordonnés (figure 2).

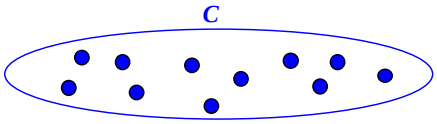

Figure 1

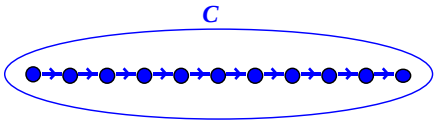

Figure 2

La recherche de nouveaux modes de présentation et de nouvelles interfaces a fait apparaître d'autres formes de présentation comme les sous-listes thématique ou les amas cartographiques, caractérisés globalement par des regroupements de réponses dans des classes construites selon une unité sémantique. L'ordre, s'il y a, n'est généralement pas total, nous appellerons cette forme hybride **le semi-ordre**. Deux cas de figure peuvent se présenter :

- Les classes sont ordonnées les unes par rapport aux autres mais les éléments à l'intérieur des classes ne le sont pas (figure 3), on parlera de *semi ordre de classe*.

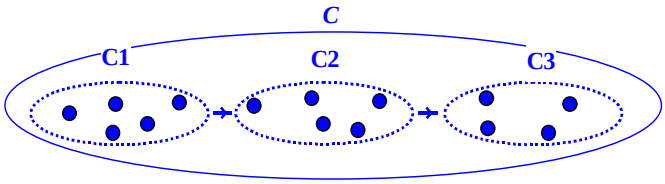

Figure 3

C'est typiquement le cas de figure qui se présente dans un système comme Spirit (Fluhr, 97). Les documents sont regroupés dans des classes selon qu'ils contiennent ou non une combinaison des mots informationnels de la question, les documents ont la même importance dans la classe car ils contiennent tous la même combinaison de mots, les classes par contre ont plus ou moins d'importance selon qu'elles sont caractérisées par plus ou moins de mots informationnels.

- Les classes ne sont pas ordonnées les unes par rapport aux autres mais les éléments à l'intérieur des classes le sont (figure 4), on parlera de *semi ordre d'éléments*.

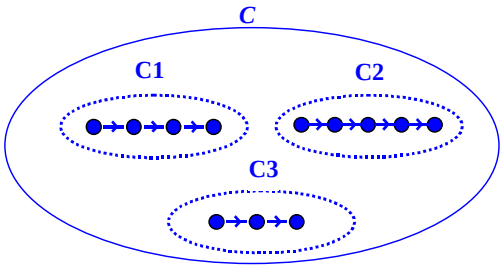

Figure 4

C'est typiquement le cas de figure qui se présente dans un système cartographique de type réseau sémantique, les classes sont formées, agrégeant les documents selon leur degré de proximité, les documents sont ensuite ordonnés dans les classes, le plus représentatif de la classe étant présenté en premier, ou bien les système qui regroupe les documents selon des méta critères comme un type de support éditorial (résumé, article de vulgarisation) ou une source (Site Web commercial, pages personnels, magazine, …) différent du contenu sémantique (Zamir, 99).

Dans tous les cas de figure les éléments ne sont présentés qu'une fois, dans le cas des figures 3 et 4 on dira que le semi ordre est **disjoint**. Cette caractéristique est propre aux réponses de systèmes en recherche d'information ; dans bon nombre d'autres contextes, les objets informationnels sont construits à partir d'éléments qui peuvent se répéter : un texte est composé de mots qui peuvent se répéter, une page Web dynamique est composé d'éléments, image ou lien qui, pour des raisons d'ergonomie, peuvent se répéter …Nous ne traiterons pas ces cas de figure et nous restreignons au contexte où les éléments ne sont présentés qu'une fois. Il est souvent nécessaire de pouvoir comparer de tels objets, par exemple il est nécessaire de pouvoir comparer les réponses de plusieurs systèmes lorsque l'on cherche à les évaluer. Comment le faire au mieux?

2- CAS D'ENSEMBLES NON ORDONNES

2-1- Les mesures de similarité fortes et faibles

La similitude entre les réponses proposées par les système se base généralement sur le nombre d'éléments qu'elles peuvent avoir en commun, calculé, si A et B sont les ensembles à comparer, par le cardinal de $A \cap B$ (noté $|A \cap B|$). La similitude est quantifiée par un nombre compris entre 0 et 1, 0 signifiant que les ensembles A et B comparés n'ont aucun éléments en commun c'est à dire que $A \cap B = \varnothing$ et 1 signifiant classiquement qu'ils sont strictement identiques c'est à dire $A = B$. En fait cette condition se révèle fausse lorsque des rôles spécifiques sont attribués à A et B comme c'est le cas pour le rappel et la précision. En effet, le rappel (R) est la proportion de documents pertinents trouvés par rapport au nombre de documents pertinents. Considérons que A est l'ensemble des documents retrouvés et B l'ensemble des documents pertinents alors le Rappel R s'écrira :

$$R(A,B) = \frac{|A \cap B|}{|B|}$$ (Grossman, 1998) **(Équation 1)**

Un rappel R=1 signifie que tous les documents pertinents sont retrouvés c'est à dire que $B \subset A$ et non pas que *seuls* les documents pertinents sont retrouvés c'est à dire que $B = A$.
De la même manière, une précision P=1 signifie que tous les documents retrouvés sont pertinents c'est à dire $A \subset B$ et non pas que seuls les documents retrouvés sont pertinents.

Cette observation a permis de séparer les mesures de similarité en deux groupes : **les mesures fortes et les mesures faibles** formellement définies dans (Egghe, 2002). Très simplement, **les mesures fortes sont caractérisées par une stricte identité des ensembles comparés si la proximité est de 1, les mesures faibles le sont par une inclusion des ensembles dans pareil cas**.

2-2- Principales mesures de similarité fortes

Le coefficient de Jaccard : $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$ **(Équation 2)**

Le coefficient de Dice : $Dice(A,B) = \frac{2 \times |A \cap B|}{|A| + |B|}$ **(Équation 3)**

Le cosinus : $Cos(A,B) = \frac{|A \cap B|}{\sqrt{|A| \times |B|}}$ **(Équation 4)**

La mesure N $N(A,B) = \sqrt{2}\frac{|A \cap B|}{\sqrt{|A|^2 + |B|^2}}$ **(Équation 5)**

Le coefficient de débordement 2 (overlap 2) : $O_2(A,B) = \frac{|A \cap B|}{\max(|A|,|B|)}$. **(Équation 6)**

Les mesures d'efficacité : construites comme combinaison ou moyenne du rappel R et de la précision P et dont la plus générale est :

Le coefficient de Dice généralisé[1] construit par Van Rijsbergen[2] : $E_\alpha = 1 - \frac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}}$ **(Équation 7)**

2-3- Principales mesures de similarité faible

---

[1] α étant le poids relatif assigné à la valeur de précision par l'utilisateur. α est une valeur comprise entre 0 et 1. Par hypothèse, le poids accordé au rappel est le complémentaire de celui accordé à la précision.

[2] VAN RIJSBERGEN C. J. - Retrieval effectiveness - Sparck Lones K ed Information retrieval experiments ; London : Butterworth - 1981 - pp32-43.

Le coefficient de débordement 1 (overlap1) : $O_1(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)}$. **(Équation 8)**

Le rappel : $R(A,B) = \frac{|A \cap B|}{|B|}$ **(Équation 9)**

La précision: $P(A,B) = \frac{|A \cap B|}{|A|}$ **(Équation 10)**

Leurs mesures dérivées :

Le bruit (Noise) : $Noise = 1 - P$ **(Équation 11)**

Le facteur d'omission (O) ou le silence : $O = 1 - R$ **(Équation 12)**

Dans ces quatre derniers cas, A est l'ensemble des documents retrouvés et B l'ensemble des documents pertinents.

3- CAS D'ENSEMBLES ORDONNES

L'arrivée de systèmes ordonnant totalement ou partiellement les documents selon une valeur de pertinence a rendu nécessaire la construction de mesures de similarités reflétant ce paramètre. On trouve dans la littérature deux manière de le faire :

- en calculant le coefficient de similarité sur les 10, 20 ou 30 premiers éléments considérés comme des ensembles sans ordre. C'est la solution choisie dans le cadre de TREC ou l'on retrouve des adaptations du Rappel et de la Précision à des listes ordonnées. Par exemple P(10) est la précision donnée par les 10 premiers documents, R-Prec est la précision donnée par les R premiers documents, R étant le nombre de document pertinents pour le sujet donné. On retrouvera plus de 80 mesures du même type et leur analyse comparative dans (Voorhees 98).
- en pondérant la mesure de similarité de deux ensembles par le coefficient de corrélation de rang R représentatif du nombre de permutations à effectuer pour replacer deux les éléments communs à deux ensembles dans le même ordre. Une expérimentation de cette combinaison est présentée dans (Tague-Sutcliffe, J, 1995).

Ces deux types de construction ne sont pas assez précises ; la première car elle ne prend en compte qu'une fraction de la réponse, la seconde car elle ne prend pas en compte l'intérêt de retrouver des éléments communs dans les premiers rangs plutôt que dans les derniers. De plus, ces deux mesures ne sont pas applicables dans le cas d'un semi-ordre. Notre propos a donc été de chercher comment construire des mesures de similarité prenant en compte le semi-ordre, avantageant les éléments présentés tôt et ayant un lien avec les mesures originelles de similarité. Notre travail initial (Michel, 1999), (Michel, 2000) (Michel, 2001) a été amélioré par une collaboration avec Leo Egghe, les résultats publié dans (Egghe, 2002) et (Egghe, 2003) permettent de construire des mesures de proximités ordonnées valides sur des ensembles en semi-ordre disjoint de classe. Bien entendu, **l'ordonnancement total et le non-ordonnancement n'étant que des cas particuliers du semi-ordre de classe, ces mesures sont applicables aussi dans ces contextes**. Nous nous proposons ici de les présenter brièvement.

3-1- Formalisation du problème pour le semi-ordre de classe

Considérons que deux ensembles C et C' à comparer sont construits selon un *semi-ordre de classes disjointes* comme le représente la figure 3. Leurs classes respectives seront notées C_i et C'_j, i variant de 1 à *m*, j variant de 1 à *m'*. L'idée initiale (Michel, 1999), (Michel, 2000) a été de construire des mesures de similarité en combinant une mesure de proximité classique calculée sur chaque couple de classes Ci et C'j puis pondérée par un coefficient représentatif des rangs i et j de chacune. Ainsi, une mesure de proximité ordonnée Q peut être construite à partir de *toute mesure de proximité classique de type faible et forte* (noté D) de la manière suivante :

$$Q(C,C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} D(C_i, C'_j) \times \varphi(i,j) \quad \textbf{(Équation 13)}$$

$\varphi(i,i)$ étant une fonction vérifiant 5 conditions définies (pour les détails on se reportera à (Michel, 2001).

Nous avons proposé 3 mesures concrètes : une basée sur le Jaccard, une sur le rappel et une sur la précision. Dans une expérimentation nous avons comparé une mesure de Jaccard classique avec la mesure de Jaccard

ordonnée. Les résultats ont montré la plus grande précision et donc l'intérêt d'utiliser une mesure de proximité ordonnée.

3-2- Les mesures de proximité ordonnées à pondération : une deuxième méthode de construction basée sur la pondération

Leo Egghe intéressé par cette problématique y a contribué en proposant une autre méthode de construction qui consiste à appliquer une fonction f sur la mesure puis à la pondérer par une fonction φ de la manière suivante[3] :

$$Q(C,C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} f\left(D_{strong}\ (C_i, C'_j)\right) \times \varphi(i,j) \qquad \textbf{(Équation 14)}$$

L'avantage de cette méthode de construction par rapport à la précédente porte principalement sur l'étendue du nombre de mesures qu'il est possible de construire. En effet, l'une des problématiques à résoudre dans l'équation 13 consiste à trouver une fonction de pondération φ respectant la normalisation de la mesure Q entre 0 et 1. Dans le cas de l'équation 14, les contraintes posées sur la fonction de pondération sont moins fortes du fait de l'application de la fonction f. Les démonstrations et définitions sont publiées dans (Egghe, 2002), il faut préciser que cette méthode n'est applicable *qu'aux mesures de type fort*.

Nous présentons les 6 mesures (Jaccard (J^{Ω}), Dice (E^{Ω}), Cosinus (Cos^{Ω}), Dice Généralisé ($E_{\alpha}{}^{\Omega}$), la mesure N (N^{Ω}), l'overlap de type 2 ($O^{2\Omega}$), construites à partir des indicateurs de type fort présentés précédemment (équation 2 à 6). On remarquera que les mesures de proximité ordonnées basées sur le Jaccard et le Dice sont identiques.

Jaccard et Dice ordonné à pondération :

$$J^{\Omega}(C,C') = E^{\Omega} = \sum_{i=1}^{m} \sum_{j=1}^{m'} J(C_i, C'_j) \times \varphi(i,j) \qquad \textbf{(Équation 15)}$$

Cosinus ordonné à pondération :

$$Cos^{\Omega}(C,C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} \frac{\left|C_i \cap C'_j\right|^2}{|C_i| . \left|C'_j\right|} \times \varphi(i,j) \qquad \textbf{(Équation 16)}$$

La mesure N ordonnée à pondération :

$$N^{\Omega}(C,C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} \frac{\left|C_i \cap C'_j\right|}{\sqrt{|C_i|^2 + \left|C'_j\right|^2 - \left|C_i \cap C'_j\right|^2}} \times \varphi(i,j) \qquad \textbf{(Équation 17)}$$

Le coefficient de débordement O_2 (overlap 2) ordonné à pondération :

$$O^{2\Omega}(C,C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} \frac{\left|C_i \cap C'_j\right|}{\max\left(|C_i|, \left|C'_j\right|\right)} \times \varphi(i,j) \qquad \textbf{(Équation 18)}$$

Le coefficient de Dice généralisé ordonné à pondération :

$$E_{\alpha}{}^{\Omega}(C,C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} \frac{\alpha\left|C_i \cap C'_j\right|}{\alpha|C_i| + (1-\alpha)\left|C'_j \setminus C_i\right|} \times \varphi(i,j) \quad si\ 0 < \alpha < \frac{1}{2} \qquad \textbf{(Équation 19)}$$

$$E_{\alpha}{}^{\Omega}(C,C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} \frac{(1-\alpha)\left|C_i \cap C'_j\right|}{\alpha\left|C_i \setminus C'_j\right| + (1-\alpha)\left|C'_j\right|} \times \varphi(i,j) \quad si\ \frac{1}{2} < \alpha < 1 \qquad \textbf{(Équation 20)}$$

[3] f et φ ne sont pas fixées mais définies comme devant respecter un certain nombre de conditions (IPM2002).

Nous avons fait une comparaison expérimentale de 6 mesures : Cos, J, Cos_l^{Ω}, Cos_p^{Ω}, J_l^{Ω} et J_p^{Ω}. Cos_l^{Ω} et J_l^{Ω} étant construites avec une fonction φ de type linéaire, Cos_p^{Ω} et J_p^{Ω} étant construites avec une fonction φ de type puissance. Nous avons pu observer que :

- la fonction de pondération atténue considérablement l'effet particulier de chaque mesure; ainsi $Cos_p^{\Omega} \approx J_p^{\Omega}$ et $Cos_l^{\Omega} \approx J_l^{\Omega}$.
- les fonctions de pondération agissent avec plus ou moins d'influence en fonction des contextes; la pondération puissance est plus précise quand les ensembles à comparer sont très différents, à l'inverse lorsque les ensembles sont très similaires la fonction linéaire est plus précise.

3-3- Les mesures structurelles de proximité ordonnées : Construction à l'aide de la théorie de la logique floue

Le problème de construction de mesures ordonnées de type faible et fort a été résolu par Leo Egghe (Egghe, 2003) en utilisant le **formalisme des ensembles flous** (Zadeh, 1979). Le principe consiste à reformuler la définition de l'ensemble semi-ordonné C grâce à la fonction d'appartenance de la logique floue (équation 21), ensuite de réécrire les intersections et unions ensemblistes grâce à ce formalisme puis de les appliquer sur les formules classiques des mesures.

Les ensembles sont définis selon le formalisme suivant : $U_C = \bigcup_{i=1}^{n} C_i$ est un ensemble équipé de la fonction d'appartenance $P_{U_C} = \varphi(i) \Leftrightarrow x \in C_i$ où φ est une fonction strictement décroissante **(Équation 21).**

En utilisant la fonction $\varphi(i) = \frac{1}{2^{i-1}}$, les quatre résultats suivants sont démontrés (IPM2003):

$$\left|U_C \cap U_{C'}\right| = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\left|C_i \cap C_j'\right|\frac{1}{2^{\max(i,j)-1}} \quad \textbf{(Équation 22)}$$

$$\left|U_C\right| = \sum_{i=1}^{\infty}\left|C_i\right|\frac{1}{2^{i-1}} \quad \textbf{(Équation 23)}$$

$$\left|U_{C'}\right| = \sum_{j=1}^{\infty}\left|C_j'\right|\frac{1}{2^{j-1}} \quad \textbf{(Équation 24)}$$

$$\left|U_C \cup U_{C'}\right| = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\left|C_i \cap C_j'\right|\frac{1}{2^{i-1}} + \sum_{i=1}^{\infty}\sum_{j=1}^{j-1}\left|C_i \cap C_j'\right|\frac{1}{2^{j-1}} + \sum_{i=1}^{\infty}\left|C_i \setminus \bigcup_{j=1}^{\infty} C_j'\right|\frac{1}{2^{i-1}} + \sum_{j=1}^{\infty}\left|C_j' \setminus \bigcup_{i=1}^{\infty} C_i\right|\frac{1}{2^{j-1}} \quad \textbf{(Équation 25)}$$

En réécrivant les mesures de similarité avec l'intersection, l'union et le cardinal formulé en logique flou il est possible de construire nombre de nouvelles mesures de proximité ordonnées. Il est très intéressant de noter qu'elles peuvent se construire **à partir de mesures de type faible et fort** comme nous pouvons le voir ci dessous.

Jaccard ordonné :

$$J_F(C,C') = \frac{\left|U_C \cap U_{C'}\right|}{\left|U_C \cup U_{C'}\right|} = \frac{\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\left|C_i \cap C_j'\right|\frac{1}{2^{\max(i,j)-1}}}{\alpha} \quad \textbf{(Équation 26)}$$

avec

$$\alpha = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\left|C_i \cap C_j'\right|\frac{1}{2^{i-1}} + \sum_{i=1}^{\infty}\sum_{j=1}^{j-1}\left|C_i \cap C_j'\right|\frac{1}{2^{j-1}} + \sum_{i=1}^{\infty}\left|C_i \setminus \bigcup_{j=1}^{\infty} C_j'\right|\frac{1}{2^{i-1}} + \sum_{j=1}^{\infty}\left|C_j' \setminus \bigcup_{i=1}^{\infty} C_i\right|\frac{1}{2^{j-1}} \quad \textbf{(Équation 27)}$$

Dice ordonné :

$$D_F(C,C') = \frac{2|U_C \cap U_{C'}|}{|U_C| + |U_{C'}|} = \frac{2\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|C_i \cap C_j'|\frac{1}{2^{\max(i,j)-1}}}{\sum_{i=1}^{\infty}|C_i|\frac{1}{2^{i-1}} + \sum_{j=1}^{\infty}|C_j'|\frac{1}{2^{j-1}}} \quad \textbf{(Équation 28)}$$

Cosinus ordonné :

$$Cos_F(C,C') = \frac{|U_C \cap U_{C'}|}{\sqrt{|U_C||U_{C'}|}} = \frac{\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|C_i \cap C_j'|\frac{1}{2^{\max(i,j)-1}}}{\sqrt{\left(\sum_{i=1}^{\infty}|C_i|\frac{1}{2^{i-1}}\right)\left(\sum_{j=1}^{\infty}|C_j'|\frac{1}{2^{j-1}}\right)}} \quad \textbf{(Équation 29)}$$

La mesure N ordonnée :

$$N_F(C,C') = \frac{\sqrt{2}|U_C \cap U_{C'}|}{\sqrt{|U_C|^2 + |U_{C'}|^2}} = \frac{\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|C_i \cap C_j'|\frac{1}{2^{\max(i,j)-1}}}{\sqrt{\left(\sum_{i=1}^{\infty}|C_i|\frac{1}{2^{i-1}}\right)^2 + \left(\sum_{j=1}^{\infty}|C_j'|\frac{1}{2^{j-1}}\right)^2}} \quad \textbf{(Équation 30)}$$

Le coefficient de débordement O_2 (overlap 2) ordonnée :

$$O_{2F}(C,C') = \frac{|U_C \cap U_{C'}|}{\max\left(|U_C|,|U_{C'}|\right)} = \frac{\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|C_i \cap C_j'|\frac{1}{2^{\max(i,j)-1}}}{\max\left(\sum_{i=1}^{\infty}|C_i|\frac{1}{2^{i-1}}, \sum_{j=1}^{\infty}|C_j'|\frac{1}{2^{j-1}}\right)} \quad \textbf{(Équation 31)}$$

Le coefficient de débordement 1 (overlap1) :

$$O_{1F}(C,C') = \frac{|U_C \cap U_{C'}|}{\min\left(|U_C|,|U_{C'}|\right)} = \frac{\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|C_i \cap C_j'|\frac{1}{2^{\max(i,j)-1}}}{\min\left(\sum_{i=1}^{\infty}|C_i|\frac{1}{2^{i-1}}, \sum_{j=1}^{\infty}|C_j'|\frac{1}{2^{j-1}}\right)}. \quad \textbf{(Équation 32)}$$

Le rappel :

$$R_F(C,C') = \frac{|U_C \cap U_{C'}|}{|U_{C'}|} = \frac{\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|C_i \cap C_j'|\frac{1}{2^{\max(i,j)-1}}}{\sum_{j=1}^{\infty}|C_j'|\frac{1}{2^{j-1}}} \quad \textbf{(Équation 33)}$$

La précision:

$$P_F(C,C') = \frac{|U_C \cap U_{C'}|}{|U_C|} = \frac{\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|C_i \cap C_j'|\frac{1}{2^{\max(i,j)-1}}}{\sum_{i=1}^{\infty}|C_i|\frac{1}{2^{i-1}}} \quad \textbf{(Équation 34)}$$

Comme précédemment nous avons fait une comparaison expérimentale prenant en compte :

- les mesures J_F , D_F , Cos_F N_F O_{2F} , O_{1F} , R_F , P_F définie ci dessus(équation 26-34)
- les mesures classiques J , D , Cos , N , O_2 , O_1 , R , P définies en début d'article (équation 2-10)
- la mesure de proximité ordonnée à pondération de type puissance J_p^{Ω} construite à partir du Jaccard comme indiquée dans l'équation 15. Rappelons que nous n'avons pas besoin de prendre en compte toutes les mesures ordonnées à pondération, en effet, un des résultats de (Egghe, 2003) est que la fonction de pondération supprime l'effet particulier de la mesure c'est a dire $J_p^{\Omega} \approx D_p^{\Omega} \approx Cos_p^{\Omega} \approx N_p^{\Omega}$

Les résultats ont montré en ce qui concerne les mesures de type fort (J_F , D_F , Cos_F N_F O_{2F}) que ces mesures sont plus **précises et plus représentatives sur les caractéristiques de similarité de rang et de documents communs** que les mesures classiques ou les mesures ordonnées à pondération. De plus, dans le cadre des mesures fortes, nous avons observé une plus grande **sensibilité** des mesures floues. En revanche, de telles conclusions n'ont pu être mises en évidence pour les mesures de type faible (O_{1F} , R_F , P_F), nous avons supposé que cela venait du corpus qui n'était pas une réelle collection test.

4- CONCLUSION

Dans le contexte de comparaison d'objets informationnels à éléments disjoints, nous avons proposé deux solutions originales permettant de construire des mesures de similarité prenant en compte l'ordre des éléments présentés et applicables aux objets informationnels construits selon un semi-ordre disjoint. Les solutions proposées s'appuient soit sur une fonction de pondération, soit sur une redéfinition structurelle des mesures en utilisant la logique floue. Le travail de recherche théorique est a notre sens complètement finalisé *dans ce contexte précis*, il ne reste à notre sens à réaliser que des expérimentations concrètes pour observer le comportement de certaines mesures, en particulier les mesures de type faible (O_{1F} , R_F , P_F). D'autres contextes sont cependant ouverts et nécessitent une réflexion théorique.

Le formalisme présenté n'est pas applicable dans le contexte de semi-ordre d'éléments disjoints c'est à dire en particulier la comparaison de réponses de systèmes basés sur des interfaces cartographique construites à partir de réseaux sémantiques, cas de figure souvent présenté dans les systèmes d'aide à la navigation. Nous pensons que ce problème peut être résolu grâce au formalisme de la logique floue, en redéfinissant l'appartenance à un ensemble comme il l'a été fait dans l'équation 21.

Un autre contexte reste complètement inexploré, celui des ensembles à éléments répétés comme les textes. Les notions de *semi ordre d'éléments ou de classes non disjointes* pourrait être particulièrement intéressantes pour ces derniers : la phase peut être considérée comme une classe composé d'éléments "mots", la racine lexicale peut permettre de créer des classes composées d'élément "mots", …. A notre connaissance, toutes les méthodes de comparaisons de texte s'appuient actuellement sur le nombre de mots qu'ils ont en commun, la prise en compte non seulement de l'ordre mais ici surtout de la granularité de classe, exprimée sous la forme de semi-ordre, devrait ouvrir considérablement les perspectives dans bon nombre de domaines comme la recherche d'information, la fouille de texte ou la bibliométrie entre autres.

5- BIBLIOGRAPHIE

Auziol, E. (2001). Problématique d'analyse de situations de communication et contextes multimédias In *Actes du colloque La Communication Médiatisée par Ordinateur : un carrefour de problématiques* - Université de Sherbrooke, 15 et 16 mai 2001.

Boyce, B.R., Meadow, C.T., & Kraft, D.H. (1994). *Measurement in information science.* Academic Press.

Egghe, L., C. Michel C. (2002). Strong similarity measures for ordered sets of documents in information retrieval, In *Information Processing & Management* 38 (6) (2002) pp. 823-848. (novembre 2002)

Egghe, L., C. Michel C. (2003). Construction of weak and strong similarity measures for ordered sets of documents using fuzzy set techniques In *Information Processing & Management* (à paraître)

Grossman, D.A., Frieder, O. (1998). *Information retrieval. Algorithms and heuristics.* Kluwer Academic Publishers, Boston

Fluhr, C. (1997).. SPIRIT.W3 : A distributed Cross.Lingual Indexing and Search Engine. *Proceeding of the INET 97 « The Seventh Annual Conference of the Internet Society »*. June 24-27 1997. Kuala Lumpur, Malaysia.

Lainé-Cruzel, S., Lafouge, T., Lardy, J.P., & Ben Abdallah, N. (1996). Improving information retrieval by combining user profile and document segmentation. *Information Processing and management*.32 (3), 305-315.

Losee, R.M. (1990). *The science of information. Measurement and applications.* Academic Press, Inc.

Michel, C. (1999). *Evaluation de systèmes de recherche d'information, comportant une fonctionnalité de filtrage, par des mesures endogènes. Réalisation et evaluation d'un prototype de système de recherche d'information avec filtre selon les profils des utilisateurs.* PhD Thesis. University Lyon II. 6 January 1999. 322 p.

Michel, C. (2000). Diagnostic Evaluation of a personalized filtering information retrieval system. Methodology and experimental results. *Proceeding of RIAO 2000 "Content-Based Multimedia Information Access"*. Paris, 12-14 april 2000.

Radasoa, H. (1988). *Méthodes d'amélioration de la pertinence des réponses dans un système de bases de données textuelles.* PhD Thesis. University Paris Sud-Orsay. 28 November 1988. 156 p.

Salton, G. & McGill, M.J. (1983). *Introduction to modern Information Retrieval.* New York : McGraw Hill.

Van Rijsbergen, C. J. (1981). Retrieval effectiveness In *Information retrieval experiments*. London : Butterworth - pp32-43.

Voorhees E.M., Harman D. – Overview of the seventh Text Retrieval Conference TREC 7. In *Proceedings of the seventh Text Retrieval Conference TREC 7. – Gaitherburg 9-11 november 1998*

Tague, J. (1990). Rank and sizes : some complementarities and contrasts. *Journal of information science*. 1990, 16 (1), 29-35.

Tague-Sutcliffe, J. (1995). *Measuring information. An information services perspectives*. Academic Press.

Zadeh, L. (1979). *Fuzzy sets and their applications to cognitive and decision processes.* Academic Press, Ney York.

Zamir O., Etzioni O.(1999) : Grouper : A dynamic clustering interface to web search results – In *Proceeding of the Eighth International World Wide Web Conference* – May 11-14 1999 – Toronto, Canada (http://www8.org)