



REDS MÉTHODOLOGIE RECHERCHE

Modèles

Wednesday 6th December, 2023

Laure Soulier



Topologie des contributions scientifiques

Topologie des contributions scientifiques

■ Proposition de modèle

- Modèle de recommandation
- Modèle de prédiction
- Modèle de recherche d'information
- ...

→ répond à une problématique

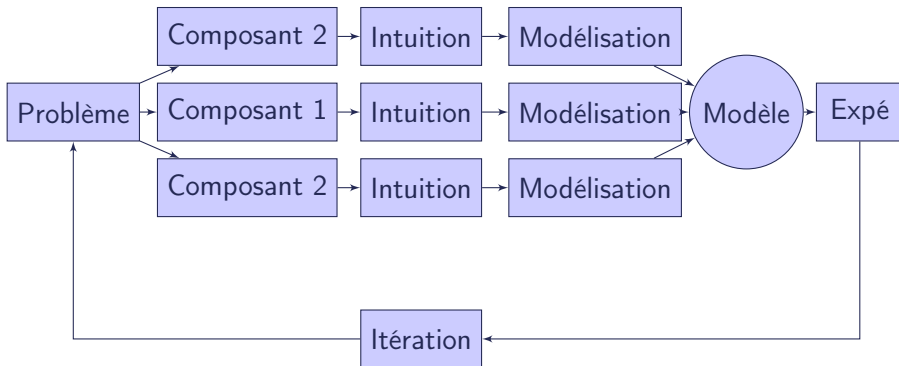
→ basé sur des intuitions et des hypothèses

→ doit raconter une "histoire"

Processus de "création" d'un modèle

Définition

Modèle : Formalisation mathématique qui répond à une problématique



Processus de "création" d'un modèle

On procède souvent de façon itérative :

- mettre en place des modèles simples (baselines)
- vérifier que le modèle surapprend sur des petites collections et des exemples simples (sanity check)
- étendre à des collections plus grandes, rajouter des contraintes
- rajouter les composants au fur et à mesure, sinon difficile de savoir ce qui fonctionne ou pas

En pratique...

- Modification de la fonction de coût
 - Exemple :

by minimizing the classification loss on the training data:

$$\sum_{i \in \mathcal{T}} \Delta(f_{\theta}(z_i), y_i) \quad (3)$$

where $\Delta(f_{\theta}(z_i), y_i)$ is the loss to predict $f_{\theta}(z_i)$ instead of the real label y_i , and \mathcal{T} is the training set.

In order to make sure that connected nodes have similar representations, we introduce the other following loss:

$$\sum_{i,j:w_{i,j} \neq 0} w_{i,j} \|z_i - z_j\|^2 \quad (4)$$

which forces the approach of the latent representation of connected nodes. The complete loss function is the aggregation of the classification and similarity loss:

$$L(z, \theta) = \sum_{i \in \mathcal{T}} \Delta(f_{\theta}(z_i), y_i) + \lambda \sum_{i,j:w_{i,j} \neq 0} w_{i,j} \|z_i - z_j\|^2 \quad (5)$$

Figure 1: [Nozza et al., ICSEFM 2014]

En pratique...

- Pré-traitement/Encodage des données
 - Exemple : Extraction de relations dans un texte avec des réseaux de neurones

Texte brut	→ sparsité
Word embedding + POS	→ prise en compte de la sémantique et de la grammaire

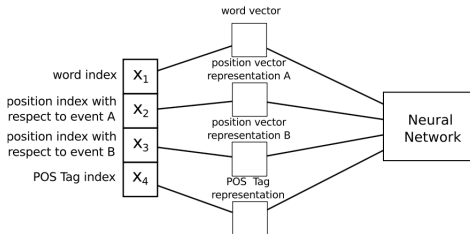


Figure 2: [Gerald, 2016]

En pratique...

- Modification d'une architecture
 - Prise en compte de la "connaissace globale" issue des textes dans un LSTM

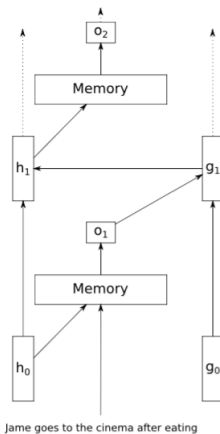


Figure 3: [Gerald, 2016]

En pratique...

- Transfert de méthodologie d'un domaine
 - Exploiter les réseaux de neurones en RI

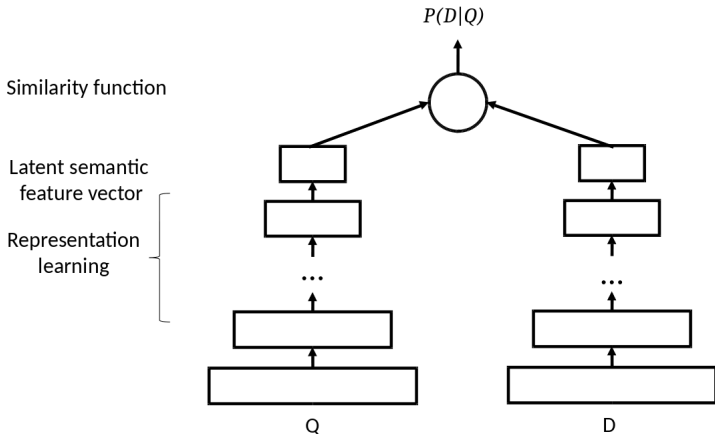


Figure 4: [DSSM - 2015]

Topologie des contributions scientifiques

■ Proposition de modèle

■ Analyse empirique

- Comparaison des différents modèles de recommandation
- Analyse comparative des architectures neuronales pour une problématique de classification
- Analyse de l'impact de l'âge sur le comportement de recherche des utilisateurs

→ Bien définir la problématique analysée

→ Identifier les questions de recherche

→ Bien définir le protocole. S'il est faux, toutes les analyses sont fausses !!

→ Définir des analyses (statistiques, erreurs, modèles qui supportent les hypothèses, ...)

→ Dégager des conclusions et des "take away messages" / "research implications" pour le design de futurs modèles

Topologie des contributions scientifiques

- Proposition de modèle
- Analyse empirique
- Prototypage
 - Moteurs de recherche
 - Interface de prédiction

→ Un prototype répond à un besoin : l'identifier

→ Un prototype permet de mettre en oeuvre un modèle : identifier l'innovation

→ Peut être mis en avant par un processus d'évaluation pour montrer l'amélioration des interactions homme-machine

Topologie des contributions scientifiques

- Proposition de modèle
- Analyse empirique
- Prototypage
- Jeux de données
 - Données labellisées (classification, nettoyage, etc...)
 - Logs de comportements utilisateurs

→ Identifier la nouveauté : tâche, population étudiée, taille plus importante, nouveau domaine, ...

→ Démontrer que le protocole de construction est valide : références de papiers, expérimentations qui valident les hypothèses, suivi des règles éthiques pour des expérimentations utilisateurs, ...

→ Fournir un cadre d'expérimentation pour les recherches futures : définition d'une tâche, de métriques, de baselines, ...

→ Montrer les perspectives d'utilisations : différentes tâches possibles, contextes d'expérimentation, ...

Topologie des contributions scientifiques

- Proposition de modèle
- Analyse empirique
- Prototypage
- Jeux de données
- Synthèse de l'état de l'art (généralement en fin de thèse)

→ bien plus difficile que ce qu'il paraît, souvent plus que l'état de l'art fait dans une thèse - sauf si état de l'art inexistant pour le domaine

→ nécessité d'organiser les travaux (pas une liste sans lien)

→ trouver un positionnement par rapport aux états de l'art existants : sciences cognitives/modèles/expérimentations, apport d'une techno dans le domaine (e.g., non supervisé, GAN, ...), papiers perspectives : que sera l'avenir dans le domaine,

A vous de jouer !

Travail à faire

- Choisir une problématique (ou en proposer une) dans votre sujet
- Développer une méthode ou améliorer un algorithme pour résoudre cette problématique (contribution modèle)
- Mettre en place le protocole de développement du modèle