Recherche et Extraction d'Information
Généralités

Mathieu Roche

Cours ECDA

2014/2015

Plan

- **Motivations**
  - Veille technologie
  - Les défis
- **Les méthodes en RI**
  - Généralités
  - Les limites des approches actuelles
  - Approches pour améliorer les résultats
- **Evaluation en RI**

Pourquoi la RI ?
Un exemple : la veille technologique (1/7)

- **Définition :**

La **veille technologique** est l'art de **repérer, collecter, traiter, stocker** des informations et des signaux pertinents (faibles, forts) qui vont permettre d'orienter le futur (technologique, commercial, etc.) et également de protéger le présent et l'avenir face aux attaques de la concurrence
[Rouach 96, Que sais-je ?]

Pourquoi la RI ?
Un exemple : la veille technologique (2/7)

La veille marketing : recueillir, sélectionner, traiter et diffuser des informations sur les produits et marchés.

De manière plus concrète la veille marketing permet d'identifier :

- les évolutions du marché de l'entreprise,
- le comportement des consommateurs,
- les retombées d'une campagne de communication,
- etc.

Pourquoi la RI ?
Un exemple : la veille technologique (3/7)

La veille concurrentielle :

• se tenir informés des diverses **activités des concurrents** (dépôts de brevets, travaux de recherche, etc.), **des techniques de vente et de distribution des concurrents et leur politique de communication**.

• **détecter des savoir-faire** de certains confrères/concurrents et d'engendrer des coopérations potentielles fructueuses.

Pourquoi la RI ?
Un exemple : la veille technologique (4/7)

La veille sociétale (aussi appelée veille socio-politique ou veille environnementale) :

- Rechercher et traiter des renseignements relatifs aux **aspects socio-économiques, politiques, géopolitiques et socioculturels de la société**.
- Etudier, en particulier, l'**évolution des moeurs et des mentalités, les risques** (désordres, conflits, etc.), **les mouvements sociaux** et de **protestation**.

Pourquoi la RI ?
Un exemple : la veille technologique (5/7)

Utilisation des techniques de TAL pour la veille :

Des outils de TAL (Traitement Automatique du Langage) sont utilisés pour **analyser les données textuelles** (groupes de discussion, bulletins électroniques, articles économiques, articles scientifiques, journaux en ligne, etc.).

Exemple : extraire des informations et concevoir de manière automatique des formulaires à partir de dépêches d'actualités économiques

Pourquoi la RI ?
Un exemple : la veille technologique (6/7)

Extraction d'informations :

Dépêche économique

L'Europe donne son feu vert au rachat de Materis par Wendel. Annoncé début janvier, le rachat par Wendel Investissement de Matreis appartenant à LBO France s'élève à 1,01 milliard d'euros. Une transaction qui valorise Materis à environ 2 MdE. Si Wendel Investissement et Materis ne sont pas présentes sur les mêmes marchés, Materis achète certains services fournis par le Bureau Veritas qui appartient à Wendel. Bureau Veritas s'occupe du contrôle et de la certification de produits, de procédés et de projets.
31/03/2006

Pourquoi la RI ?
Un exemple : la veille technologique (7/7)

Extraction d'informations :

Trois étapes :

(a) **Analyser les textes** (analyse lexicale et syntaxique).

(b) **Extraire des éléments pertinents** dans les textes (noms de personnes, de société, de lieu, etc.).

(c) **Déterminer les relations entre ces éléments**

Les défis internationaux

- **MUC** : les conférences MUC (Message Understanding Conferences) permettent depuis une quinzaine d'années de confronter divers systèmes d'Extraction d'Informations autour de quatre tâches :
  - "Named Entity" : extraction des formes linguistiques contenant un nom propre ou un nombre (autre qu'un cardinal d'ensemble) pour faire référence à une entité.
  - "Coreference" : calcul de chaînes de référence.
  - "Template Element" : remplissage de fiches contenant pour chaque entité concernée une simple liste attributs-valeurs.
  - "Scenario Template" : remplissage de fiches contenant pour un type d'événement caractéristique du domaine traité la mise en relation entre ces événements et les entités impliquées.

- **TREC** (Text REtrieval Conference) : TREC Genomics, Novelty, Spams, etc.

Les défis francophones

- **DEFT (DEfi Francophone de Fouille de textes)** :
  - Identification d'auteurs
  - Segmentation thématique
  - Classification de textes d'opinion

- **EASY (Evaluation des Analyseurs SYntaxiques)**

... et des exemples d'autres tâches

- **Analyse d'articles scientifiques,**
- **Classification thématique, d'opinions,**
- **Constitution de dictionnaires (vocabulaire spécialisé, sigles, etc)**
- ***etc.***

Plan

- **Motivations**
  - Veille technologie
  - Les défis
- **Les méthodes en RI**
  - Généralités
  - Les limites des approches actuelles
  - Approches pour améliorer les résultats
- **Evaluation en RI**

Les méthodes de RI - Généralités

- **Types de méthodes :**
  - Statistiques
  - Linguistiques
  - Mixtes
- **Apprentissage supervisé / non supervisé**
- **Méthode sac de mots**

Les limites des méthodes de RI

- Limites liées aux langues étudiées
- Complexité du traitement du langage naturel (polysémie, traitement des anaphores, etc.)
- Quantité et qualité des données disponibles
- Qualité des systèmes de TAL

Les améliorations possibles

- Ajouter des **connaissances sémantiques** (généralistes, spécialisées, comment obtenir ces informations ?)
- Ajouter des **connaissances lexicales**
- Ajouter des **connaissances syntaxiques**
- Leur combinaison ?

Plan

- **Motivations**
  - Veille technologie
  - Les défis
- **Les méthodes en RI**
  - Généralités
  - Les limites des approches actuelles
  - Approches pour améliorer les résultats
- **Evaluation en RI**

Evaluation des méthodes de RI

- Notion générale de **précision** et de **rappel**

$$précision = \frac{\text{nombre d'exemples positifs couverts}}{\text{nombre d'exemple couverts}}$$

Une précision de 100% signifie que tous les exemples couverts sont positifs.

$$rappel = \frac{\text{nombre d'exemples positifs couverts}}{\text{nombre d'exemples positifs}}$$

Une couverture de 100% signifie que tous les exemples positifs sont couverts.

Evaluation des méthodes de RI

- **Mesure pour combiner Rappel et Précision :**

la F-mesure (ou F-score).

$$F = \frac{2 \cdot (\text{précision} \cdot \text{rappel})}{(\text{précision} + \text{rappel})}$$

- **Variantes de la F-mesure**

Evaluation des méthodes de RI

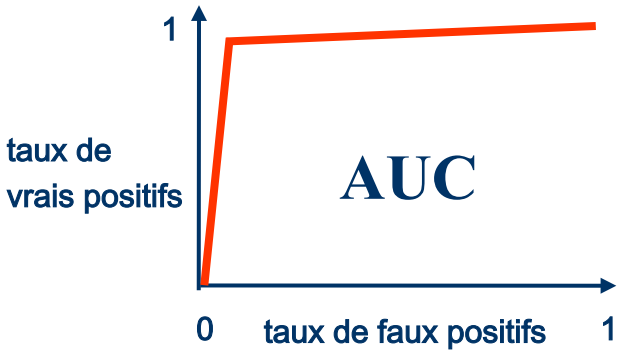
- **Front de Pareto (courbe Rappel/Précision) : plusieurs critères sont abordés**
- ➔ Le front de Pareto est défini par l'ensemble des approches telles qu'il n'existe pas une solution qui soit la meilleure pour tous les critères (ici précision et rappel).
- ➔ Les approches qui ne sont pas sur le front de Pareto sont dites "dominées".
- ➔ Classement possible par front de Pareto

Evaluation des méthodes de RI

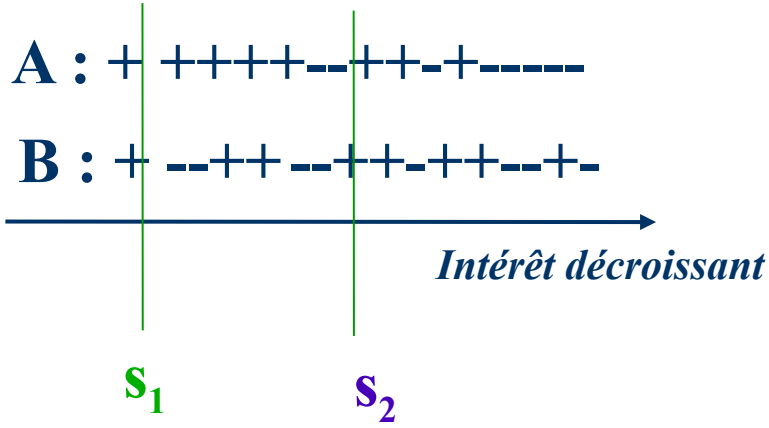
- **Evaluation des fonctions de rang**

- **Utilisation des courbes ROC (Receiver Operating Characteristic) : courbe dont le taux de vrais positifs est représenté en ordonnées et le taux de faux positifs est représenté par l'axe des abscisses**

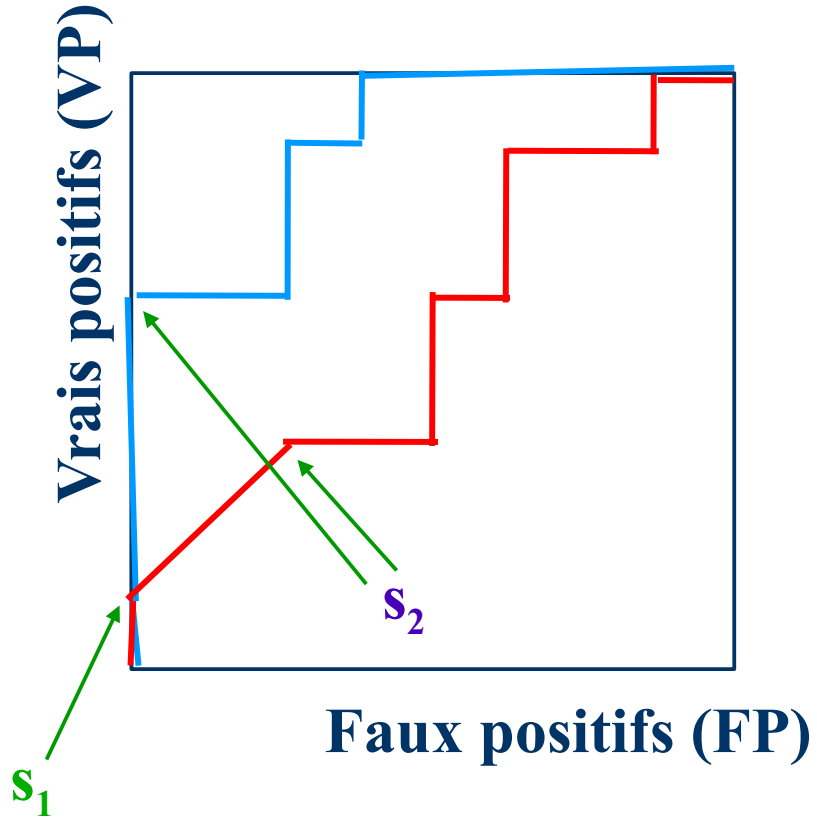
- ***Avantage :* pas de sensibilité dans le cas d'un déséquilibre entre les classes.**



Evaluation des méthodes de RI



	S_1	S_2
A	VP = 1/8 FP = 0	VP = 5/8 FP = 0
B	VP = 1/8 FP = 0	VP = 3/8 FP = 2/8



Evaluation des méthodes de RI

- **Evaluation des fonctions de rang**

Exercice : Soit les deux listes ci-dessous :

liste 1 : - - + + -

liste 2 : + - - - +

Quelle est la liste la plus pertinente selon le "critère AUC" ?

Evaluation des méthodes de RI

Et dans le cadre de la problématique de la classification de textes ?

Evaluation de classifieurs (1/3)

Évaluation du test : Matrice de confusion

Prédit \ Réel	Pos	Neg
Pos	TP	FP
Neg	FN	TN

- TP : True Positive
- FP : False Positive
- FN : False Negative
- TN : True Negative

Précision, Rappel, Accuracy

- $Precision = \frac{TP}{TP+FP}$
- $Rappel = \frac{TP}{TP+FN}$
- $Accuracy = \frac{TP+TN}{TP+FN+FP+TN}$

Evaluation de classifieurs (2/3)

Matrice de confusion multi-classes

Prédit \ Réel	C_1	C_2	...	C_i	...	C_n
C_1	c_1^1	c_1^2		c_1^i		c_1^n
C_2	c_2^1					
...	...					
C_i	c_i^1			c_i^i		
...	...					
C_n	c_n^1					

- Prédiction correcte : c_i^i
- Prédiction incorrecte : c_i^j avec $i \neq j$

Précision, Rappel, Accuracy

- $Precision(C_i) = \frac{c_i^i}{\sum_{j=1}^{n} c_i^j}$
- $Rappel(C_i) = \frac{c_i^i}{\sum_{j=1}^{n} c_j^i}$
- $Accuracy = \frac{\sum_{i=1}^{n} c_i^i}{\sum_{i,j=1}^{n} c_i^j}$

Evaluation de classifieurs (3/3)

- La validation croisée sert à estimer l'erreur réelle d'un modèle (**adaptée aux méthodes supervisées comme les KPPV**)

.

Validation croisée (S,x)

// S est un ensemble x est un entier

Découper S en x parties égales $\{S_1,...,S_x\}$

Pour i de 1 à x

Construire un modèle M avec l'ensemble S-Si

Evaluer l'erreur ei de M avec Si

Fin Pour

Retourner la moyenne des $e_i = \Sigma_{i=1..x} e_i/x$