

Analyse de données textuelles
Panorama des fonctions, des méthodes et des usages

Sylvie Dalbin

Assistance & Techniques Documentaires

DocForum, Le 17 Novembre 2005

Déroulé de l'intervention (1)

1. Définition et historique

2. Des usages qui se multiplient

3. Fonctions et techniques de traitement

4. Compétences et perspectives

Analyse de données textuelles (ADT) De quoi parle-t-on ?

Démarches, méthodes et techniques
offrant une **vision globale** d'un corpus
textuel, et mettant en exergue des **relations**
entre les données, relations jusque-là non
repérées et potentiellement utiles

- **Produire de la connaissance à partir de données brutes**

Ancrage historique

1. Les bases -

Avant - 1969 - 1970

- ◆ Avant l'informatique
- ◆ 1960-1969 : Sociologie des sciences
- ◆ 1970: Mathématiques
Traitement automatique des langues

2. Développements « localisés »

Avant - 1970 - 1990

- ◆ 1970 : Informatique décisionnelle > Data mining
- ◆ 1979-1990 : Sociologie des sciences > Text mining

3. Déploiements - Arrivée du Web

Depuis 1990

- ◆ Datamining and knowledge discovery
- ◆ Composants logiciels

Evolution terminologique

❑ Extraction, découverte de connaissances dans des bases de données (KDD)

- IA et apprentissage automatique
- Incorporer les connaissances du domaine

❑ Data mining

- Statisticiens, analystes de données
- Exploitation d'algorithmes
- Analyse statistique exploratoire des données multidimensionnelles, numériques, qualitatives et textuelles

✓ ***Text mining***

Constats

- ❑ Les informations (masse, flux) sont sous-exploitées et difficiles à traiter
- ❑ L'accès à l'information « à l'unité » est insuffisant - analyse inter-documents
- ❑ Les activités fortement liées à l'information sont complexes et « mangeuses de temps »

✓ ***Des outils d'aide***

✓ ***Des produits d'information synthétiques***

Déroulé de l'intervention (2)

1. Définition et historique

2. Des usages qui se multiplient

3. Fonctions et techniques de traitement

4. Compétences et perspectives

Usages

- Finalités d'actions variées
 - ◆ Evaluation et analyse prospective
- Spécificité des indicateurs à étudier et des ressources à exploiter
- Des usages différenciés - tout secteur
 - ◆ IA, Stratégie, Gestion des connaissances, Qualité (dans son travail), Aide à la recherche d'information

Déroulé de l'intervention (3)

1. Définition et historique
2. Des usages qui se multiplient
3. Fonctions et techniques de traitement
4. Compétences et perspectives

Analyse de données textuelles

3 fonctions clés

Fonctions-clés		Traitements
Modéliser	1	Normaliser et nettoyer Traitements linguistiques Etablissement d'une représentation modélisée
Faire émerger du sens activer le processus de fouille	2	Fouille, extraction, traitements. A plat, dans le temps
Visualiser et interpréter	3	Graphes et cartographies

Technologies mises en oeuvre

1. Traitements linguistiques et sémantiques

- ◆ Passer des chaînes de caractères - mots aux notions
- ◆ Prendre en compte des usages et des contextes (ressources terminologiques)

2. Analyse de données, statistiques

- ◆ Mettre en relief les structures pertinentes d'un ensemble de données

Technologies mises en oeuvre

3. Technique de classification automatique

- ◆ Organiser des masses de documents
Explorer et mettre en valeur des contenus

4. Techniques de schématisation - cartographie

- ◆ Produire une représentation visuelle, synthétique
> aide à l'analyse et à la compréhension de phénomènes

5. Approche « modélisation » de situation

- ◆ Prise en compte d'un contexte précis

Evolution des modes de représentation

☐ Mots clés

- ◆ Indicateur de contenu d'un document

☐ Classes

- ◆ Indicateurs de thèmes ou centres d'intérêts

☐ Cartes

◆ Indicateurs de relations

- Positionnement dans l'espace (noeuds et liens) : Avec ou Sans signification
- Valeur d'un noeud : un document, regroupement de documents, regroupement d'information

Apports de ces méthodes et techniques

- En amont, structurer sa pensée
 - ◆ Identifier les éléments à traiter, choisir les modes de représentation
- Support de réflexion et de questionnement
 - ◆ Explorer, faire émerger et rendre visible
 - *Mais fouiller n'est pas trouver*
- En aval, support d'idées, de présentation
 - ◆ outil de communication
 - ◆ Un dessin plutôt qu'un rapport
 - *Mais une carte n'est pas le territoire*

Limites et biais

- ❑ Constitution des corpus
- ❑ Traitements amont
- ❑ Techniques d'analyse
- ❑ Infographie ; lecture "erronée" de graphiques

Déroulé de l'intervention (4)

1. Définition et historique
2. Des usages qui se multiplient
3. Fonctions et techniques de traitement
4. Compétences et perspectives

Des compétences nécessaires

- ❑ Intégrer les principes sous-jacents
- ❑ Interpréter les résultats
- ❑ Comprendre l'ensemble des techniques mathématiques mises en œuvre

✓ ***S'approprier, faire sien les résultats***

Perspectives

- ❑ Consolider les techniques existantes
- ❑ Prendre en charge toutes les ressources utiles
- ❑ Etendre les usages