Modèle LSI (Latent semantic indexing)

Les autres modèles présentés dans ce cours utilisent les mots-clés pour représenter le contenu d'un document (ou d'une requête). On se pose souvent la question sur la qualité de cette représentation. En effet, les mots-clés ne sont pas une représentation idéale. Le but de LSI est de transformer une représentation par des mots-clés en une autre représentation qui est "meilleure". Le mot "meilleur" est compris dans le sens suivant: les documents et les requêtes sémantiquement similaires seront plus proches avec la représentation transformée qu'avec les mots-clés. La transformation par LSI est comme suit:

1. Au début, chaque document et requête est représenté comme un vecteur de mots-clés.
2. LSI utilise la SVD (singular value decomposition) pour créer un nouvel espace vectoriel:

$$X = T_0 S_0 D_0'$$

où X est la matrice de document-terme originale (de taille t× d)
T_0 est une matrice t× m
S_0 est une matrice m× m diagonale (seulement les élément en diagonal sont non-nuls)
D_0' est une matrice m× d.
La valeur m est choisi comme une valeur ≤ min(t,d).
En plus, on trie les valeurs dans S_0 dans l'ordre décroissant. Il existe juste une seule décomposition de cette façon.

3. On pense que la représentation par des mots-clés contiennent beaucoup de bruits. Typiquement, ces bruits se retrouvent dans les dimensions de S_0 qui ont des valeurs faibles. Ainsi, la technique de LSI veux supprimer ces dimensions de valeurs faibles (ou de les ramener à la valeur 0), ce qui ramène les dimensions de S_0 à k, et cette matrice réduite est notée par S. En conséquence, les matrices T_0 et D_0 nettoyées deviennent T et D.
La matrice reconstituée correspondrait à une matrice de document-terme nettoyé. Mais LSI ne fait pas cette reconstitution. Plutôt, les matrices décomposées vont rester. En particulier S correspond à un nouvel espace vectoriel. Des exemples montrent que certains documents sémantiquement similaires se seront rapprocher dans ce nouvel espace.

4. Quand une requête est soumise, elle est aussi traduite dans cette nouvel espace. Dans l'article de Deerwester et al., un requête est transformée d'abord en un pseudo-document comme suit:

$$D_q = X_q' T S^{-1}$$

où X_q est le vecteur de mots-clés de la requête (dans l'espace des mots-clés).
Ensuite, ce pseudo-document est ajouté dans la matrice D comme un nouveau "document". Le calcul de similarité entre chaque paire de documents peut se faire par: $D S^2 D'$
Ainsi, après ce calcul, on peut connaître la similarité de ce pseudo-document (ou la requête) avec tous les autres documents.

Ce modèle a montré des performances très intéressantes. Pour un corpus de petit ou moyenne taille, la performance est très supérieure au modèle vectoriel classique, et est un des meilleurs modèles. Quand la taille de corpus augmente, la différence avec les autres modèles classiques semble diminuer.

Il y a aussi la question sur la valeur de k (les dimensions à garder dans S_0). Il n'y a que des expérimentations qui peuvent déterminer une valeur optimale. Il se trouve que la valeur d'environ 300-500 fonctionne bien.
Pour une description plus détaillé, lire l'article suivant:
http://citeseer.nj.nec.com/deerwester90indexing.html
Indexing by Latent Semantic Analysis (1990)
Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman
Journal of the American Society of Information Science