Fouille de Données Textuelles

Indexation Automatique

Huaizhong KOU

Responsable: Prof. Amedeo NAPOLI

ORPAILLEUR, Laboratoire LORIA, Nancy

Plan

I. Fouille de Données Textuelles(FDT)
II. Indexation Automatique

Partie I. Fouille de Données Textuelles(FDT)

- Scénario
- Problématiques
- Qu'est ce que c'est FDT?
- FDT vs Data Mining

Scénario

- Informations textuelles volumineuse
- Croissance Phénoménale
- Connaissances riches codées dans des textes par des auteurs
- Gestion de connaissances nécessaire
  - Manière traditionnelle infaisable
  - Acquisition automatique vois le jour

Problématique

- Textes libres en langage naturel
- Facile de lire et comprendre pour des lecteurs
- MAIS difficile d'identifier des connaissances y cachées et codées par les auteurs pour applications.

Qu'est ce que c'est FDT?

- Ensemble de techniques permettant de fouiller des textes libres pour découvre des connaissances cachées.
- BUT: Analyser des histoires comportementales pour aider la gestion des connaissances et prévoir l'avenir.
- Branches principales
  - Extraction d'information
  - Résumé de textes
  - Groupement de textes
  - Catégorisation de textes
  - etc.

Extraction d'information

- Processus permettant d'identifier des informations avec des catégories prédéfinies à partir de textes
  - nome de personne, lieu, organisation, morceaux d'informations pertinentes à un domaine d'application,...
- BUT: transformer des informations en forme structurée et rendre ses sémantiques accessibles aux applications

(suite)

Extraction d'information (suite)

- Exemple: analyse des textes abordant des évènements de terroriste.

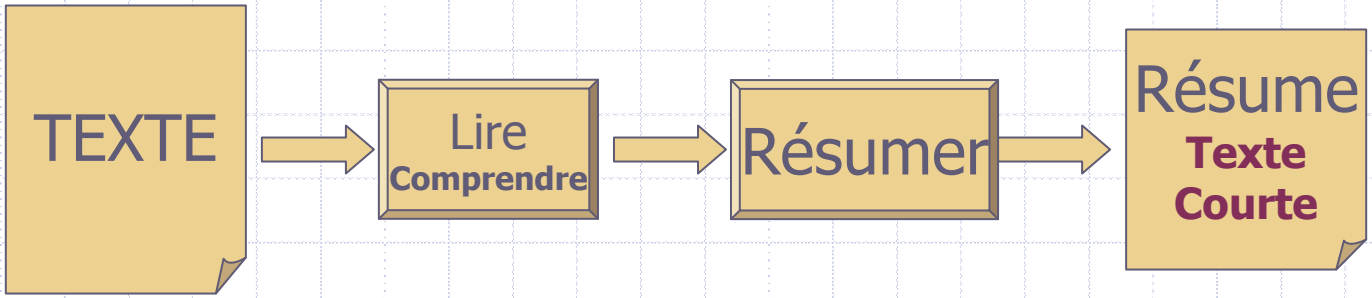
« the <victim> *U.S. embassy* </victim> in <lieu> *Bogota* </lieu> was bombed <date> *yesterday* </date> by <auteur> *FMLN gurrillas* </auteur> »

DATE	AUTEUR	VICTIME	LIEU
21/09/1990	FMLN gurrillas	U.S. embassy	Bogota

- Techniques: étiquetage POS, analyse partial de phrase, interprétation sémantique, analyse de discours,...
- Application: support de recherche information, etc

Résumé de textes

- Processus essayant de comprendre des textes et générer de nouvelles ***textes courtes*** pour synthétiser des sujets principaux abordés.



- BUT: Aider aux hommes à comprendre des textes efficacement sans les lire complètement et soulager des acquisitions d'informations souhaitées.

(suite)

Résumé de textes(suite)

- En théorie des connaissances linguistiques et syntaxiques et lexicales et spécifiques à un domaine nécessaires
- **MAIS DIFFICILE** de les coder dans des applications.
- Techniques répandus: statistiques
  - Identifier des phrases représentatifs
    - Mesurer des pertinences
      - Fréquences, positions et etc.
- Application: Alertes d'informations vers PDA, Portable et etc.

FDT vs Data Mining(DM)

	Data Mining	FDT
Objet	numérique & catégorique	textuel
Structure	structuré	non-structuré
Représentation	simple	complexe
Dimension	<dizaine milles	>dizaine milles
Maturité	Implémentation vaste dès 1994	Implémentation vaste dès 2000

Part II. Indexation Automatique (INDA)

- Que-ce que c'est INDA?
- Architecture d'INDA
- Unité d'index
- Construction de vocabulaire d'index
- Modèles de représentation
- Structure des fiches inversées

Que-ce que c'est INDA?

- Processus automatique permettant de sélectionner des identités correspondant aux mots représentative dans des textes et de les représenter.
- Étape de base pour analyser et traiter des textes.
- Exemple:

« l'association Bernard Grégory lance la campagne 2004 du programme " valorisation de compétence " pour aider les doctorants en fin de thèse. le dossier doit être visé par votre directeur de thèse et le directeur de l'école doctorale. »

« associ, lance, campagne, programme, valorise, doctorant, thèse, dossier, viser, directeur»

Architecture d'INDA

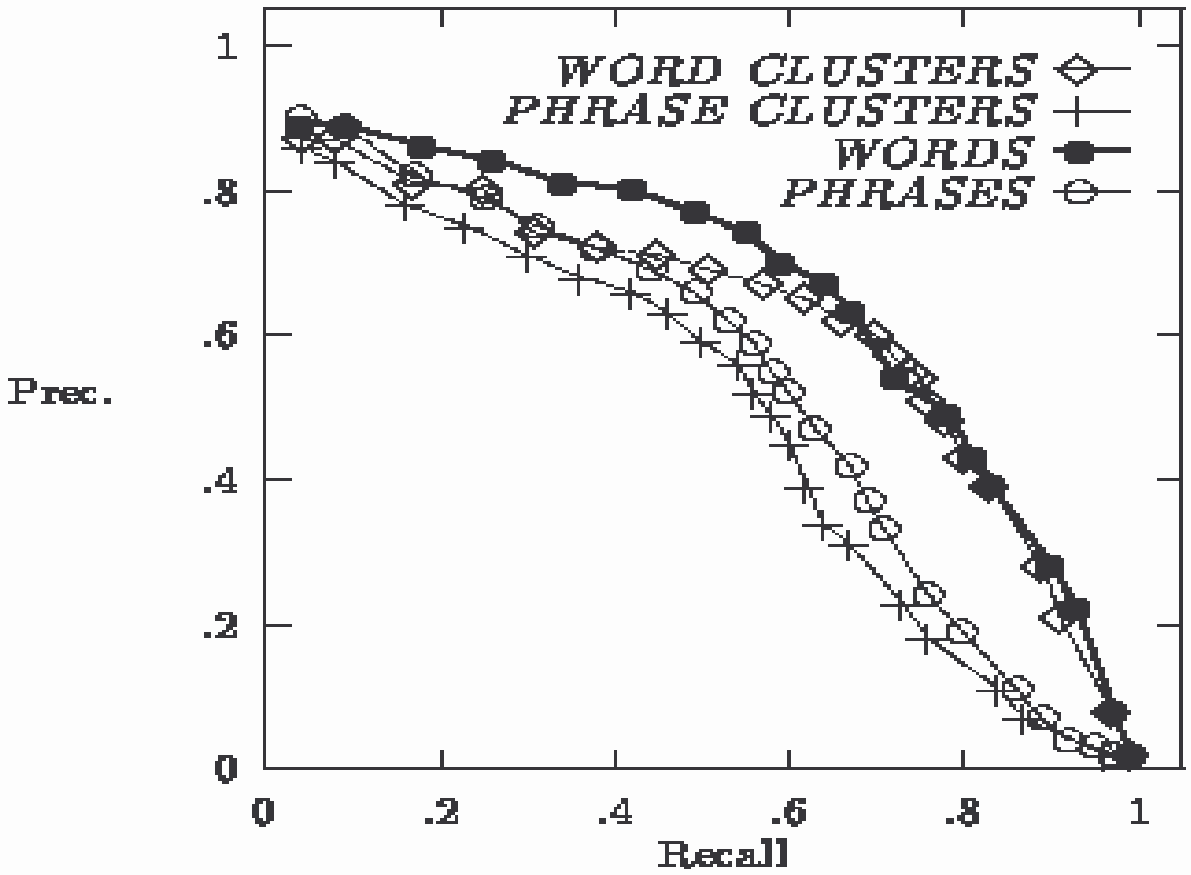
- Pré-traitement
  - Identifie des mots individus
  - Enlève des mots vides (sur, le, la, est, ont, dans...)
  - Enlève symboles non-alphabétiques (#,*,-,_,&,\,...)
  - Mettre des lettres en minuscule(« Bière » → « bière »)
  - D'autres canonisations
- Calcul de poids
  - Mesure des importances des mots dans des textes
- Représentation

Unité d'index

- Mots simples et phrases
  - Au cas du recherche d'information (RI)
    - L'utilisation de phrases améliore des performance de 5% à 10% par rapport à celles de mots simples
  - Au cas du catégorisation de textes (CT)
    - Mots simples marches bien pour CT
    - L'utilisation de phrases dégrade parfois des performances

(suite)

Unité d'index(suit)



Phrases dégradant des performances CT

D.D.Lewis,1992

(suite)

Unité d'index(suite)

- Lemmatisation (stemming)
  - Processus automatique permettant d'enlever récursivement suffixes et/ou préfixes des mots et de sortir ses racines.
    - dictionnaire des suffixes: (ly, ness, ion, ant, ent,ical, able, ary, ence,ing, etc.)
    - dictionnaire des préfixes: (anti,bi, co, contra, intra, micro,mini,pro,semi,tri, etc.)
    - dictionnaire des exceptions
    - règles morphologiques: « hopping →hope » à lieu de « hopp »
  - mots → racines

Unité d'index(suite)

- Lemmatisation (stemming)
  - Algorithmes Porter et Lovins
  - Exemple:
    - connection→connect, connected→connect, connecting→connect, etc.
- Hypothèse: des sens des mots avec même racine sont proches.
- Organise des mots en différents groupes dont des mots ont des même racines.
- Racines comme unités d'index
- Avantages:
  - Réduit la taille de vocabulaires d'index
  - Augmente des capacités représentative de termes d'index
  - Améliore des rappels.

(suite)

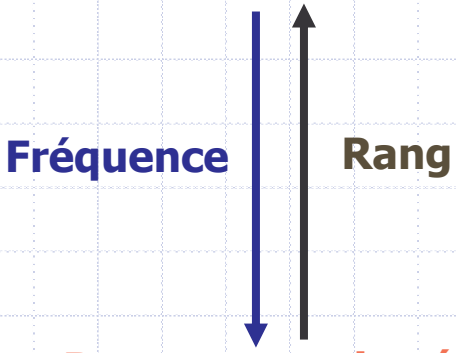
Unité d'index(suite)

- Thésaurus: organise des mots en différant classes de la manière que des mots dans même classe ont même sens par rapport à un domaine donné.
  - identités des classes comme unités d'index.
  - génération automatique ou manuelle
- Exemple

Identité de classe	entré	Identité de classe	entré
760	(permission)	761	(prohibition)
	permission		prohibition
	leave		veto
	sanction		disallowance
	allowance		injunction
	tolerance		ban
	authorization		taboo
	warrant		

Construction de vocabulaire d'index

- Fréquences
  - Fréquence de terme (tf)
    - Nombre d'occurrence de terme dans des textes
    - Mesure des importances des termes dans un texte
  - Fréquence de document de terme (df)
    - Nombre des textes contenant un terme
    - Mesure des capacités des termes pour différencier des textes
  - Règle Zipf: fréquence•rang ≈ constant



(suite)

Construction de vocabulaire d'index(suite)

- Poids des termes dans des textes
  - Mesure d'importance des termes pour représenter des textes et différencier l'un de l'autre.
  - tf-idf formule

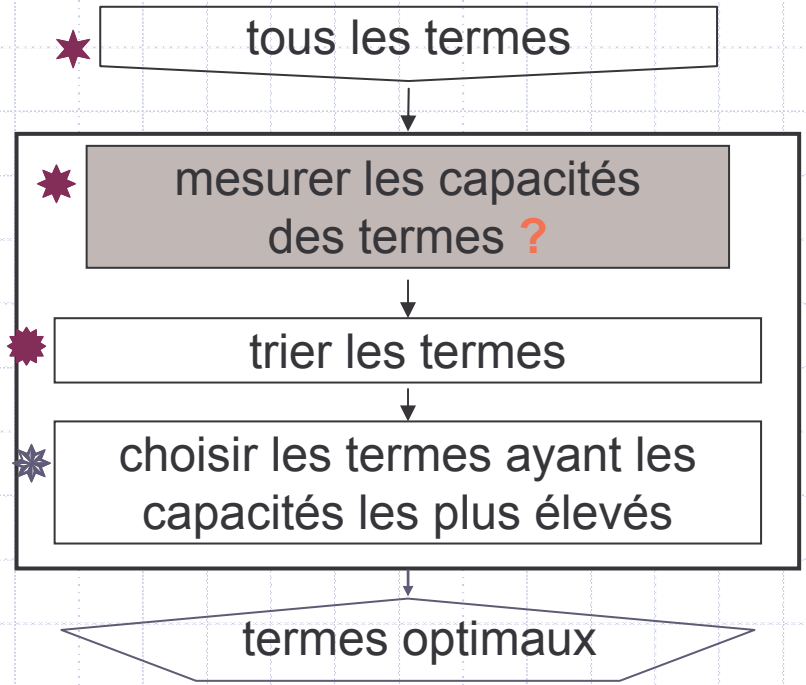
$$w_{ij} = tf_{ij} * \left\lfloor \log\left(N / df_j\right) + 1 \right\rfloor$$

- Sélection de termes
  - Trop de termes pour l'applications
    - Dizaine de milles pour un corpus modéré
    - Emploie de tous les termes n'améliore pas les performances
  - BUT: sélectionner un sous-ensemble optimal des termes capable de représenter des contenus abordés dans des textes.

(suite)

Construction de vocabulaire d'index(suite)

- Sélection de termes
  - Méthode de filtrage



(suite)

Construction de vocabulaire d'index(suite)

- Sélection de termes pour IR
  - À l'aide des fréquences tf ou df
    - Élimine des termes avec les rangs très élevés et très peu élevés
    - Garde des termes avec les rangs modérés
- Sélection de termes pour CT
  - Mesures des capacités des termes pour identifier des catégories
    - Gain informationnel (IG)
    - Information mutuelle (MI)
    - CHI-test ($\chi 2$): performant
    - Approches basant sur concept et indépendance (CBA et IBA) (thèse)

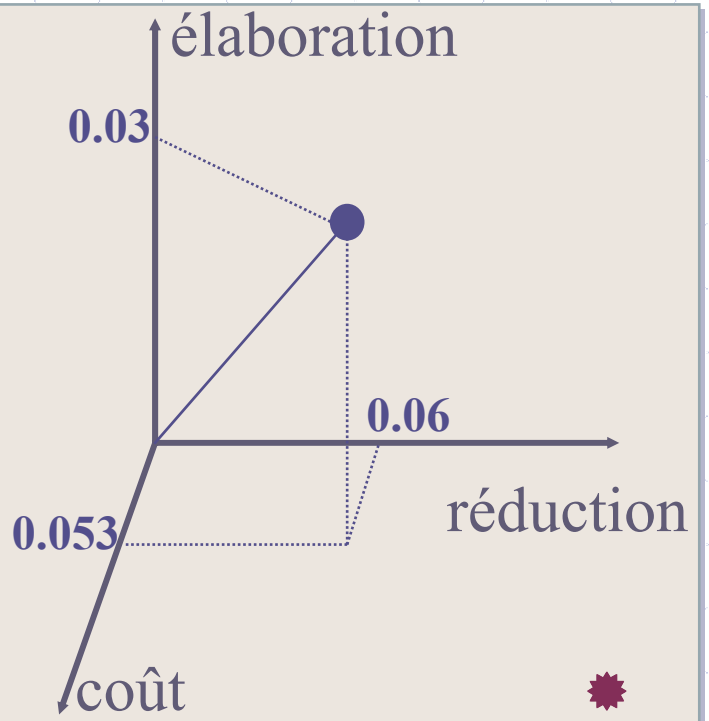
Modèles de représentation

- Modèle vectoriel
  - Ensemble de termes et aucune structure considérée
  - Coordonnées, poids de termes et formules *tf-idf*

$$d_i = (w_{i1}, w_{i2}, \ldots, w_{il})$$

la direction d'Airbus a confirmé les information contenues dans le Financial Times mentionnant l'élaboration d'un plan de réduction de coûts...

(airbus 0.13, coût 0.053, direction 0.02, élaboration 0.03, information 0.01, plan 0.1, réduction 0.06, ...)



(suite)

Modèles de représentation(suite)

- Modèle vectoriel
  - Modèles de similarités
    - Distance Euclidienne

$$simil(d_1, d_2) = \|d_1 - d_2\|^2$$

    - Fonction cosinus

$$simil(d_1, d_2) = \frac{d_1 \bullet d_2}{\|d_1\|^2 \times \|d_2\|^2}$$

    - Produit scalaire

$$simil(d_1, d_2) = d_1 \bullet d_2$$

  - Association de termes
    - Modèle mathématique(thèse)

(suite)

Modèles de représentation(suite)

- Modèle probabiliste: Génération des textes
  - Modèle multi-variable
    - document := un évènement;
    - termes := attributs;
  - Modèle multi-nominal
    - document:= un ensemble d'évènements;
    - termes:= évènements

Structure des fiches inversées

- Structure orientée vers des termes
  - Pour chaque terme dans le vocabulaire, une liste des pairs de (texte, position) dont le texte contient le terme
  - Pour RI traditionnel

terme1	txt_{i1}, $poid_{i1}$; txt_{j2}, $poid_{j2}$;…; txt_{in}, $poid_{in}$
terme2	txt_{k2}, $poid_{k2}$; txt_{s1}, $poid_{s1}$;…; txt_{rl}, $poid_{rl}$
.	.....
.	.....
.	.....

Merci