Expansion de requête

1. Problématique

Les approches classiques présentées dans les sections précédentes sont toutes basées sur l'utilisation de "mot-clé". Un mot-clé est supposé de représenter une partie de contenu du document et de la requête. C'est raisonnable, en tenant compte de la représentativité des mots-clés pour le contenu et la simplicité de leur manipulation. Cependant, les opérations de recherche définies dans les modèles de RI posse la notion de mot-clé plus loin, bien que implicitement. Il est implicitement supposé qu'un mot-clé est l'unique représentant d'une signification unique. Autrement dit, il est supposé qu'il y a une correspondance du type 1:1 entre les mots-clés et les sens. C'est évidemment faux. En réalité, un mot peut avoir plusieurs sens, et un sens peut être exprimé par des mots différents.

Afin de traiter cette correspondance n:n entre les mots-clés et les sens, on propose généralement de faire deux traitements suivants:
1. considérer des mots (ou termes) reliés pour étendre la requête.
2. considérer une désambiguisation des mots afin de reconnaître le sens dénoté.

Le premier traitement tente de considérer le fait qu'un sens peut être véhiculé par des mots différents. Le deuxième traitement considère le phénomène dans le sens opposé. Dans cette section, nous considérons juste le premier traitement.

Une expansion de requête est aussi vue comme un traitement pour "élargir" le champ de recherche pour cette requête. Une requête étendue va contenir plus de termes reliés. En utilisant le modèle vectoriel, par exemple, plus de documents seront repérés. Ainsi, ce traitement est souvent vu comme un moyen d'augmenter le taux de rappel. Cependant, nous savons qu'il a pas de sens de parler du rappel sans considérer en même temps la précision. Ainsi, cette affirmation que l'expansion de requête va conduire à un meilleur rappel n'est pas tout à fait juste. Il faut plutôt dire que, en sélectionnant les documents selon un seuil de similarité entre un document et une requête, nous avons la chance de sélectionner plus de documents pertinents avec une requête étendue.

L'utilité de l'expansion de requête dépend fortement de deux facteurs:
1. Quels mots doit-on utiliser pour étendre la requête?
2. Comment les nouveaux mots doivent-ils être ajoutés dans la requête?

2. Méthodes proposées

L'expansion de requête est souvent utilisée dans le modèle booléen et le modèle vectoriel.

Pour le modèle booléen, il est assez facile de voir le processus d'expansion: Si on considère qu'il y a une relation forte entre A et B (ou s'il y a la relation d'implication $B \rightarrow A$), et que A apparaît dans une requête booléenne, alors on remplace A par $(A \vee B)$. Typiquement, cette expansion utilise des synonymes. Si on considère que B est un bon substitut de A, alors la requête étendue ne change pas la sémantique de la requête initiale. En général, une requête booléenne n'est pas pondérée. Il n'y a donc pas de question de pondération pour les nouveaux mots.

Pour le modèle vectoriel, dans la même situation, on va simplement ajouter B dans le vecteur de la requête (s'il n'y est pas déjà). Cette méthode d'ajout est généralement utilisée. La question qu'on se pose est plutôt sur la pondération des nouveaux mots dans le vecteur. Elle peut être:
- un mot ajouté B est pondéré comme le mot initial A en relation.
- Un mot ajouté B est pondéré comme la pondération de A multiplié par un facteur. Ce facteur peut être fixe (par exemple, 0.5), ou bien déterminé selon le nombre de B en relation avec A (L'idée est que si A conduit à beaucoup de mots B reliés, ces mots reliés doivent être pondérés plus faiblement).

Cette méthode de pondération a été expérimenté par plusieurs chercheurs, entre autres, Voorhees (1994). Elle utilise le thésaurus Wordnet pour déterminer les mots à ajouter dans le vecteur. Cependant, le résultat est négatif: avec cet ajout, la performance est dégradée.

Ici, on doit se poser la question sur cette méthode naïve de faire l'expansion dans un vecteur. On peut observer que l'expansion n'est pas uniforme pour tous les mots de la requête. Seulement certains mots seront étendus, et certains sont étendus plus fortement (par plus de mots) que d'autres. En conséquence, un concept qui est fortement étendu sera renforcé dans le vecteur obtenu, car il est maintenant représenté plusieurs fois - par le mot initial et par tous les mots ajoutés. Est-ce que ces concepts renforcés sont réellement importants dans la requête? Pas nécessairement.

Étant donné la manière d'évaluer la similarité, il est possible qu'un document retrouvé ne concerne qu'un seul concept - il contient plusieurs représentation de ce concept, mais aucun autre concept. Ce document pourrait être mieux classé qu'un autre document qui concerne tous les concepts de la requêtes (selon la pondération des mots). Dans le livre de Salton & McGill, 1983, on parle de la spécificité et de l'exhaustivité. La spécificité d'un document détermine si tout le contenu du document est concentré sur le thème de la requête, alors que l'exhaustivité veut mesurer si tous les aspects de la requête a été abordé dans le document. Pour notre exemple, on voir que le premier document peut être spécifique, mais pas exhaustive. Le deuxième document est plus exhaustive.

3. Quels mots ajouter?

Les mots utilisés pour faire l'expansion de requête doivent être fortement reliés à la requête. Typiquement, on utilise un dictionnaire de synonyme, ou un thésaurus. Les mots reliés avec des mots de la requête par certains types de relation (e.g. IS_A) sont choisis pour étendre la requête.

Il y a aussi des études qui essaient de trouver automatiquement les mots fortement reliés. La plupart de ces approches exploitent les co-occurrences: Plus deux mots co-occurent dans des textes, plus on suppose qu'ils sont fortement reliés. Une fois ces relations statistiques choisies, on peut les utiliser dans un processus d'expansion de requête.

La plupart d'approches d'expansion considère chaque mot de la requête en isolé. Dans l'article de Qiu et Frei (1993), ils pensent qu'il vaut mieux choisir des mots qui sont reliés à la requête qu'aux mots individuels de la requête. Autrement dit, ils calculent la relation entre un mot et la requête dans son ensemble, et choisissent à utiliser les mots les plus fortement reliés. Ils montrent que cette approche est meilleure que de faire expansion de mots.

Il est aussi suggéré que le processus d'expansion soit interactif: L'usager peut filtrer les mots proposés par le système. Cette approche est utilisée dans certains systèmes, par exemple, Medline qui intègre un thésaurus du domaine médical.

4. Expansion comme une inférence logique

Il est possible de voir le processus d'expansion de requête comme une étape d'inférence en utilisant des connaissances (relations stockées dans un thésaurus, par exemple). Ce point de vue est présenté dans l'article mis dans la semaine 6. Dans ce document, il y a aussi plus de références sur les travaux reliés.

Références

Qiu, Y. and Frei, H. P. (1993). Concept based query expansion. Research and Development in Information Retrieval, ACM-SIGIR, 160-169.
Voorhees, E. M. (1994). Query expansion using lexical-semantic relations. Research and Development on Information Retrieval - ACM-SIGIR, Dublin, 61-70.