Etude d'hypertextes construits sur les relations « descripteur » et « référent »[1]

Luc-Olivier Pochon et Alain Favre

RÉSUMÉ. *Cette communication propose un modèle d'hypertexte inspiré du modèle décrit par J.-P. Balpe en 1990. Aux notions de descripteur et d'unité d'information de celui-ci sont ajoutés les éléments qui servent d'ancrage (les référents). L'hypertexte est donc construit par la juxtaposition de deux relations. L'une donne les descripteurs d'une unité d'information, l'autre les référents. Ces deux relations qui indexent doublement des documents sont données par deux matrices R et D.*

Une matrice de structure, la matrice de circulation peut se calculer à partir de R et D. Elle permet de définir des équivalents des coefficients classiques de structure qui prennent en compte les particularités du modèle de l'hypertexte et donc d'une certaine sémantique des liens.

Ce modèle « abstrait » peut également servir dans des « instanciations » qui étendent la représentation documentaire de l'hypertexte comme, par exemple, à un ensemble de tâches mathématiques.

ABSTRACT. *This communication proposes a hypertext model inspired by the one described by J. - P. Balpe in 1990. To the concepts of descriptor and unit of information in the Balpe model, « referents » are added which are used as anchoring elements. The hypertext is thus built by the juxtaposition of two relations. The first one gives the descriptors of a unit of information, the second one their referents. These two relations which doubly index documents are given by two matrices R (for referents) and D (for descriptors).*

A matrix describing the structure called the circulation matrix is derived from R and D. It makes it possible to define equivalents of the usual structural coefficients which take into account the characteristics of the hypertext model and thus of certain semantic aspects of the links.

This "abstract" model can also be useful in "instanciations" which extend the documentary representation of an hypertext like, for example, the whole set of mathematical tasks.

MOTS-CLÉS : *structure d'hypertexte, graphe, indexation, pagerank.*

KEYWORDS: *hypertext structure, graph, indexation, pagerank.*

[1] Proposition de communication courte refusée à H2PTM'09

1. Introduction

Nous présentons dans cet article un modèle particulier d'hypertexte pour lequel nous étudions quelques indicateurs de structure.

L'origine du projet remonte à la publication d'un ouvrage d'hommage (Cornali-Engel & Weiss, 1996) où, pour créer avec les outils de l'époque trois documents (livre, CD-ROM, site Internet), un modèle d'hypertexte a été imaginé, inspiré du modèle décrit par Balpe (1990)[2]. A ce dernier basé sur les notions de descripteurs et d'unités d'information, nous ajoutons les éléments qui servent d'ancrage (les référents). Les descripteurs et les référents sont issus d'un même ensemble de concepts. Cela conduit à introduire systématiquement des coefficients duaux à ceux proposés par Balpe. Par exemple D*(**u**) est l'ensemble des concepts référent de l'unité d'information **u**. Il est le dual de D(**u**), ensemble des concepts descripteur de **u**. Finalement, le modèle se réduit à la juxtaposition de deux relations entre l'ensemble des unités d'information et l'ensemble des concepts. L'une donne les descripteurs d'une unité d'information, l'autre les référents.

Pour ce modèle d'hypertexte nous proposerons la définition de la matrice de circulation qui est un indicateur de la structure globale. Nous donnerons quelques informations tirées de cette matrice comparables aux indicateurs classiques de structure. Nous terminerons par donner une application possible (un autre « habillage ») de ce modèle « abstrait ».

2. Exemple

La figure 1a présente cinq unités d'information (U_1- U_5) et six concepts (C_1- C_6). Les concepts ont été associés aux unités d'information comme descripteurs (marqués par un trait vertical) ou comme référents (trait horizontal). L'hypertexte qui en résulte en liant référent et descripteur est représenté dans la figure 1b.

Les relations peuvent être données sous forme matricielle. La première, celle des descripteurs correspond à un indexage classique en documentation : une famille de documents (chacun correspondant à une ligne de matrice) est indexé par un certain nombre concepts (ou mots-clés). L'indexage peut être dichotomique (on ne traite que ce cas ici) ou alors pondéré. La matrice de cette première relation sera notée par *D*.

La deuxième matrice, *R*, correspond à une seconde indexation, qui correspond dans un modèle documentaire à signaler tous les termes (concepts) faisant référence à d'autres documents. En référence à ces deux matrices, nous appelons ce modèle RD.

La juxtaposition représentée par le produit matriciel $R*D^T$ donne la matrice d'adjacence du graphe pondéré et orienté sous-jacent de l'hypertexte. Les sommets en

---

[2] L'hypertexte à la base de ce modèle a été baptisé UTOPIA en référence au titre de l'ouvrage réalisé.

sont les unités d'information (l'aspect dual : le graphe des concepts ne sera pas traité ici).

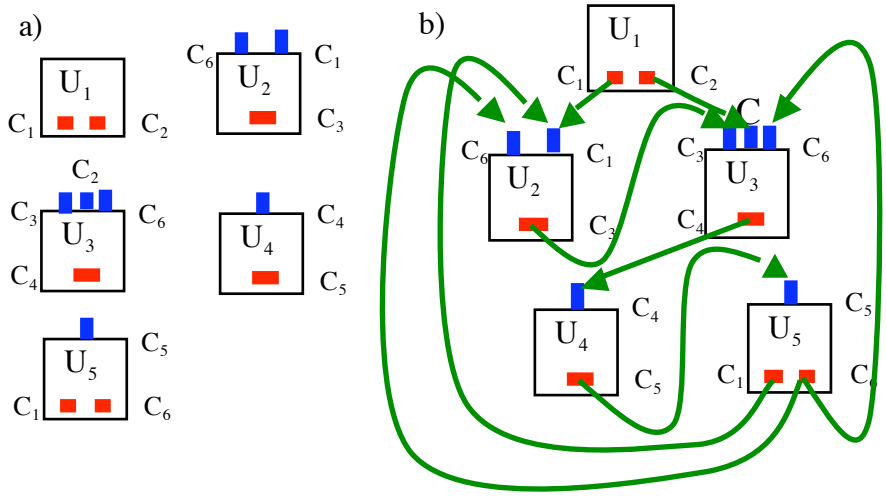

Figure 1. *a) Unités avec descripteurs et référents ; b) L'hypertexte associé*

Dans le cas représenté dans la figure 1, on a les matrices R et D associées suivantes :

$$R = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad D = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \quad G = RD^T = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 2 & 1 & 0 & 0 \end{pmatrix}$$

La matrice de circulation va être calculée à partir des matrices des deux relations de base.

3. La matrice de circulation

Il s'agit de caractériser le graphe de l'hypertexte par des valeurs qui en « mesurent » la morphologie à partir des informations locales.

Notre approche se fonde sur une idée probabiliste. Le but est de représenter la « probabilité » d'atteindre un certain type d'information, c'est-à-dire un concept **c**, à partir d'une unité **u** donnée. La procédure est la suivante : choisir au hasard (équiprobabilité) un référent de **u**, à partir de ce référent choisir au hasard l'unité d'information reliée et procéder de même jusqu'à atteindre **c**. Chaque fois on admet

une probabilité de 1/2 de continuer le processus. Ceci permet de tenir compte du fait que l'information obtenue via d'autres unités d'information est moins accessible.

Si l'unité **u** possède **c** comme référent on donnera comme probabilité d'atteindre **c** via ce référent la valeur 1/Di*(u). A cette valeur viendra s'ajouter la possibilité d'atteindre **c** via les unités d'informations liées. En formule:

$$p_u(c) = \frac{1}{Di*(u)}\left(\delta_u(c) + \frac{1}{2}\sum_{c'\in D*(u)}\sum_{u'\in U(c')}\frac{p_{u'}(c')}{\mathrm{Re}(c')}\right) \quad \text{[I]}$$

Dans la formule (I) : D*(u) est l'ensemble des concepts référents dans **u** ; Di*(u)=#D*(u) est l'ouverture de **u**, nombre de concepts référents dans u ; U(c) est l'ensemble des unités d'information dont **c** est descripteur ; Re(c)=#U(c) est le rendement ou la valence de **c**, nombre d'unités d'information décrites par **c** ; $\delta_u(c)$ vaut 1 si **c** est un "référent" de **u** et 0 sinon (coefficient de R).

La solution de ce système peut être donnée sous forme matricielle de la façon suivante : on forme $\overline{R}$ à partir de R dont on pondère les coefficients par les totaux marginaux de lignes (une ligne nulle reste telle quelle). De même on obtient $D|$ en pondérant les coefficients de D par les totaux de colonne. Puis on forme $G = R\,D|^{T}$ ($D|^{T} = \overline{D^{T}}$) qui est une matrice d'adjacence pondérée du graphe sous-jacent.

La solution de [I] est donnée par : $(I + S)\overline{R}$ [II]

Dans cette solution intervient la matrice S d'adjacence du graphe complété avec une pondération de 1/2. Cette matrice sera prise comme « mesure » de la structure globale de l'hypertexte. Nous appelons S la matrice de circulation de l'hypertexte.

$$S = \sum_{n>0}\left(\frac{1}{2}\hat{G}\right)^{n} = \frac{1}{2}\hat{G} + \frac{1}{4}\hat{G}^{2} + \ldots = \left(I - \frac{1}{2}\hat{G}\right)^{-1} - I \quad \text{[III]}$$

$\overline{R}$ et $\overline{D^{T}}$ sont sous-stochastiques par ligne (leurs coefficients sont positifs et la somme des lignes est inférieure ou égale à 1). Il en va de même pour $\hat{G}$ et donc pour toutes ses puissances successives. Cela assure la convergence de la série (II). De plus S est également sous-stochastique par ligne. Pour le calcul de S, on peut donc compter sur l'optimisation des systèmes de calcul pour la détermination de l'inverse[3].

Plusieurs propriétés de l'hypertexte peuvent être déduites de la considération de S, par exemple :

- S est stochastique ⇔ toute unité d'information possède au moins un référent (c'est-à-dire n'est pas un « puit »).

- Un élément diagonal est non nul ⇔ l'unité d'information correspondante est sur un circuit.

---

[3] Notre ambition est d'étudier des hypertextes de quelques milliers d'unités d'information ce que Matlab, par exemple, permet facilement de traiter.

Par ailleurs, en tronquant (mise à 0 de tous les coefficients inférieurs à une valeur ε donnée) et en « dichotomisant » S (mise à 1 de tous les coefficients non nuls), il est possible d'utiliser les techniques standard de la théorie des graphes pour définir la structure du graphe sous-jacent avec divers degrés d'approximation.

Dans le cas de l'exemple de la figure 1, on trouve :

$$S = \begin{pmatrix} 0 & 0.288 & 0.407 & 0.203 & 0.102 \\ 0 & 0.051 & 0.542 & 0.271 & 0.136 \\ 0 & 0.101 & 0.085 & 0.542 & 0.271 \\ 0 & 0.203 & 0.169 & 0.085 & 0.542 \\ 0 & 0.407 & 0.339 & 0.169 & 0.085 \end{pmatrix}$$

La structure donnée par S dépend d'une certaine sémantique des liens. En restreignant le nombre de concepts, ce qui revient à masquer (mettre à 0) des colonnes de R et D, on obtient une structure liée aux concepts restants.

4. Les simulations

Des simulations montrent, comme on peut s'y attendre, que les hypertextes de type RD sont d'autant plus fortement connectés que le rapport entre nombre d'unités et de concepts est grand et que les concepts sont également répartis dans les unités. Par exemple avec 1000 unités d'informations et 100 concepts avec en moyenne 10 descripteurs et 20 référents par unité, la connectivité est en moyenne supérieure à 40. Elle est inférieure à 10 si le nombre de concepts est 1000.

Le CORE[4] maximum de ces deux hypertextes sont constitués respectivement de 960 et 930 unités. En diminuant les densités des descripteurs et référents de moitié, le nombre d'unités du CORE maximum est de 770.

Par ailleurs, lorsque l'on considère les hypertextes décomposables en une partie de structure (les menus avec peu de descripteurs), de contenu (tous les descripteurs sont référents et vice-versa) et d'annotations (avec peu de référents) sous la forme : $H = M + C + A$ (Lowe & Hall, 1999). La matrice de circulation S de H est approchée par celle de C.

5. Hypertexte quelconque et modèle RD

Tout hypertexte peut se mettre sous la forme RD simplement en associant un concept à chaque lien[5]. Il est possible alors de comparer l'information apportée par S

[4] Le CORE d'un graphe orienté est une classe d'équivalence de sommets (d'unités) donnée par : deux unités u et u' sont en relation si et seulement si il existe un chemin (pas forcément direct) de u à u' et vice versa.

[5] Voir Pochon et Favre, 2007.

par rapport à des coefficients classiques de structure qui eux se calculent à partir de la matrice des distances converties (Botafogo, Revlin & Schneiderman, 1992). Les coefficients dont on obtient des correspondants sont ceux de *centralité relative entrante* (somme des colonnes de *S*), de *compacité* (rapport de la somme des coefficients de *S* au nombre d'unités) et de *stratification* (voir <auteurs> pour les calculs).

Le coefficient de centralité relative entrante peut en particulier être utilisé comme un « page ranking » comparable à celui donné par la version élémentaire de l'algorithme présenté par Bryan & Leise (2006) tout en tenant compte du facteur d'éloignement.

6. Pour conclure

Le modèle « abstrait RD peut s'avérer utile lorsqu'une analyse locale doit pouvoir se propager de façon globale. L'application qui retient actuellement notre attention est une approche locale des « espaces de connaissance » (Falmagne & al, s.d.) définis par un corpus de tâches, mathématiques en l'occurrence. Pour cela on considère un ensemble de tâches (les documents) et un ensemble de savoirs et de compétences (les concepts). Chaque tâche est doublement analysée : les savoirs élémentaires nécessaires à son accomplissement (les référents) et les savoir globaux mis en évidence par l'accomplissement de la tâche (les descripteurs). La structure *S* qui en résulte met en évidence la hiérarchie des tâches.

7. Bibliographie

Balpe, J.-P., *Les hyperdocuments*, Paris : Eyrolles, 1990.

Botafogo, R.A., Revlin, E., Schneiderman, B., « Structural Analysis of Hypertexts: Identifying hierarchies and Useful Metrics », *ACM Transactions on Information Systems*, vol. 10, n°2, 1992, p. 142-180.

Bryan, K., Leise, T., « The $ 25.000.000.000 Eigenvector. The Linear Algebra behind Google », *SIAM Reviews*, vol. 48, n°3, 2006, p. 569-581.

Cornali-Engel, I., Weiss, J. (dir.), *Des utopies à construire. Hommage à Jacques-André Tschoumy*, Neuchâtel & Lausanne : IRDP & LEP, 1996.

Falmagne, J.-C., Cosyn, E., Doignon, J.-P., Thiéry, N., *The assessment of knowledge in theory and in practice*. (Science_Behind_ALEKS.pdf sur http://www.aleks.com/, consulté: janvier 2005)

Lowe, D., Hall, W., *Hypermedia and the web*, New York : John Wiley & Sons, 1999.

Pochon, L.-O., Favre, A. *Connaissance, théorie de l'information et hypertextes: histoire d'une lecture sélective*. Neuchâtel : IRDP, 2007