Fouille du contenu web:

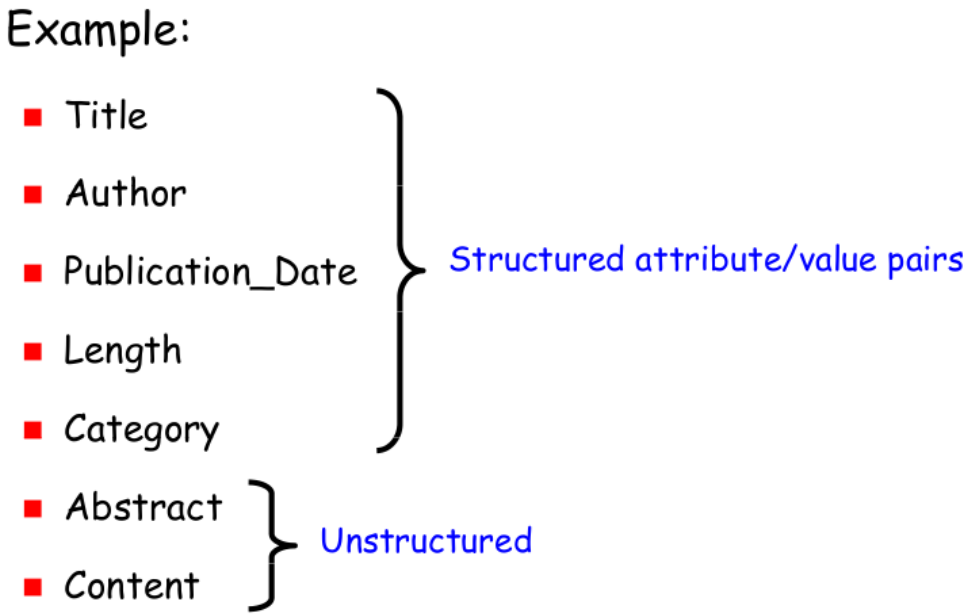
modèle d’espace vectoriel, recherche web, indexation sémantique latente (LSI), extraction automatique de thèmes.

Introduction

Web Content Mining utilise les idées et les principes de data mining pour découvrir de connaissances a partir du contenu des pages Web et des résultats de la recherche sur le Web.

Si nous pouvons représenter des documents par un ensemble d'attributs, nous pourrons utiliser les méthodes d'exploration de données existantes

Les données Web sont principalement semi-structurées et/ou non structurées, tandis que l'exploration de données traite principalement des données structurées



Prétraitement du texte et des pages Web

l’objectif du prétraitement est de minimiser l’espace de recherche.

- Tokenisation : term, handling of digits, hyphens, punctuations
- Normalisation : cases of letters
- Suppression de stopword : des mots fréquents et insignifiants. Articles, prepositions and conjunctions comme the, and, , a, an, is, of, that, are, ...
- Stemming: réduire les mots à leur racine
- ,,.

Pour pages Web

- Suppression des balises HTML
- identification des blocs de contenu principaux

Modèle d’espace vectoriel 1/2

Comment représenter un document ?

La représentation d’un ensemble de documents sous forme de vecteurs dans un espace vectoriel commun est connue sous le nom de **modèle d’espace vectoriel** ou “bag of words”

Elle utilise des statistiques pour ajouter des dimensions numériques à du texte non structuré.

Comme : fréquence des mots, fréquence des documents, la longueur du document…

Elle est fondamentale pour un grand nombre d'opérations de recherche d'information, allant de l’évaluation de documents pour une requête, à la classification de documents et au regroupement de documents.

Modèle d’espace vectoriel 2/2

- La représentation **booléenne**
  - Chaque entrée décrit un document
  - Attribut indiquant si un terme apparaît ou non dans le document
- La **fréquence des termes** noté $tf_{t,d}$: Les attributs représentent la fréquence à laquelle un terme t apparaît dans le document d
- La **Fréquence relative**: Nombre d’occurrences/Nombre de mots dans le document

	word1	word2	word3
doc 1	1	0	0
doc 2	1	1	1
doc 3			

	word1	word2	word3
doc 1	5	0	1
doc 2	7	4	12
doc 3			

Schéma de pondération des fréquences des termes

Pondération TF*IDF : donner plus de poids aux termes rares

- N: le nombre total de documents
- tf : fréquence des termes (augmente le poids des termes fréquents)
- df_t : la fréquence des documents, le nombre de documents qui contiennent un terme t.

- idf : Fréquence inverse des documents $idf_t = log \frac{N}{df_t}$

Si un terme est fréquent dans un grand nombre de documents, il n'est pas **discriminatoire**.

- *tf_idf:* le schéma de pondération attribue au terme t un poids dans le document d

$$tf\_idf_{t,d} = tf_{t,d} \times idf_t \qquad (1)$$

Maintenant, nous pouvons voir chaque document comme un vecteur dont chaque composante correspond à un terme du dictionnaire, avec un poids pour chaque composante qui est donné par l'équation (1).

Localiser les documents pertinents

- Nous pouvons traiter une requête comme un document très court.

Utiliser la mesure de similarité/distance pour trouver des documents similaires/pertinents.

Classer les documents en fonction de leur pertinence/similarité.

- Distance euclidienne (exemple de distance euclidienne) :

■ is defined as:

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- Cosinus de l'angle entre les vecteurs représentant le document et la requête

Les documents " dans la même direction " sont étroitement liés.

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

recherche web 1/2

La recherche sur le web trouve son origine dans la recherche d'informations (IR)

L'unité d'information est un document, et une grande collection de documents est disponible pour former la base de données textuelle.

Un crawler Web (robot Web) est un programme ou un script automatisé qui parcourt le Web de manière méthodique et automatisée.

Applications telles que l'intelligence économique, la surveillance des sites web et des pages d'intérêt, et les moteurs de recherche.

Types de crawlers

- universels
- préférentiels(ciblés) utilise la similarité et la classification

recherche web 2/2

Les opérations d'un moteur de recherche sont

Parsing: Analyse syntaxique

Indexing:

Searching and Ranking:

1. Prétraitement des termes de la requête
2. Trouver les pages qui contiennent les termes de la requête
3. Classer les pages et les renvoyer à l'utilisateur.

Classer les pages : in-links, Occurrence Type, Count, Position:,,

indexation sémantique latente (LSI) 1/2

La représentation de l'espace vectoriel souffre de son incapacité à traiter deux problèmes : la synonymie et la polysémie.

L'analyse sémantique latente (LSA) est une technique de traitement du langage naturel pour extraire et représenter la signification contextuelle des mots par des calculs statistiques appliqués à un grand corpus de texte. L'idée sous-jacente est que la totalité des informations sur tous les contextes de mots dans lesquels un mot donné apparaît et n'apparaît pas fournit un ensemble de contraintes mutuelles qui déterminent en grande partie la similitude de sens des mots et des ensembles de mots entre eux.

LSA utilise une matrice terme-document qui décrit les occurrences des termes dans les documents, généralement, tf-idf

LSA transforme la matrice d'occurrence en une relation entre les termes et certains concepts, et une relation entre ces concepts et les documents.

indexation sémantique latente (LSI) 2/3

Réduction du rang

LSA permet de trouver une matrice de rang plus faible, qui donne une approximation de la matrice des occurrences (fusionner les termes de sens proches.)

On effectue alors une décomposition en valeurs singulières sur X, qui donne deux matrices orthonormales U et V et une matrice diagonale Σ. $X = U\Sigma V^T$

Lorsqu'on sélectionne les k plus grandes valeurs singulières, ainsi que les vecteurs singuliers correspondants dans U et V, on obtient une approximation de rang k de la matrice des occurrences

Pour comparer une requête q dans l'espace des concepts il faut la traduire dans l'espace des concepts

$$\hat{q} = \Sigma_k^{-1} U_k^T q$$

indexation sémantique latente (LSI) 3/3

Applications de LSA

- la comparaison de documents dans l'espace des concepts (classification, clustring, …)
- la recherche de documents similaires entre différentes langues, (en utilisant un dictionnaire)
- la recherche de relations entre les termes (résolution de synonymie et de polysémie)
- traduire les termes de la requête dans l'espace des concepts, pour retrouver des documents liés sémantiquement (IR)
- trouver la meilleure similarité entre petits groupes de termes, de façon sémantique

extraction automatique de thèmes

est une technique d'apprentissage automatique qui organise et comprend de grandes collections de données textuelles, en attribuant des "balises" ou des catégories en fonction du sujet ou du thème de chaque texte.

L'extraction automatique de sujets peut être utile à des applications telles que l'exploration intelligente du Web, l'analyse de motifs topiques, l'extraction d'opinions, le résumé de résultats de recherche et le filtrage du spam sur le Web.