Recherche d'Information

Cours « Recherche d'Information et Graphes de Données »

Nicolas Delestre, Nicolas Malandain

INSA ROUEN NORMANDIE

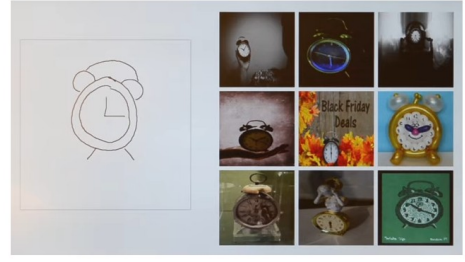
Plan...

INSA ROUEN NORMANDIE

Contexte 1 / 2

Moteur de recherche

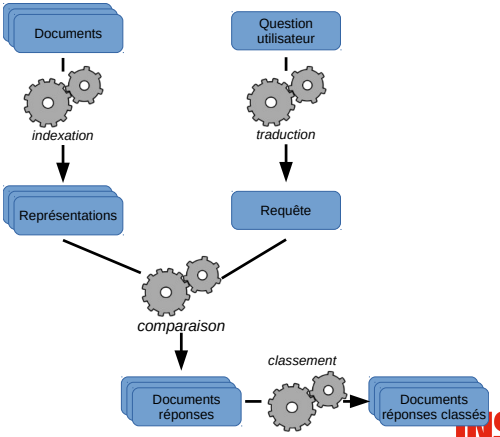
- Type document : texte
  - sonore :
    - exact : Shazam, SoundHound
    - approché : SoundHound, midomi
  - image (dessin ou photo)
- Sur le Web
- Corpus généraliste
- Indexation automatique



Contexte 2 / 2

Deux phases

- Phase d'indexation
  - Parcours du corpus
  - Création pour chaque document d'une représentation manipulable
- Phase de requêtage
  - Création d'une représentation de la question utilisateur
  - Sélection des documents « pertinents »
  - Classement de ces documents
- Chaque représentation nécessite des traitements du texte à indexer



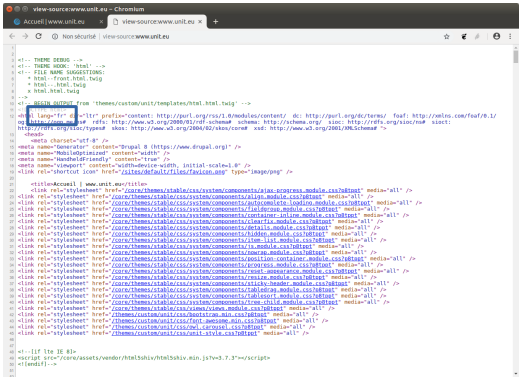
Du texte au mots

Pré-traitements principaaux avant l'indexation

- Identification de la langue
- Segmentation
- Normalisation
- Filtrage

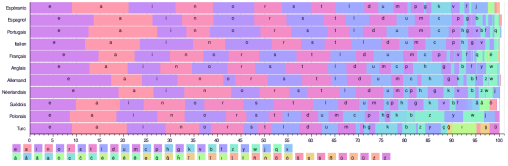
Identification de la langue

- Information présente sous forme de métadonnées
- Inféré à partir de statistique
  - Taux compression
  - Calculer la fréquence d'apparition de n-gram (succession de n caractères) : méthode tolérante aux fautes d'orthographe
- Difficultés lorsque le document
  - est court
  - est multilingue
  - contient du texte qui n'est pas de la langue



Identification de la langue

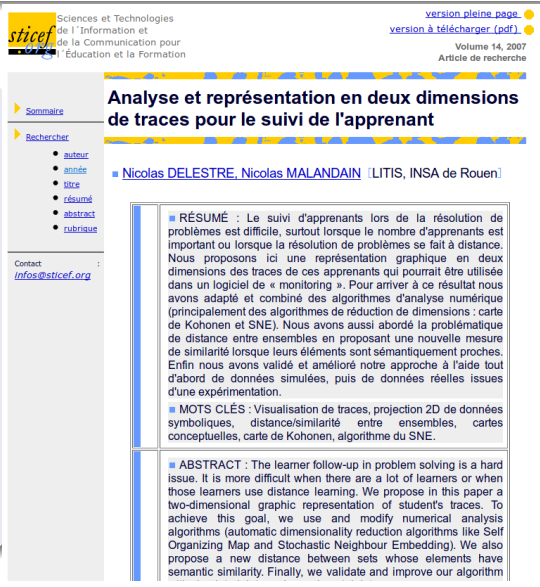
- Information présente sous forme de métadonnées
- Inféré à partir de statistique
  - Taux compression
  - Calculer la fréquence d'apparition de n-gram (succession de n caractères) : méthode tolérante aux fautes d'orthographe
- Difficultés lorsque le document
  - est court
  - est multilingue
  - contient du texte qui n'est pas de la langue



https://fr.wikipedia.org/wiki/Fréquence_d'apparition_des_lettres_en_français

Identification de la langue

- Information présente sous forme de métadonnées
- Inféré à partir de statistique
  - Taux compression
  - Calculer la fréquence d'apparition de n-gram (succession de n caractères) : méthode tolérante aux fautes d'orthographe
- Difficultés lorsque le document
  - est court
  - est multilingue
  - contient du texte qui n'est pas de la langue



sticef — Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation

version pleine page
version à télécharger (pdf)
Volume 14, 2007
Article de recherche

Sommaire
Rechercher
- auteur
- année
- titre
- résumé
- abstract
- rubrique

Contact : infos@sticef.org

Analyse et représentation en deux dimensions de traces pour le suivi de l'apprenant

Nicolas DELESTRE, Nicolas MALANDAIN [LITIS, INSA de Rouen]

RÉSUMÉ : Le suivi d'apprenants lors de la résolution de problèmes est difficile, surtout lorsque le nombre d'apprenants est important ou lorsque la résolution de problèmes se fait à distance. Nous proposons ici une représentation graphique en deux dimensions des traces de ces apprenants qui pourrait être utilisée dans un logiciel de « monitoring ». Pour arriver à ce résultat nous avons adapté et combiné des algorithmes d'analyse numérique (principalement des algorithmes de réduction de dimensions : carte de Kohonen et SNE). Nous avons aussi abordé la problématique de distance entre ensembles en proposant une nouvelle mesure de similarité lorsque leurs éléments sont sémantiquement proches. Enfin nous avons validé et amélioré notre approche à l'aide tout d'abord de données simulées, puis de données réelles issues d'une expérimentation.

MOTS CLÉS : Visualisation de traces, projection 2D de données symboliques, distance/similarité entre ensembles, cartes conceptuelles, carte de Kohonen, algorithme du SNE.

ABSTRACT : The learner follow-up in problem solving is a hard issue. It is more difficult when there are a lot of learners or when those learners use distance learning. We propose in this paper a two-dimensional graphic representation of student's traces. To achieve this goal, we use and modify numerical analysis algorithms (automatic dimensionality reduction algorithms like Self Organizing Map and Stochastic Neighbour Embedding). We also propose a new distance between sets whose elements have semantic similarity. Finally, we validate and improve our algorithm

Identification de la langue

- Information présente sous forme de métadonnées
- Inféré à partir de statistique
  - Taux compression
  - Calculer la fréquence d'apparition de n-gram (succession de n caractères) : méthode tolérante aux fautes d'orthographe
- Difficultés lorsque le document
  - est court
  - est multilingue
  - contient du texte qui n'est pas de la langue

Dès lors, l'algorithme d'apprentissage d'une carte de Kohonen devient un algorithme d'initialisation (Cf. figure 12).

Données en entrée :

X : les cartes conceptuelles d'apprentissage (de l'enseignant)

N : les neurones

Données en sortie :

W : les prototypes de la carte de Kohonen

début

Initialiser les W_i avec $\emptyset$

$W_{N(X_i)} \leftarrow \{X_i\}, i \in [1..|X|]$

Pour chaque neurone $n \notin N$ **faire**

Calculer les cartes influentes $C_i \subset X$ avec $i > 0$

Calculer les attributs att de l'ensemble des cartes C_i

Calculer le nombre d'attributs nb_{att} des cartes pour W_n

$W_n \leftarrow$ l'ensemble des cartes générées à partir de att et nb_{att}

fin

Figure 12 • Phase d'initialisation d'une carte de Kohonen de cartes conceptuelles

Précisons quelques points de cet algorithme :

- pour un neurone *n* de coordonnées *(i,j)*, les cartes d'influence sont les cartes se trouvant dans le cercle de centre *(i,j)* et de rayon *r*. Ce rayon est par défaut fixé au nombre maximal d'attributs que possèdent les cartes d'apprentissage,

Segmentation

Découper le texte en mots

- utiliser les séparateurs de mots
  - espaces, ponctuations, apostrophe
    - l'enfant
    - aujourd'hui
  - trait d'union
    - Mont-Saint-Michel
    - qu'en est-il ?
- Cela peut être difficile avec des langues acceptant des mots composés
  - « Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz »
  - « loi sur le transfert des obligations de surveillance de l'étiquetage de la viande bovine »

En recherche d'information, les moteurs de recherche comme Google ou Bing indexent des milliards de documents. En 2022-2023, ces systèmes ont évolué vers des approches basées sur l'IA (deep-learning, LLMs, etc.).

⇓

□

Normalisation

Objectifs

- Simplifier l'indexation
- Représenter tous les mots dans des formes plus simples
- Deux types de normalisation :
  - textuelle
  - linguistique
- Deux types de normalisation linguistique :
  - racinisation : mot → racine. Par exemple les mots *motoriser*, *motrice*, *motoriste*, *motricité* ont pour racine *moteur*
  - lemmatisation : mot → représentation encyclopédique

Normalisation textuelle + lemmatisation

Thomas avait choisi Me Dupont pour le défendre Mais depuis son procès il n a plus confiance dans son avocat il est véreux

⇓

Thomas avoir choisir Me Dupont pour le defendre mais depuis son proces il ne avoir plus confiance dans son avocat il etre véreux

Filtrage

Objectifs

- Simplifier l'indexation
- Supprimer les mots les plus fréquents de la langue, car supposer peu significatifs
- Mots les plus fréquents de la langue française :

1-10	de, la, le, et, les, des, en, un, du, une
11-20	que, est, pour, qui dans, a, par, plus, pas, au
21-30	...

Thomas avoir choisir Me Dupont pour le defendre mais depuis son proces il ne avoir plus confiance dans son avocat il etre véreux

⇓

Thomas choisir Me Dupont defendre proces confiance avocat véreux

INSA ROUEN NORMANDIE

Indexation

Rappel, deux étapes

1. Parcours du corpus
2. Représentation du document après l'application des algo. de TAL

Parcours

- Un programme nommé « Robot » part de pages référencées
- Il analyse le code HTML pour en extraire
  - Le contenu textuel du document (+ métadonnées) → représentation du document
  - Les liens vers les autres documents → continuer le parcours

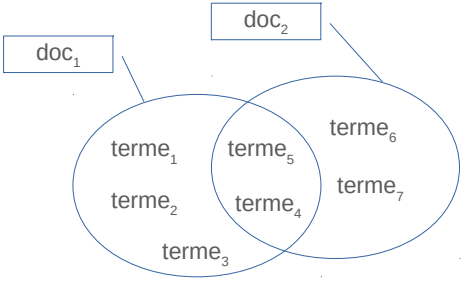
Deux types de représentation

1. Représentation vectorielle : chaque document est représenté par un vecteur mathématique (cf. cours sur les *embeddings*)
2. Représentation ensembliste : chaque mot est représenté par un ensemble de documents

Représentation ensemblistes 1 / 2

Constat

- Document = Suite de termes normalisés
- Création d'un index
  - Termes sont associés aux documents
  - Possibilité d'ajouter des informations aux termes
- Problèmes
  - On ne recherche pas les termes qui appartiennent aux documents
  - On cherche des documents qui contiennent certains termes



doc_1 : $terme_1$ (pos_1, pos_2, ...)
$terme_2$ (...)
$terme_3$ (...)
$terme_4$ (...)
$terme_5$ (...)

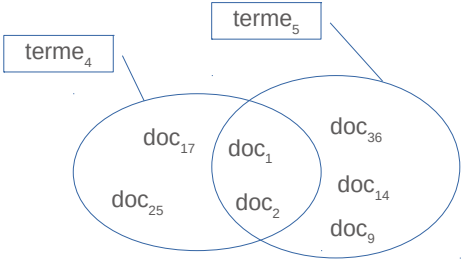
doc_2 : ...

INSA ROUEN NORMANDIE

Représentation ensemblistes 2 / 2

Index inversé

- Les documents sont associés aux termes
- Opérations ensemblistes :
  - Intersection : documents qui partagent certains termes
  - Union : documents qui contiennent certains termes
  - Soustraction : documents qui ne contiennent pas certains termes
- Avantage
  - L'ajout d'un nouveau document ne pose pas de problème



terme$_4$: doc$_1$ (pos$_1$, pos$_2$, ...)
doc$_2$ (...)
doc$_{17}$ (...)
doc$_{25}$ (...)
terme$_5$: ...

INSA ROUEN NORMANDIE

Phase de requétage

Rappels

- Création d'une représentation de la question utilisateur
- Sélection des documents « pertinents »
- Classement de ces documents

Question → requête

- Requête :
  - Représentation de la question de l'utilisateur
- La question peut être :
  - Une phrase en langue naturelle
  - Une suite de mots clés
  - Une expression booléenne de mots clés
- La requête est une expression booléenne de termes normalisés
  - Par défaut utilisation de l'opérateur *et*

Sélection des documents pertinents 1 / 2

Représentation vectorielle

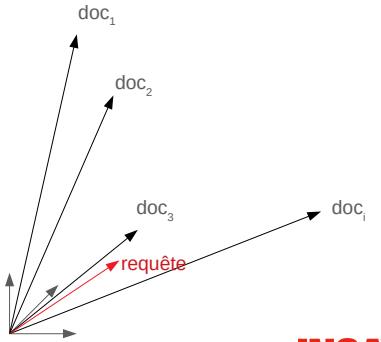
- Documents et requêtes représentés par un vecteur mathématique
- Comparaison et classement : suivant une similarité entre la requête et les documents, souvent similarité cosinus

Avantages

- Modèle homogène
- Classification des documents possibles

Inconvénients

- Nombreux calculs
- Conjonction uniquement



Sélection des documents pertinents 2 / 2

Représentation ensembliste

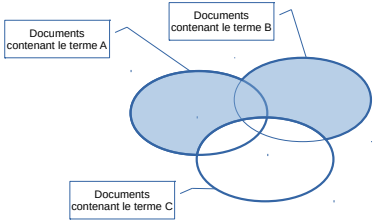
- Comparaison :
  - Traduction des opérateurs booléens de la requête par des opérations ensemblistes :
    - Et → intersection
    - Ou → union
    - Non → Soustraction

Avantages

- Opérations très efficaces
- Permet d'ajouter facilement de nouveau documents

Inconvénients

- Pas de classement



Classement dans la représentation ensembliste 1 / 3

Historiquement

- Classement fonction du calcul d'un score qui prend en compte les termes de la requête dans les documents :
  - nombre d'occurrences
  - positions (titre, sous-titre, etc.)

Page Rank

- Classement indépendant de la requête
  - Peut être calculé en « tâche de fond » et régulièrement
- Tire parti de la topologie du graphe du Web
- Fonction de la popularité des documents du corpus
  - Plus il y a des liens qui pointent sur un document d plus d est populaire
  - Plus les documents d_i qui référencent d sont populaires, plus d est populaire

Classement dans la représentation ensembliste 2 / 3

Principe de l'algorithme du *page rank*

- Imaginer un « surfeur » éternelle qui parcourt le Web. La notoriété d'une page p est le nombre de fois que ce surfeur a parcouru la page p (cf. [Abi12])
- Dans les faits, les n documents d_i et les liens forment un graphe orienté :
  1. Calcul de la matrice M de transferts (matrice d'adjacence tel que les poids d'un lien allant de d_a à d_b est $1/|arcsSortant(d_a)|$. La matrice est de dimension $n \times n$
  2. Initialisation d'un vecteur N_0 de taille n représentant la notoriété de chaque document (initialisation aléatoire ou équiprobable)
  3. Mise à jour du vecteur N ($N_{i+1} \leftarrow N_i M$) jusqu'à un point fixe ($N_{i+1} - N_i < \epsilon$)

INSA ROUEN NORMANDIE

Classement dans la représentation ensembliste 3 / 3

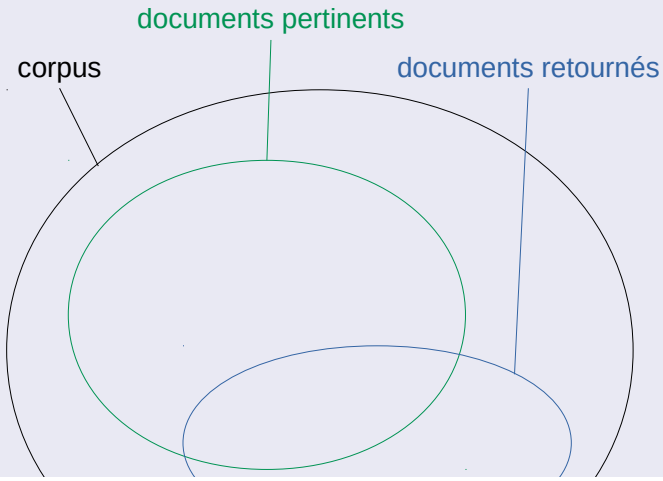
Un classement objectif ?

- Le classement effectué par certains moteurs de recherche est aussi fonction :
  - du contenu des documents (en dehors du texte)
  - de votre profil
  - des enjeux commerciaux

INSA ROUEN NORMANDIE

Évaluation de la recherche d'information 1 / 4

Ensembles pour la mesure de la qualité



Évaluation de la recherche d'information 2 / 4

Résultats non ordonnés : deux couples d'indicateurs

1. Précision, rappel :

$$précision = \frac{|Nb\ documents\ pertinents\ retournés|}{|Nb\ documents\ retournés|}$$

$$rappel = \frac{|Nb\ documents\ pertinents\ retournés|}{|Nb\ documents\ pertinents|}$$

2. Bruit, Silence :

$$bruit = \frac{|Nb\ documents\ non\ pertinents\ retournés|}{|Nb\ documents\ retournés|}$$

$$silence = \frac{|Nb\ documents\ pertinents\ non\ retournés|}{|Nb\ documents\ pertinents|}$$

Évaluation de la recherche d'information 3 / 4

Résultats non ordonnés : F-Mesure

- Très souvent l'amélioration d'un indicateur est au dépend de l'autre :
  - Lorsque l'on essaye de réduire le bruit, le silence grandit
  - Lorsque l'on essaye d'augmenter la précision, le rappel diminue
- F-mesure
  - Critère de qualité utilisant le couple rappel-précision Moyenne harmonique :

$$F_1 = 2 \times \frac{precision \times rappel}{precision + rappel}$$

INSA ROUEN NORMANDIE

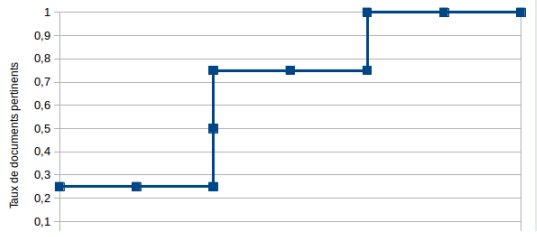
Évaluation de la recherche d'information 4 / 4

Résultats ordonnées

1. La courbe ROC (*Receiver Operating Characteristics*) :
   - Taux de documents non pertinents en abscisse (bruit), Taux de documents pertinents en ordonnée (rappel)
   - « Cette courbe se construit en estimant le taux de documents pertinents à chaque rang, en fonction du taux de documents non pertinents [à ce même rang] » [**?**]
   - Non sensible au ratio des deux classes [**?**]

Exemple avec un corpus de 10 documents [?]

Rang	Pertinence	Taux doc. non pert.	Taux doc. pert.
1	1	0	1/4
2	0	1/6	1/4
3	0	1/3	1/4
4	1	1/3	1/2
5	1	1/3	3/4
6	0	1/2	3/4
7	0	2/3	3/4
8	1	2/3	3/4



De la recherche documentaire à la recherche d'information

Constat

- Que contiennent les documents sur le Web ? des informations
- Mais pourquoi utilisons nous un moteur de recherche ?
  - Pour avoir une réponse à une question, pour obtenir une information
  - Plutôt que de retourner des documents qui contiennent « peut-être » la réponse à la question posée, il faudrait que les moteurs de recherche retournent l'information désirée
  - C'est ce qu'essaye de faire :
    - Certains moteurs de recherche pour des questions simples
    - Les assistants personnels (Google Home, Siri, Alexa, etc.)
- Difficulté supplémentaire : les informations ne sont pas seulement dans le texte des documents, mais aussi dans des tableaux, figures, etc.

Compétition SQuAD 1 / 2

The Standford Question Answering Dataset

- https://rajpurkar.github.io/SQuAD-explorer/
- La version 2.0 : 150000 questions/réponses sur plus de 500 extrait d'articles Wikipédia anglophone
- Corpus
  - Public : texte + questions/réponses + un script d'évaluation
  - Privé : questions/réponses pour évaluer les contributions
  - Réponse = terme(s) d'un article

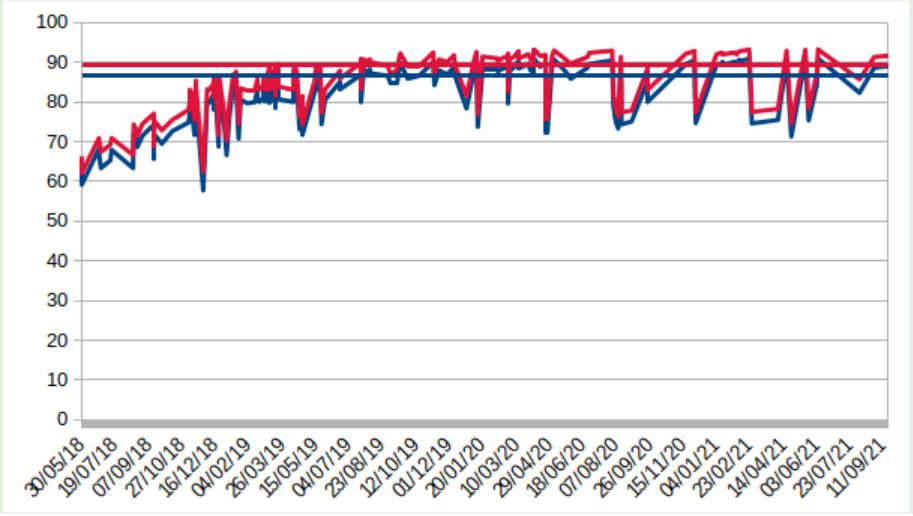
ex : Normans

The Normans (Norman : Nourmands ; French : Normands ; Latin : Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia. Through generations of assimilation and mixing with the native Frankish and Roman-Gaulish populations, their descendants would gradually merge with the Carolingian-based cultures of West Francia. The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries.

- In what country is Normandy located ? France
- What century did the Normans first gain their separate identity ? the first half of the 10th century, the 10th century, the 10th

Compétition SQuAD 2 / 2

Résultats en mars 2022



Limite des approches classiques de la RI

Problèmes

- Compréhension de la langue naturelle
  - Synonymie, polysémie
  - Ambiguïtés syntaxiques et sémantiques
- Problèmes liés au corpus
  - Bruit dans les documents
  - Documents non textuels
- Problèmes liés à l'utilisateur
  - Formulation de la requête
  - Subjectivité de la pertinence

INSA ROUEN NORMANDIE

Conclusion

Résumé

- La recherche d'information est un domaine en constante évolution
- Les approches classiques restent utilisées dans de nombreux moteurs de recherche
- Les avancées en TAL et en IA ouvrent de nouvelles perspectives pour améliorer la recherche d'information : pour représenter le sens, raisonner mais aussi générer des réponses