Organisation d'une masse documentaire électronique présentée à des lecteurs potentiels

David Wartel, Pascal Francq, Alain Delchambre

CAD/CAM – Université Libre de Bruxelles – 1050 Bruxelles – Belgique
dwartel@ulb.ac.be, pfrancq@ulb.ac.be, adelch@ulb.ac.be

Abstract

Since the early 90's, and thanks to the facilities of communication technologies developed such as the Internet, the number of electronic documents has continuously increased so that it seems more and more difficult and time-consuming for a user, drowned in this documentary space, to find relevant information. To reduce this effect of critical documentary mass and to decrease the access time to relevant resources, we describe in this article a method to organize electronic documents. This organization contains two levels: the first level being described by the concept of "points of view", the second one by the concept of "expert key-words". The method has to be a few time-consuming algorithm in order to be included into a real project.

Résumé

À l'heure où le nombre de documents électroniques ne cesse d'augmenter, et ce depuis plus de dix ans, grâce aux facilités de communication qui se sont développées, notamment l'Internet, il semble de plus en plus difficile et de plus en plus long pour un utilisateur, noyé dans cet espace documentaire, de trouver de l'information pertinente. Pour réduire cet effet de masse documentaire et diminuer le temps d'accès aux ressources importantes, nous décrivons dans cet article une méthode d'organisation de documents électroniques. La méthode doit être très simple et rapide pour s'inclure dans un système informatique réel. L'organisation est à deux niveaux : le premier niveau est décrit par la notion de « points de vue », et le second niveau par la notion de « mots experts ».

Mots-clés : organisation documentaire, information, rapidité d'exécution, « points de vue », « mots experts ».

1. Introduction

Que ce soit pour des recherches dans le domaine du World Wide Web (Berners-Lee *et al.*, 1994), ou au sein des réseaux internes aux entreprises (intranet), la quantité d'information stockée est généralement très importante, et croît quotidiennement, ce qui rend fastidieux l'accès aux informations. Des outils, déjà existants ou en cours de développement, permettent de stocker ces informations de manière à ce que l'utilisateur puisse accéder plus rapidement aux données qui l'intéressent. Parmi eux, le projet GALILEI[1] développe une solution d'analyse et de groupement permettant à des utilisateurs de partager des ressources intéressantes suivant leurs centres d'intérêt.

La méthode que nous décrivons dans cet article s'intègre dans le projet GALILEI. Elle consiste à organiser un ensemble d'éléments (documents électroniques, utilisateurs, ou groupes d'utilisateurs) en un système de mots clés à deux niveaux, que nous appellerons « points de vue » et « mots experts ». Des méthodes déjà existantes, comme les cartes auto organisatrices basées sur les réseaux de neurones (Kohonen, 2000) auraient répondu au problème d'organisation des données, mais la méthode souhaitée doit être rapide en temps de

[1] Subventionné par la région wallonne sous le contrat 01/1/4675 - http://www.galilei.ulb.ac.be

calcul pour pouvoir s'intégrer dans la partie « client » du projet. Elle doit permettre d'alléger la présentation à un utilisateur d'une grande quantité de données, essentiellement des documents, en les organisant sous forme de mots-clés. Il est important de bien noter que le faible temps d'exécution de la méthode est, dans notre cadre, la nécessité la plus importante.

Nous décrirons dans un premier temps le contexte de recherche (section 0), puis les représentations utilisées pour représenter les différents types d'information (section 0), suivront les règles de calcul des « points de vue », des « mots experts » et les méthodes d'organisation (section 0). Nous présenterons enfin la détermination des paramètres et les résultats obtenus sur nos bases de données (section 0).

2. Le projet GALILEI

Le projet GALILEI (Francq, 2003) est un projet utilisant l'indexation de documents électroniques de plusieurs types, permettant le groupement d'utilisateurs et la création de communautés virtuelles. Il a pour objectif d'aboutir à un produit fini, intégrable dans tout système de gestion documentaire. C'est dans le cadre de ce projet que la méthode décrite ici est implémentée, elle doit donc être performante et relativement peu gourmande en temps de calcul.

Pour être reconnu par le système GALILEI, un utilisateur doit se définir comme une personne possédant un ou plusieurs centres d'intérêt. Pour chacun d'entre eux, des profils devront être créés, qui permettront de se connecter au système. La définition d'un profil se fait juste par le choix arbitraire d'un nom que choisir l'utilisateur pour utiliser ce profil. Une fois connecté, le profil peut émettre des jugements sur des documents électroniques locaux, ou des documents accessibles à distance (intra ou inter réseaux). Le jugement des documents se fait de manière rapide et simple : alors qu'il consulte un document, l'utilisateur peut, en un click de souris, définir son document comme intéressant, peu intéressant, ou complètement hors sujet. C'est cette définition que nous utiliserons dans notre méthode pour définir un utilisateur : une personne permettant d'alimenter le système en source d'informations et effectuant un jugement sur certaines d'entre elles.

Les documents jugés sont analysés par le système, ce qui permet de décrire les profils et de les grouper en communautés de même intérêt. Les utilisateurs se verront alors proposer les documents provenant de leurs communautés et qu'ils n'ont pas jugés : les documents les plus intéressants seront donc partagés au sein de chaque communauté. Ces documents peuvent s'avérer être nombreux (plus de 100). Pour éviter d'être équivalente au résultat d'une simple requête dans un moteur de recherche, la méthode présentée va organiser les documents échangés entre utilisateurs et ainsi rendre plus convivial l'accès aux informations pertinentes. Il n'est pas nécessaire que l'ensemble des documents à échanger soit organisé, mais les plus intéressants d'entre eux doivent l'être absolument.

3. Représentations utilisées

L'algorithme d'organisation des documents est un outil qui se greffe au projet GALILEI. Il se base donc sur la représentation utilisée dans le projet : la représentation vectorielle (Salton et McGills, 1983). Cette représentation a principalement été choisie pour sa rapidité au niveau du temps de calcul.

3.1. *Représentation des documents*

Un document électronique est décrit par sa représentation vectorielle dans l'espace des mots. Cet espace est de très grande dimension. Pour réduire l'espace des mots, des traitements addi-

tionnels sont appliqués, comme l'utilisation d'anti-dictionnaires, ou d'algorithmes de désuffixation comme Porter (1980), ou son équivalent français Carry (Paternostre *et al.*, 2002).

Un document est donc perçu comme un vecteur, fortement creux, constitué de poids représentant le nombre d'occurrence du mot dans le document.

$$D_i = \{d_{i,j}\} \qquad \text{avec } d_{i,j} \text{ le nombre d'occurrence du mot } j \text{ dans le document } i$$

3.2. *Représentation des utilisateurs*

Les utilisateurs jugent certains documents. A partir des documents jugés comme relevants, nous calculons un vecteur de mots qui représente l'utilisateur. Ces mots représentent les N premiers termes du vecteur somme de tous les documents jugés par l'utilisateur, dont les poids ont été multipliés par un facteur de fréquence inverse (idf) qui représente le logarithme du nombre total de documents divisé par le nombre de documents contenant le mot. Les coordonnées de ce vecteur sont ensuite ordonnées de manière décroissante (les termes de plus fort poids en premier).

$$U_i = \{u_{i,j}\} \qquad \text{avec} \qquad u_{i,j} = \sum_k \frac{d_{k,j}}{\max_n(d_{k,n})} \times \log\left(\frac{D}{D_j}\right)$$

avec D le nombre total de documents, et D_j le nombre de documents contenant le mot j.

Tous comme les documents, les utilisateurs sont aussi représentés sous forme de. Remarquons cependant que, contrairement aux documents, les utilisateurs bénéficient du facteur idf. Celui-ci est nécessaire pour le système GALILEI afin d'améliorer la création de groupes. La méthode, en soit, reste la même : les utilisateurs sont identifiés par un vecteurs de mots ordonné par occurrences décroissantes.

3.3. *Représentation des groupes*

Les utilisateurs étant définis, ils sont regroupés en communautés virtuelles grâce à des algorithmes de groupement comme le k-Means (MacQueen, 1967 ; Wartel 2002) ou le *Genetic Virtual Community Algorithm* (Francq, 2003) Ces méthodes de groupements utilisent entre autre la similarité, définie comme étant le cosinus entre deux vecteurs mots. Les utilisateurs partageant les mêmes centres d'intérêt sont alors groupés ensemble. Là encore, nous calculons une représentation vectorielle des groupes ainsi obtenus.

4. Organisation

Les documents, les utilisateurs et les groupes sont donc représentés par des vecteurs de mots dans l'espace des mots. La méthode d'organisation a pour but de cartographier ces vecteurs en deux niveaux. Que les vecteurs soient des documents, des groupes ou des utilisateurs, la méthode d'organisation reste la même. La méthode est basée sur une analyse statistique de l'occurrence des mots au sein des données à organiser. Nous n'allons retenir des vecteurs d'entrée que les mots de plus forte occurrence. Intuitivement, on ressent que ces mots représentent les éléments les plus importants, nous allons donc les utiliser pour définir des « points de vue » et des « mots experts » qui permettront d'organiser l'ensemble des vecteurs.

4.1. *Paramètres*

La méthode d'organisation requiert certains paramètres de configuration :

- ***N***, représentant le nombre de mots clés à calculer.
- ***P***, représentant le pourcentage de « points de vue » désiré.

4.2. *Calcul des « points de vue » – « mots experts »*

Les « points de vue » correspondent aux *N* mots les plus importants (ou les plus « représentatifs » sur l'ensemble des vecteurs d'entrée *Ve*. Rappelons que ces vecteurs d'entrées peuvent être l'ensemble des groupes utilisateurs, l'ensemble des utilisateurs d'un groupes, ou l'ensemble des documents d'un utilisateur.

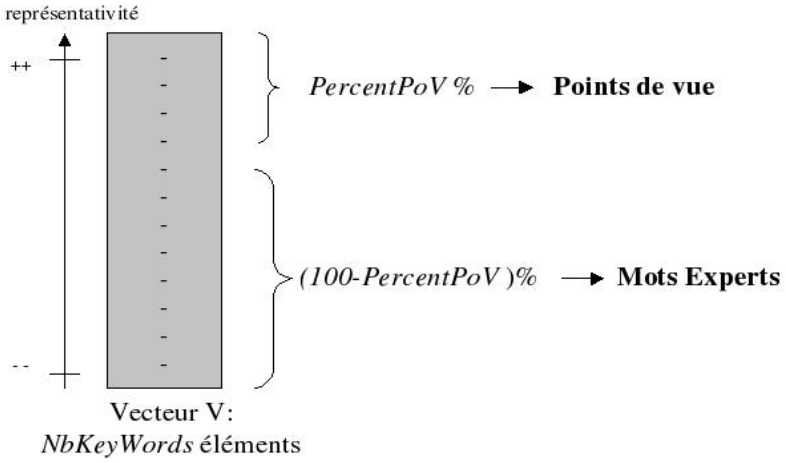

Figure 1. Séparation des «points de vue» et «mots experts»

Comme pour le calcul des utilisateurs, nous allons identifier les mots les plus importants de tous les vecteurs en calculant le vecteur *V* somme des vecteurs d'entrée, chaque terme étant multiplié par son facteur idf :

$$V = \{v_i\} \qquad \text{où} \qquad v_i = \sum Ve_i \times \log\left(\frac{N_v}{Nv_i}\right)$$

avec N_v le nombre total de documents et N_{vi} le nombre de documents contenant le mot i.

Nous obtiendrons donc *N* mots.

Parmi ceux-ci, les *P* premiers pourcents seront considérés comme « points de vue » pour nos vecteurs d'entrée. Les (1-P) autres pourcents seront leurs « mots experts » (*Figure 1*).

4.3. *Organisation des « mots experts »*

Pour obtenir notre organisation à deux niveaux, il faut associer les « mots experts » aux « points de vue ». Pour cela, nous allons calculer pour chaque « mot expert » le « point de vue » avec lequel il a la cooccurrence la plus forte sur l'ensemble des vecteurs d'entrée. La cooccurrence d'un mot est définie comme le nombre de fois que les mots apparaissent ensemble sur l'ensemble des vecteurs d'entrée *V*:

$$Co(m_1, m_2) = card(v_1 \cap v_2)$$

avec v_1 l'ensemble des vecteurs contenant le mot *m1* et v_2 l'ensemble des vecteurs contenant le mot *m2* :

$$v_1 = \{V_1 \subset V \mid \forall \vec{v} \in V_1, \vec{v}.\overrightarrow{m_1} \neq 0\}$$

$$v_2 = \{V_2 \subset V \mid \forall \vec{v} \in V_2, \vec{v}.\overrightarrow{m_2} \neq 0\}$$

Un « mot expert » ne pourra donc être associé qu'à un et un seul « point de vue » (*Figure 2*).

Il peut arriver qu'un « mot expert » ne trouve pas de « point de vue » lui correspondant : cela voudra dire que ce « mot expert » n'apparaît avec aucun « point de vue » sur l'ensemble des vecteurs d'entrée. Il est important de quantifier cette « perte » de « mots experts », et de vérifier que celle ci reste faible. Au cas où celle-ci devient trop forte, les paramètres *N* et *P* devront être réajustés. De même, il arrive qu'un « point de vue » reste vide, c'est-à-dire qu'aucun « mot expert » ne lui est attribué. Ce cas est moins grave : le « point de vue » peut être supprimé si l'organisation des données le permet.

```
M={ensemble des Mots Experts}
Tant que M non vide
    m=premier mot de M
    P={ensemble des Points de vue}
    best_occurrence = 0
    best_pov = 0
    Tant que P non vide
        p=element pris au hasard dans P
        Calculer C=co_occurrence(m,p)
        Si C > best_occurrence
            best_occurrence=C
            best_pov=p
        Supprimer p de P
    Si best_pov different de 0
        affecter m à best_pov

    Supprimer m de M
```

Figure 2. Algorithme d'organisation des «mots experts»

4.4. *Organisation des données*

La hiérarchie à deux niveaux réalisée, les vecteurs d'entrée, qui correspondent aux données à catégoriser, peuvent être classés suivant les « mots experts ». L'organisation des données est simple : un vecteur d'entrée est affecté à un « mot expert » si et seulement si le «mot expert» apparaît dans ce vecteur et sa fréquence est la plus importante de toutes les fréquences des « mots experts » pour ce vecteur d'entrée.

Ainsi, soit $\vec{m}_e$ un « mot expert », $\vec{v}$ un vecteur d'entrée :

$$\vec{v} \in \vec{m}_e \iff \vec{v}.\vec{m}_e \neq 0 \text{ et } \forall \vec{m}_i \in \{mots\ \exp\ erts\}, \vec{m}_e = \arg\min_i (\vec{v}.\vec{m}_i)$$

Nous avons choisi dans notre cas de n'affecter un même vecteur d'entrée qu'à un et un seul « mot expert ». Il suffit juste que ce mot apparaisse dans le document. Si on veut qu'un vecteur puisse appartenir à plusieurs mots experts différents, on pourra juste vérifier que le mot expert appartient à ce vecteur :

$$\vec{v} \quad \text{affecté} \quad \vec{m}_e \iff \vec{v}.\vec{m}_e \neq 0$$

De même que pour l'organisation des « mots experts », il est ici aussi important de mesurer la perte de vecteurs d'entrée, c'est-à-dire le nombre de vecteurs d'entrée non affectés à au moins un « mot expert ». Si ce nombre excède une valeur seuil, les paramètres *N* et *P* devront être

modifiés. La valeur seuil pourra être le nombre moyen de documents affectés à un même « mot expert ».

5. Dimensionnement : application sur un ensemble de groupe de nouvelles

Nous avons vu que l'algorithme dépend principalement de deux paramètres : *N* et *P*. Rappelons que l'objectif est d'obtenir une organisation rapide, sans trop de perte de données et minimisant le nombre de « points de vue » et de « mots experts » perdus (i.e., sans élément affecté).

L'algorithme d'organisation a été testé sur deux bases de données. La première (appelée news1500), issue de groupes de nouvelles (newsgroups) contient environ 1500 documents réunissant 76 utilisateurs en 18 groupes. La seconde (appelée news20000), toujours issue de groupes de nouvelles, contient environ 20000 documents, et regroupe 140 utilisateurs en 18 groupes. L'algorithme a été écrit en C++, et les résultats traités et analysés avec le logiciel R. Nous nous attachons, dans ce dimensionnement, à l'organisation des documents pour chaque groupe.

Pour les tests, le nombre de mots clés *N* a évolué de 10 à 100 pour le classement des documents. En effet, au delà de cette valeur, nous considérons que le nombre de « points de vue » devient trop important pour l'utilisateur. Le pourcentage de « points de vue » *P varie* de 10 % à 50 %. Au delà, nous aurions forcé certains « points de vue » à être vides.

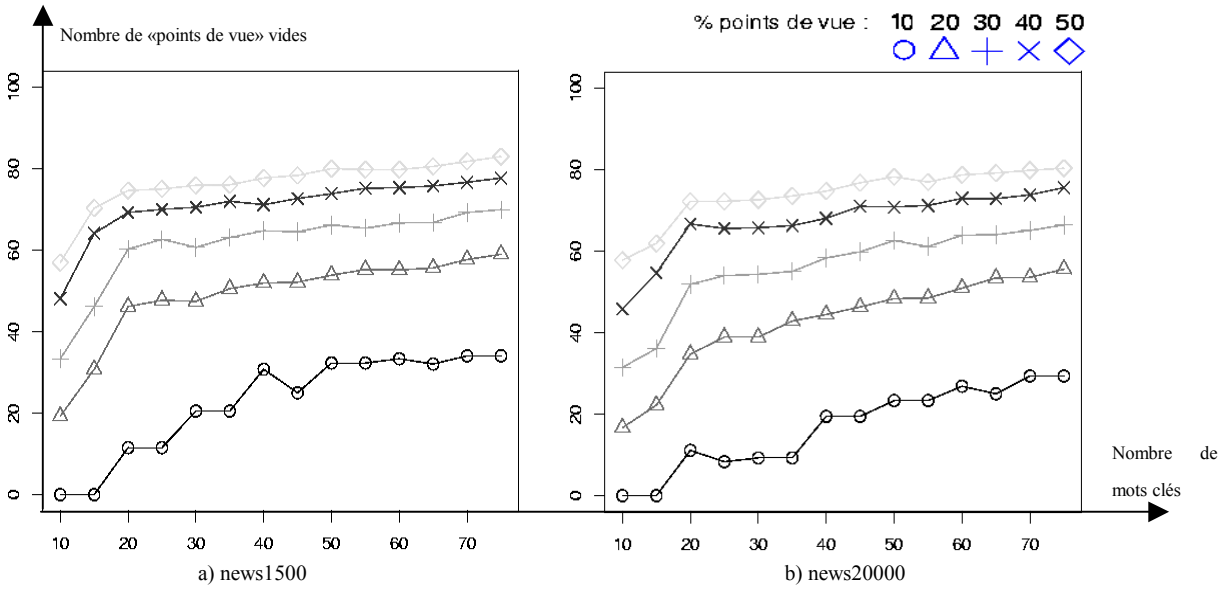

Figure 3. Nombre de «points de vue» vides pour différents pourcentages de «points de vue»

Si le nombre *N* devient trop grand, on peut arriver à des pourcentages élevés de « points de vue » vides, aussi bien sur la base de données news1500 que la base de données news20000.

Cet effet est légèrement amplifié si on augmente le pourcentage de « points de vue » parmi le nombre total de mots clés (*Figure 3.*). Ce constat est relativement normal : plus N est grand, plus nous allons considérer des mots peu importants comme mots clefs, et donc comme points de vue. Ces mots n'ont pas de cooccurrence assez élevée avec les mots experts, ils restent alors vides.

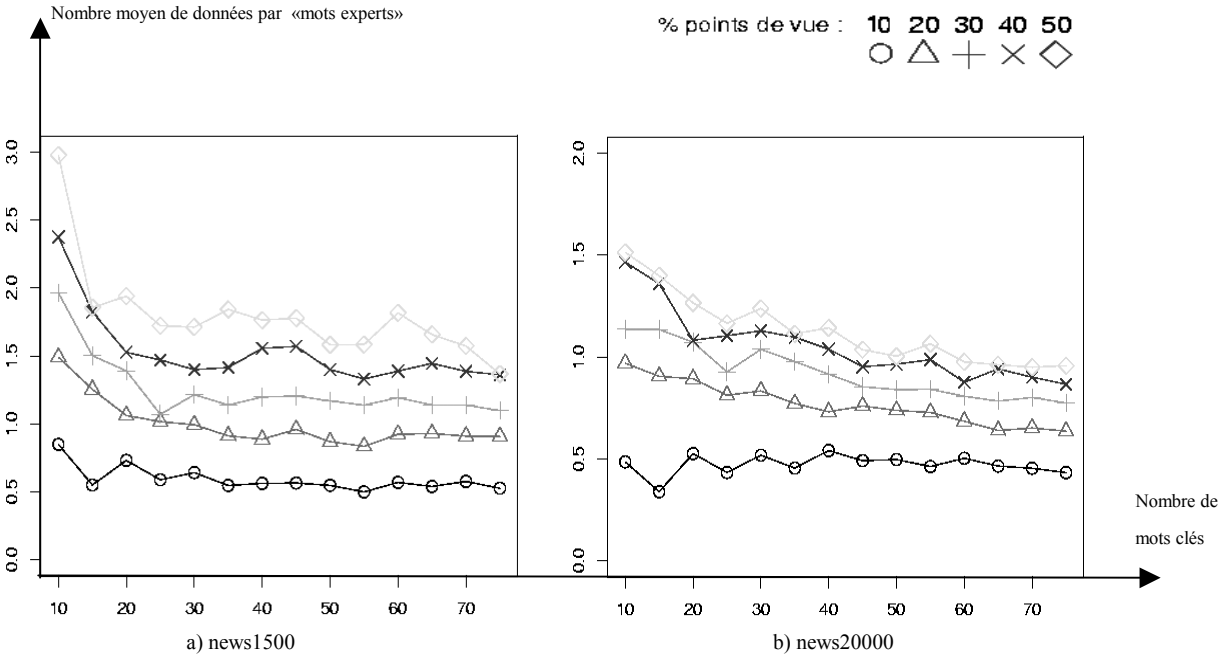

Figure 4. Nombre moyen de données par «mots experts» pour différents pourcentages de «points de vue»

Le fait que certains « points de vue » soient vides n'est, en soi, pas problématique : on peut les supprimer mais il faut cependant vérifier que le nombre de « points de vue » restants soit acceptable, et que le nombre de « mots experts » non affectés à un « point de vue » et le nombre de données non classées soient peu élevés. Dans notre cas, le nombre de « mots experts » non affectés est nul pour les deux bases de données: chaque « mot expert » a trouvé un « point de vue » lui correspondant.

De plus, on pourrait penser que le fait d'augmenter le nombre de mots clés conduirait à une organisation plus rigoureuse. Cependant, le nombre moyen de documents par « mots experts » varie peu (*Figure 4.*). En effet, l'augmentation du nombre de mots clés s'accompagne d'une augmentation du nombre de « mots experts » vides (*Figure 5.*). L'explication est la même que pour les « points de vue » vides : en augmentant le nombre de mots clefs, et donc de « mots experts », on retient des mots dont la fréquence d'apparition dans les documents n'est pas assez importante. Aucun document ne leur est donc affecté, ils restent vides.

Le nombre de données non affectées à des « mots experts » est également un paramètre important. Il correspond au nombre de données que le système n'indexera pas, elles seront donc considérées comme inexistantes ou perdues. Le nombre de ces données diminue avec l'augmentation du nombre de mots clés (*Figure 6.*). Si ce nombre reste trop faible (20), nous pouvons obtenir entre 20 et 40 % de perte pour la base de données news20000. Ces pertes deviennent inférieures à 10 % pour un nombre moyen de 50 mots clés. Il faut donc, pour cette base de données, utiliser un nombre assez important de mots clés pour minimiser cette perte,

le peu de documents perdus pourront être classés sous un autre mot expert crée pour regrouper les données non classées.

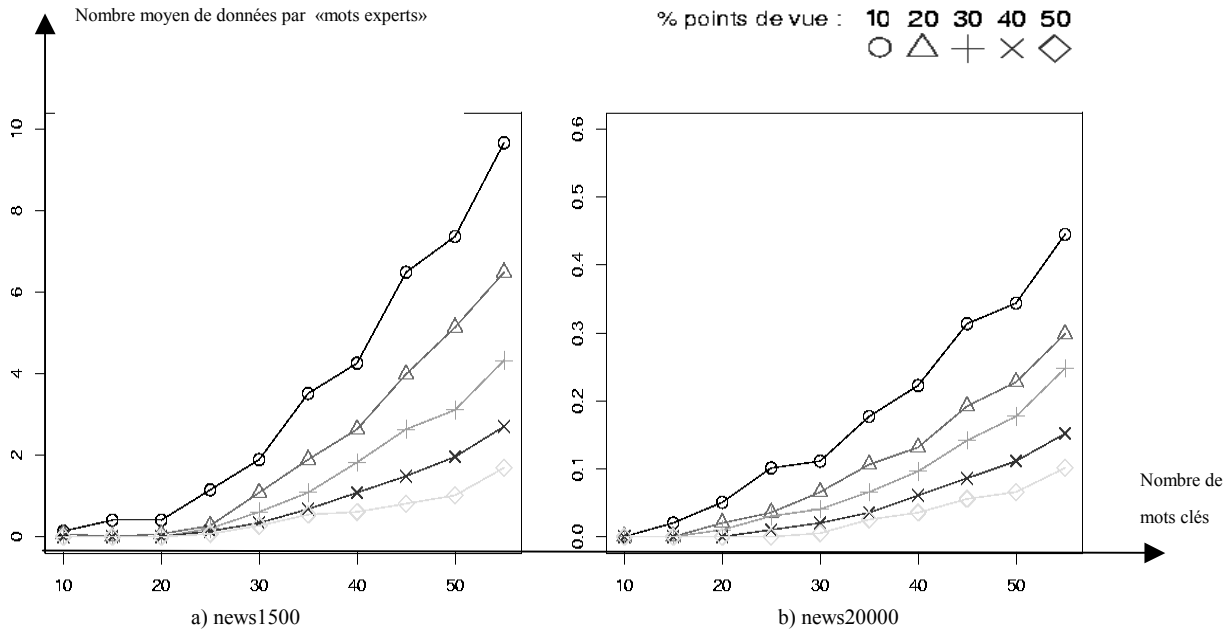

Figure 5. Nombre de « mots experts » vides pour différents pourcentages de « points de vue »

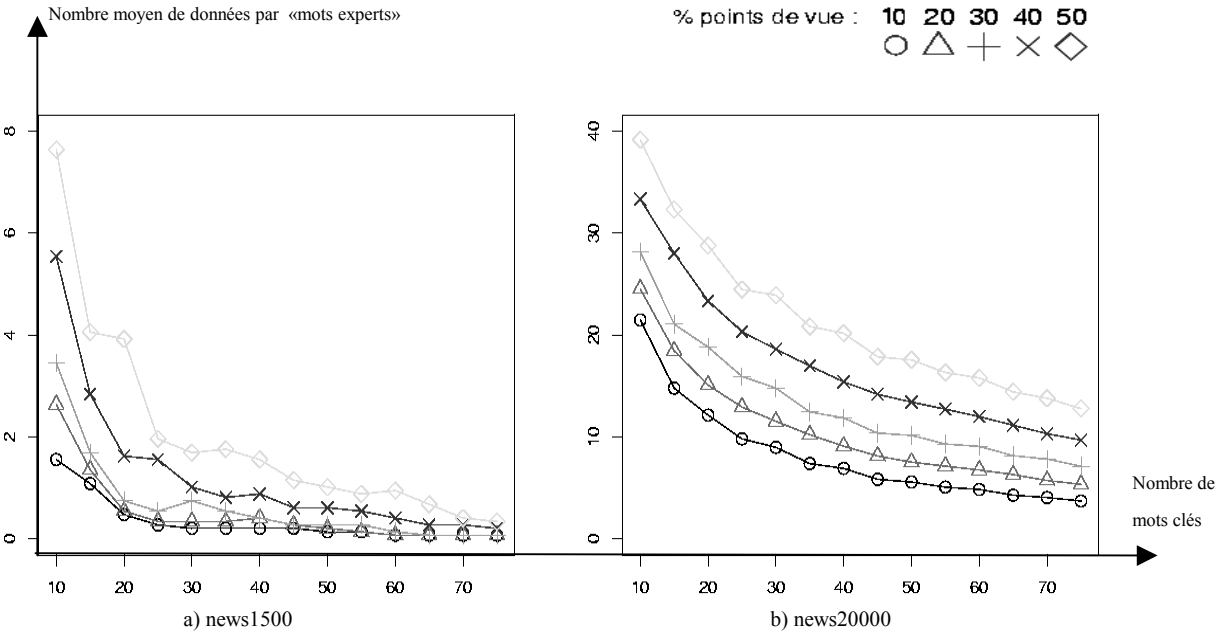

Figure 6. Nombre de documents non classés pour différents pourcentages de « points de vue »

Précisons cependant que ces documents non classés sont en grande partie les documents les moins intéressants à l'intérieur des communautés d'utilisateurs. En effet, ceux-ci sont perdus car ils ne contiennent aucun « mot expert », qui sont les mots importants sur l'ensemble des documents. Nous pouvons donc penser que les documents perdus sont des documents qui

n'apportent pas l'information pertinente à l'intérieur du groupe. Leur perte n'est donc pas un problème majeur.

Le pourcentage de « points de vue » doit rester, sur l'ensemble des résultats, relativement faible. Ainsi, il permet de minimiser le nombre de « points de vue » vides (*Figure 3.*), et d'obtenir un nombre faible de documents non affectés, et donc peu de perte de données (*Figure 6.*). Dans notre cas, un pourcentage de « point de vue » de 20 permet d'obtenir de bons résultats.

Dans cette configuration, nous obtenons des résultats très rapidement ce qui satisfait aux conditions d'intégrations dans notre système.

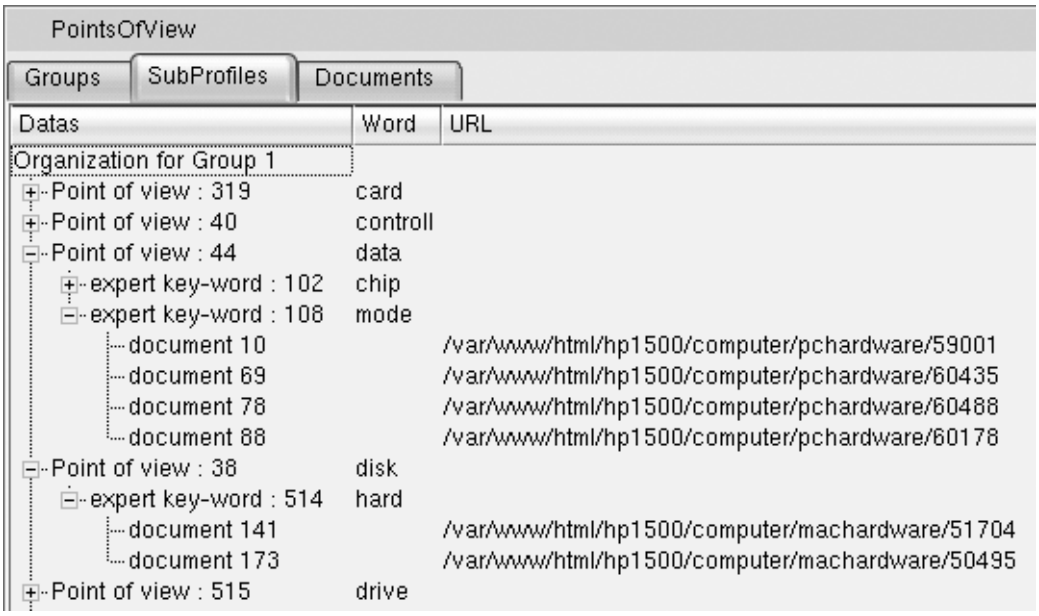

Figure 7. Présentation des points de vue dans GALILEI

La *Figure 7.* montre le résultat du découpage en points de vue et mots experts de l'ensemble des documents jugés par un utilisateur au sein du système GALILEI. L'intérêt global de l'utilisateur est l'informatique. Ainsi, nous remarquons une organisation en points de vue tels que « card », « controller », « data » ou encore « disk », et un sous découpage, pour chacun de ces points de vue, en mots experts tels que « chip », « mode » ou « hard ». L'organisation n'est certainement pas parfaite mais là n'est pas l'objectif recherché : les documents sont organisés de manière à adoucir l'effet de la masse documentaire ce qui permet une navigation plus conviviale pour l'utilisateur.

6. Conclusion

La méthode présentée dans cet article permet une organisation paramétrable d'un ensemble de documents. Cette organisation se fait à deux niveaux, un niveau de « points de vue » et un niveau de « mots experts », et, grâce à la même représentation vectorielle utilisée, elle peut s'appliquer sur trois types d'entités informationnelles : des groupes, des utilisateurs, et des documents. Pour chaque type d'entité, une étape de dimensionnement permet de définir les paramètres à utiliser pour l'organisation des données. Actuellement, des études sont menées pour trouver un dimensionnement automatique. Dans le cas étudié, il a été montré que pour garder une organisation significative et ne pas obtenir un nombre trop important de données

perdues, il nous faut un nombre de mots clés relativement faible (20) pour la base de données news1500, et moyennement élevé (45) pour la base de données news2000. Ces configurations conduisent tout de même à des pertes de documents, pertes qui ne sont pas très importantes vue qu'elles correspondent à des documents peu pertinents. Par ailleurs, le pourcentage de « points de vue » parmi ces mots clés doit lui rester faible (20 %), pour éviter d'accentuer certains effets néfastes du nombre de mots clés comme le nombre de points de vue vides ou le nombre de données perdues.

L'organisation des documents est rapide (moins de 3 secondes pour une base de données de 1500 documents et 18 groupes) et permet donc une implémentation dans le projet GALILEI.

Références

Berners-Lee T., Caillau R., Luotonen A., Nielsen H. et Secret A. (1994). The world-wide web. In *Communication of the ACM* : 76-82.

Francq P. (2003). *Structured and Collaborative Search: An integrated approach to share documents among users*. PhD Thesis, Université Libre de Bruxelles.

Kohonen T. (2000). *Self-Organizing Maps* (3rd edition). Springer Verlag.

MacQueen J.(1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical statistics and probability*, vol.(1) : 281-297.

Paternostre M., Francq P., Saerens M., Lamoral J. et Wartel D. (2002). *Carry, un algorithme de désuffixation pour le français*, URL : http://www.galilei.ulb.ac.be

Porter M.F. (1980). An Algorithm for suffix stripping. *Program*, vol. (14) : 130-137.

Salton G. et McGills M. (1983). *Modern Information Retrieval*, McGraw-Hill Book Co.

Wartel D.(2003). *Les algorithmes de clustering dans GALILEI,* rapport scientifique, Université Libre de Bruxelles.