Préambule

Ce document fait un résumé rapide de l'analyse sémantique latente. Il s'agit d'une méthode statistique permettant d'extraire automatiquement des relations conceptuelles entre les termes d'une collection de textes (par exemple : mobylette et voiture sont deux types de vehicule et sont donc reliés conceptuellement).

L'analyse sémantique latente

Principe

Le principe de cette analyse est très simple. Elle nécessite tout simplement d'avoir à disposition un logiciel permettant la décomposition en valeurs singulières de matrices.

Encore une fois, l'objectif de cette procédure est de créér un espace de représentation de mots dans lequel les relations de synonymie et d'hypernymie sont modélisées adéquatement tout comme la possibilité de polysémie (un même mot qui prend plusieurs sens différents).

Le principe est le suivant : soit $X = [\mathbf{x}_1 \cdot \mathbf{x}_m]^\top$ la matrice où chaque $\mathbf{x}_i$ est un vecteur-colonne qui est la représentation vectorielle du document i de la collection (cette représentation peut être binaire, fréquentielle ou bien TF-IDF). L'analyse sémantique latente consiste simplement à calculer la décomposition en valeurs singulière de X :

$$X = U\Sigma V^\top$$

où U et V sont des matrices orthonormales et Σ est une matrice diagonale contenant les valeurs singulières de X. Il suffit alors de se fixer un espace de dimension K et préserver la valeur des K premières valeurs singulières de Σ (i.e. de X) et annuler les autres. La matrice modifiée des valeurs singulières $\hat{\Sigma}$ permet de calculer un nouveau codage des textes dans l'espace sémantique latent par :

$$\hat{X} = U\hat{\Sigma}V^\top.$$

Exemple

Voici un exemple tiré du livre *Modeling the Internet and the Web* de Pierre Baldi, Paolo Frasconi, Padhraic Smyth.

d1: Indian government goes for open-source software
d2: Debian 3.0 Woody released
d3: Wine 2.0 released with fixes for Gentoo 1.4 and Debian 3.0
d4: gnuPOD released: iPOD on Linux... with GPLed software
d5: Gentoo servers running at open-source mySQL database
d6: Dolly the sheep not totally identical clone
d7: DNA news: introduced low-cost human genome DNA chip
d8: Malaria-parasite genome database on the Web
d9: UK sets up genome bank to protect rare sheep breeds
d10: Dolly's DNA damaged

Voici la matrice $X^\top$:

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
open-source	1	0	0	0	1	0	0	0	0	0
software	1	0	0	1	0	0	0	0	0	0
Linux	0	0	0	1	0	0	0	0	0	0
released	0	1	1	1	0	0	0	0	0	0
Debian	0	1	1	0	0	0	0	0	0	0
Gentoo	0	0	1	0	1	0	0	0	0	0
database	0	0	0	0	1	0	0	1	0	0
Dolly	0	0	0	0	0	1	0	0	0	1
sheep	0	0	0	0	0	1	0	0	0	0
genome	0	0	0	0	0	0	1	1	1	0
DNA	0	0	0	0	0	0	2	0	0	1

et, en conservant les 2 premières valeurs singulières on obtient la nouvelle matrice :

- Σ = diag(2.57, 2.49, 1.99, 1.9, 1.68, 1.53, 0.94, 0.66, 0.36, 0.10)

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
open-source	0.34	0.28	0.38	0.42	0.24	0.00	0.04	0.07	0.02	0.01
software	0.44	0.37	0.50	0.55	0.31	-0.01	-0.03	0.06	0.00	-0.02
Linux	0.44	0.37	0.50	0.55	0.31	-0.01	-0.03	0.06	0.00	-0.02
released	0.63	0.53	0.72	0.79	0.45	-0.01	-0.05	0.09	-0.00	-0.04
Debian	0.39	0.33	0.44	0.48	0.28	-0.01	-0.03	0.06	0.00	-0.02
Gentoo	0.36	0.30	0.41	0.45	0.26	0.00	0.03	0.07	0.02	0.01
database	0.17	0.14	0.19	0.21	0.14	0.04	0.25	0.11	0.09	0.12
Dolly	-0.01	-0.01	-0.01	-0.02	0.03	0.08	0.45	0.13	0.14	0.21
sheep	-0.00	-0.00	-0.00	-0.01	0.03	0.06	0.34	0.10	0.11	0.16
genome	0.02	0.01	0.02	0.01	0.10	0.19	1.11	0.34	0.36	0.53
DNA	-0.03	-0.04	-0.04	-0.06	0.11	0.30	1.70	0.51	0.55	0.81

On voit clairement la séparation des termes entre deux sujets ! ! !

Conclusion

Méthode extrêmement simple et très efficace pour l'extraction de relations conceptuelles entre termes. La question de la dimension à conserver reste un paramètre à régler. L'utilisation de cette méthode nécessite par ailleurs la disponibilité d'une bibliothèque d'algèbre linéaire permettant de gérer des matrices de grande dimension (et éventuellement creuses).