Texte

Bases de données documentaires et distribuées
Cours NFE04
Recherche avec classement

Auteurs : Raphaël Fournier-S'niehotta, Philippe Rigaux, Nicolas Travers
prénom.nom@cnam.fr

Département d'informatique
Conservatoire National des Arts & Métiers, Paris, France

Classement par pertinence

Les requêtes Booléennes classent les documents en deux catégories : ceux qui satisfont la requête, et les autres. C'est 1, ou c'est 0.

Le **classement par pertinence** (*ranked search*) trie un résultat en fonction d'un "poids" (*weight*) mesurant le degré de pertinence d'un document pour une recherche.

Plusieurs approches complémentaires pour évaluer la pertinence d'un document d pour une recherche q :

- par **similarité** entre d et q ;
- par **l'importance** de d, évaluée par exemple par sa position dans une collection structurée (e.g., graphe, cf. PageRank) ;
- en prenant en compte l'utilisateur (**profil**, **boucles de pertinence**, etc.). Cf. l'analyse des **clics** !

Recherche par similarité

La requête q **et** le document d sont placés dans un même **Espace métrique**, doté d'une fonction de distance m_E.

Une fonction f produit un **descripteur** (le plus souvent sous la forme d'un vecteur appelé *features vector*) à partir d'un document.

La **similarité**, ou **score**, est l'inverse de la distance.

$$sim(q,d) = \frac{1}{m_E(f(q), f(d))}$$

Langage de requêtes ultra-simplifié

La requête est exprimée par un ensemble de mots-clés, et interprétée comme un "document" simplifié.

Quelques caractéristiques

Avec l'approche par similarité, le résultat (c.à.d. les documents qui ont un score non nul) est potentiellement **très grand**.

Il est impératif de **classer** le résultat par ordre de score croissant, et de présenter les k prtemiers à l'utilisateur (typ., $k \simeq 10 - 20$).

Souvenez-vous : rappel et précision

- La **précision** est la fraction du résultat qui est vraiment pertinente.

$$precision = \frac{|relevant| \cap |retrieved|}{|retrieved|}$$

- Le **rappel** est la fraction des documents pertinents présente dans le résultat.

$$recall = \frac{|relevant| \cap |retrieved|}{|relevant|}$$

Vocabulaire usuel

On place **documents** et **requête** dans un **espace métrique**, souvent un **espace vectoriel**.

L'essentiel : une fonction de **similarité**, définie à partir d'une distance, ou directement.

Les éléments de cet espace sont appelés **descripteurs** ou **vecteurs de caractéristiques** (*features vector*).

Le **score** $sim(q, d)$ mesure la pertinence d'un document d par rapport à une requête (besoin) q.

Le résultat est **classé** sur le score ; les k premiers documents sont présentés.

Soyons concrets : première approche pour la recherche plein texte

Supposons connu l'ensemble de tous les termes $V = \{t_1, t_2, \cdots, t_n\}$ de tous les termes utilisables pour la rédaction d'un document.

Exemple : V = {"papa", "maman", "gateau", "chocolat", "haut", "bas"}

On définit $E = \{0,1\}^n$ comme l'espace de tous les vecteurs constitués de n coordonnées valant soit 0, soit 1. Ce sont nos descripteurs.

Exemple : vecteurs constitués de 6 coordonnées valant soit 0, soit 1.

Fonction f associant un document d à son descripteur $v = f(d)$.

$$v[i] = \begin{cases} 1 \text{ si } d \text{ contient le terme } t_i \\ 0 \text{ sinon} \end{cases}$$

Jusque là, rien de nouveau

C'est la représentation déjà vue pour les matrices d'incidences.

Exemple(s)

Rappel : on a $V = \{$"papa", "maman", "gateau", "chocolat", "haut", "bas"$\}$

Soit le document d_{maman}=`maman est en haut, qui fait du gateau`

alors $f(d_{maman}) = [0,1,1,0,1,0]$

Test

À vous de jouer : d_{papa}= `papa est en bas, qui fait du chocolat`

Deux remarques :

- On ignore certains mots (les "mots inutiles" ou **stop words**)
- L'ordre des mots dans le document est **ignoré** (approche "*bag of words*")

La distance

Dans un espace vectoriel, on peut penser à prendre la **distance Euclidienne**. Si v_1 et v_2 sont deux vecteurs :

$$E(v_1, v_2) = \sqrt{(v_1^1 - v_2^1)^2 + (v_1^2 - v_2^2)^2 + \cdots + (v_1^n - v_2^n)^2}$$

Et la similarité est l'inverse de la distance

$$sim(v_1, v_2) = \begin{cases} \infty \text{ si } v_1^i = v_2^i \text{ pour tout } i \\ \frac{1}{E(v_1, v_2)} \text{ sinon} \end{cases}$$

Exemple, pour $q =$ "`maman haut chocolat`", $v_q = [0,1,0,1,1,0]$

$$sim(v_q, d_{maman}) = \frac{1}{\sqrt{2}}$$

Test

À vous de calculer $sim(v_q, d_{papa})$

Quelques points à retenir

Important

Une différence concrète très sensible (illustrée ci-dessus) avec les requêtes Booléennes est qu'il n'est pas nécessaire qu'un document contienne tous les termes de la requête pour que son score soit différent de 0.

- "chocolat", un des mots-clés de q, n'apparaît pas dans le document d_{maman}, malgré tout classé en tête ;

Approche présenté ici : simple mais **nombreux inconvénients** (exercices).

Ignore l'impact de : la taille des documents, la taille du vocabulaire, le nombre d'occurrences d'un terme dans un document et la rareté de ce terme dans la collection.