Introduction à la recherche d'information

ISN 2 - 2017/2018

Lorraine Goeuriot (LIG, UGA)

lorraine.goeuriot@imag.fr
http://mrim.imag.fr/User/lorraine.goeuriot/isn2

Préambule

- Trois séances :
  - Séance 1 : cours d'introduction à la recherche d'information
  - Séances 2 et 3 : TP de mise en place d'un système simple de recherche d'information

Plan

1. Introduction
2. Elements clés en recherche d'information (RI)
3. Modèles de RI
4. Systèmes de recherche d'information (SRI)
5. Evaluation des SRI

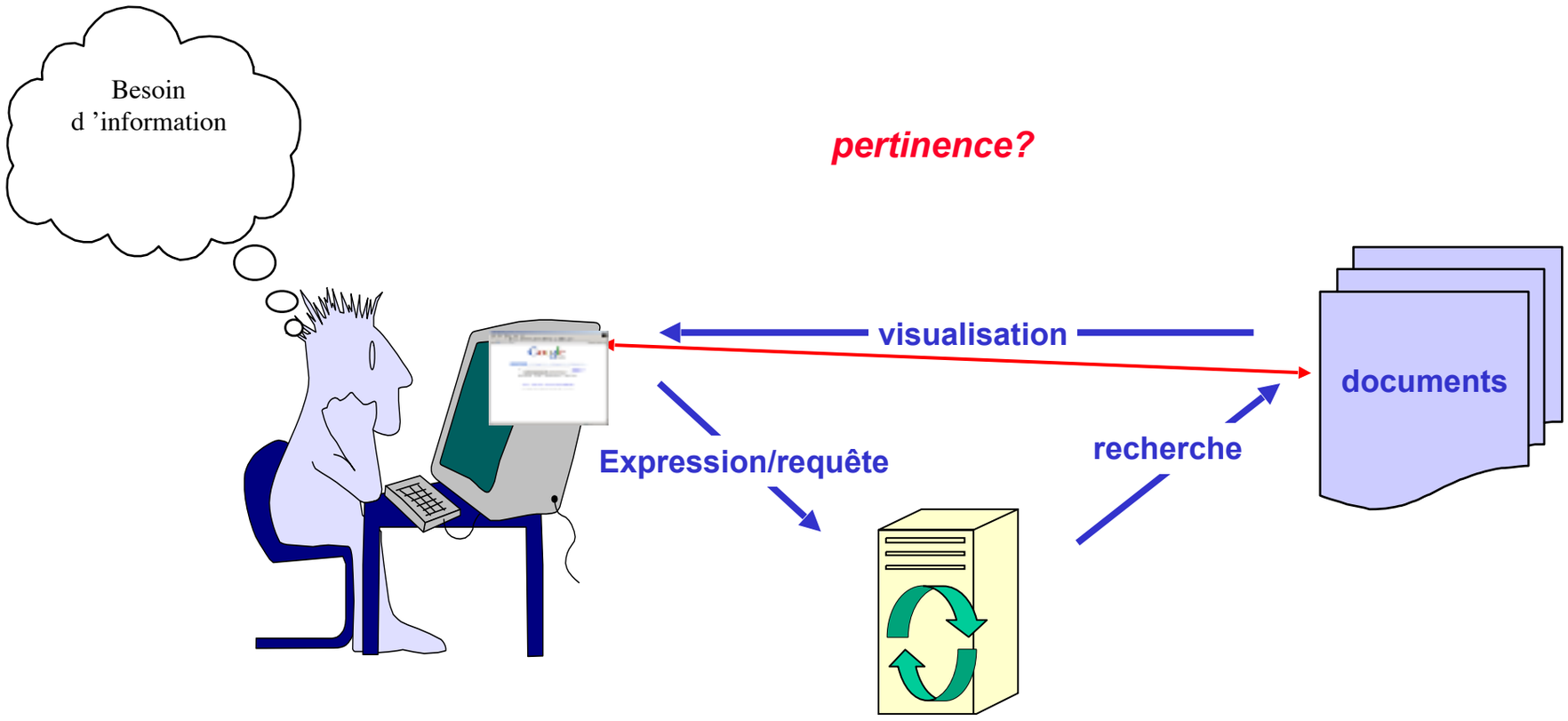
Plan

1. Introduction
2. Elements clés en recherche d'information (RI)
3. Modèles de RI
4. Systèmes de recherche d'information (SRI)
5. Evaluation des SRI

Introduction

Problématique de la RI :

- Accès par le contenu à des documents satisfaisant un besoin d'information d'un utilisateur

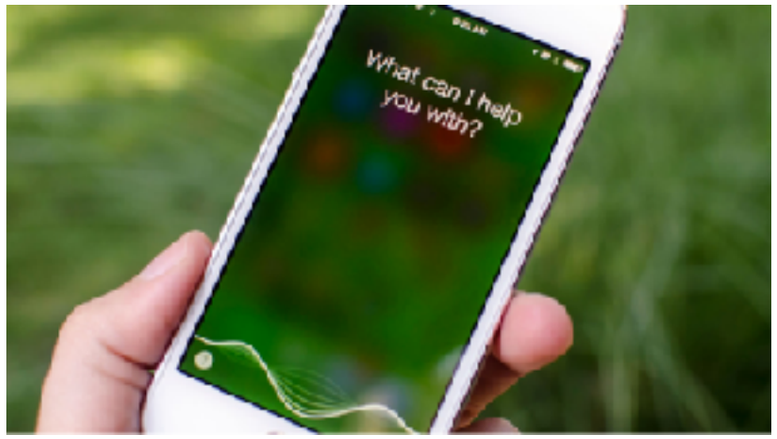


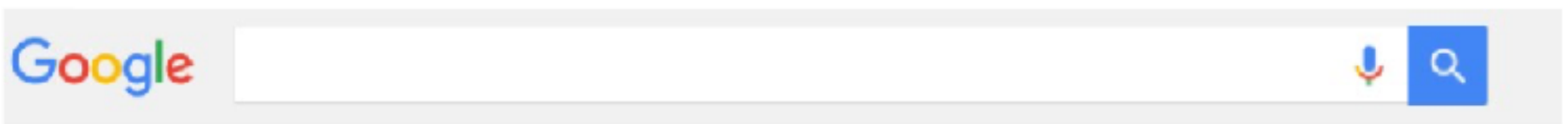
Introduction

Qu’est-ce que la recherche d’information ?

Introduction

Qu’est-ce que la recherche d’information ?





Introduction

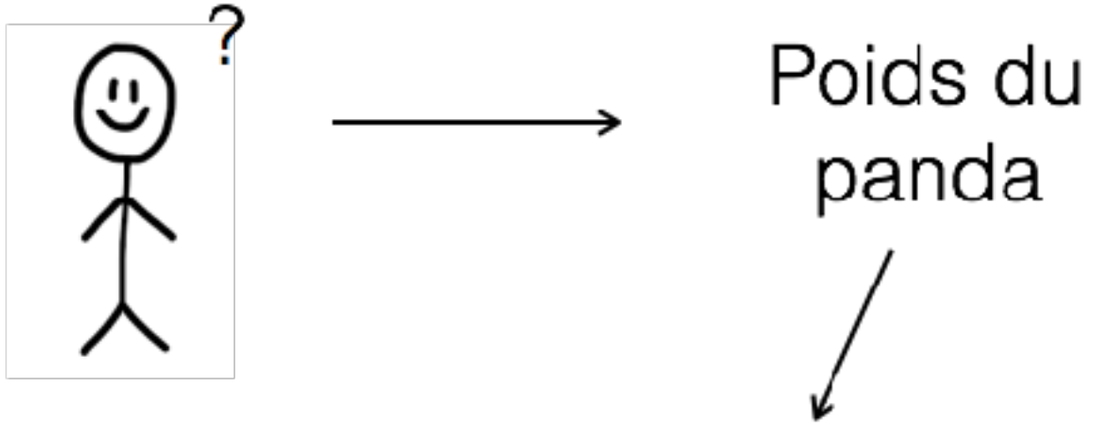
- Objectif : évaluer, comprendre, et améliorer le processus

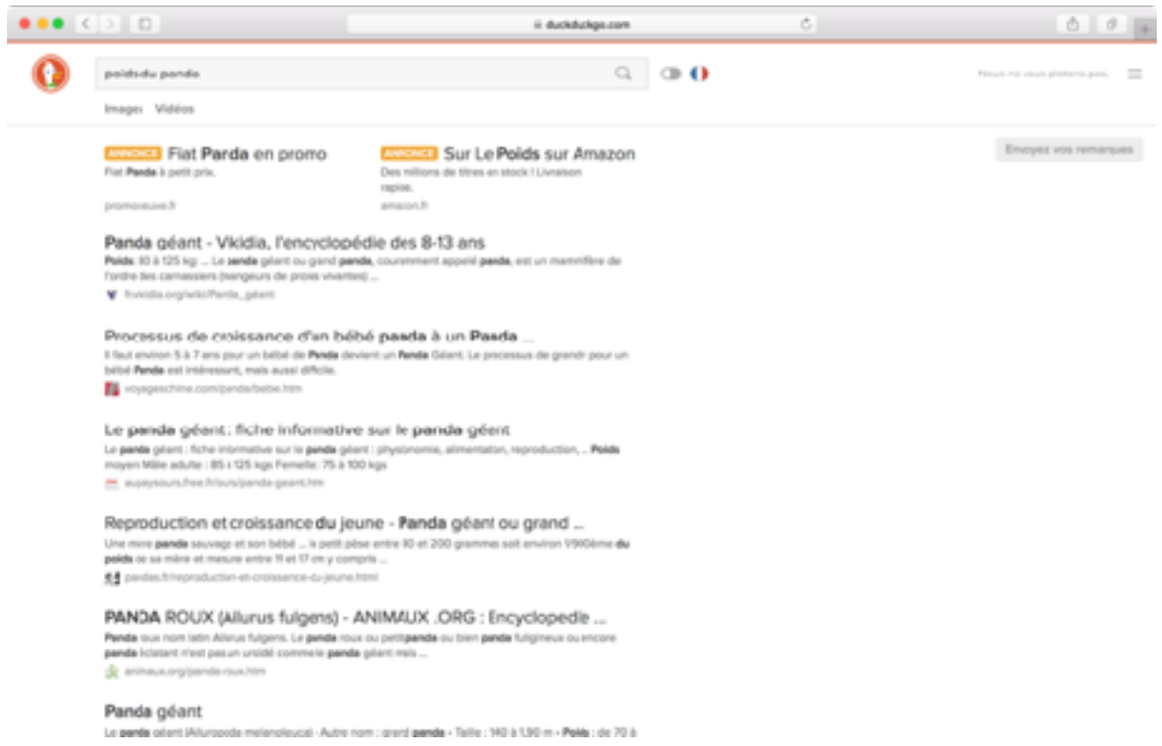
- Comment exprimer un besoin d’information ?

- Comment traiter la requête ?

- Comment chercher parmi une collection de documents ?

- Qu’est-ce qu’un bon résultat ?





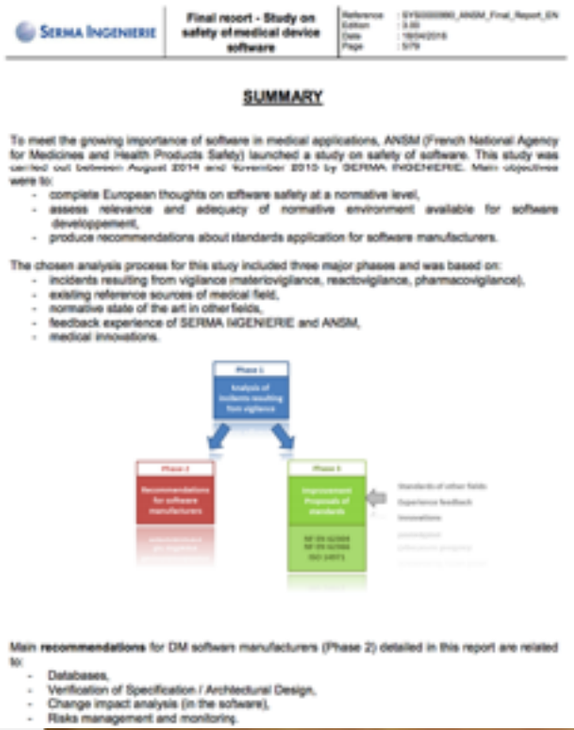
Plan

1. Introduction
2. Elements clés en recherche d'information (RI)
3. Modèles de RI
4. Systèmes de recherche d'information (SRI)
5. Evaluation des SRI

Eléments clés de la RI

- Documents
- Contenu des documents
- Besoin d’information d’un utilisateur
- Satisfaction

Eléments clés de la RI


SERMA INGENIERIE | Final report - Study on safety of medical device software

SUMMARY

To meet the growing importance of software in medical applications, ANSM (French National Agency for Medicines and Health Products Safety) launched a study on safety of software. This study was carried out between August 2014 and November 2015 by SERMA INGENIERIE. Main objectives were to:

- complete European thoughts on software safety at a normative level,
- assess relevance and adequacy of normative environment available for software development,
- produce recommendations about standards application for software manufacturers.

The chosen analysis process for this study included three major phases and was based on:

- incidents resulting from vigilance materiovigilance, reactovigilance, pharmacovigilance),
- existing reference sources of medical field,
- normative state of the art in other fields,
- feedback experience of SERMA INGENIERIE and ANSM,
- medical innovations.

Main recommendations for DM software manufacturers (Phase 2) detailed in this report are related to:

- Databases,
- Verification of Specification / Architectural Design,
- Change impact analysis (in the software),
- Risks management and monitoring.


Grenoble

45° 11' 16" N 5° 43' 37" E carte

Pour les articles homonymes, voir Grenoble (homonymie).

Grenoble (prononcé [gʁə.'nɔbl] Écouter, *Grenòble* en arpitan) est une commune du sud-est de la France, chef-lieu du département de l'Isère, ancienne capitale du Dauphiné. Troisième commune de la région Auvergne-Rhône-Alpes, derrière Lyon et Saint-Étienne, Grenoble est la commune-centre de la deuxième agglomération de cette région après celle de Lyon. Elle est également la plus grande métropole alpine, devant Innsbruck et Bolzano, ce qui lui vaut le surnom de « capitale des Alpes ».

L'histoire connue de Grenoble couvre une période de plus de deux mille ans. À l'époque gallo-romaine, le bourg gaulois porte le nom de Cularo, puis celui de Gratianopolis. Son importance s'accroît durant le XIe siècle lorsque les comtes d'Albon choisissent la cité comme capitale de leur province, le Dauphiné. Ce statut, consolidé par l'annexion à la France, lui permet de développer son économie. Grenoble devient alors une ville parlementaire et militaire, à la frontière avec la Savoie.

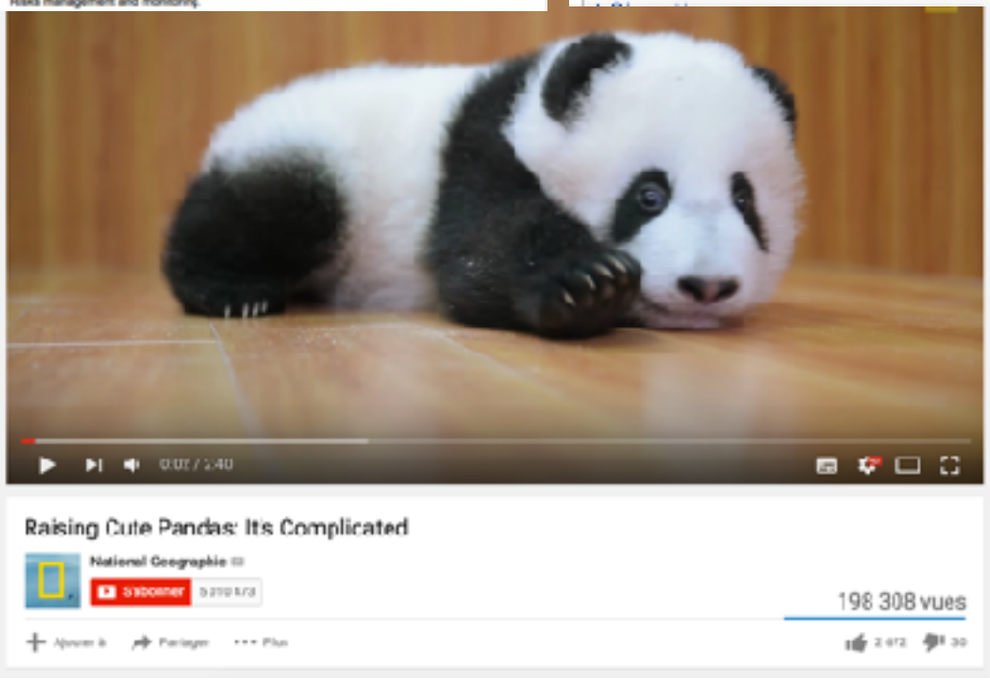
En 1788, la population s'est mobilisée pour défendre ses parlementaires lors de la Journée des Tuiles en attaquant les troupes royales. Antoine Barnave et Jean-Joseph Mounier (à l'origine du fameux Serment du jeu de Paume) furent d'illustres acteurs grenoblois de la Révolution française.

Le développement industriel de Grenoble commence véritablement au XVIIe siècle avec la ganterie et s'accentue dans la deuxième partie du XIXe siècle, avec la découverte de la houille blanche. Mais Grenoble connaît sa croissance la plus forte durant les Trente Glorieuses. La tenue des Jeux olympiques d'hiver symbolise cette période de grands bouleversements pour la ville. Son développement continuant, Grenoble s'affirme aujourd'hui comme un grand centre scientifique européen[1,2]. Pour ce qui est de la population, Grenoble était en 2013 la seizième commune de France avec 160 215 habitants[3], son agglomération la dixième de France en 2013, avec 506 201 habitants[4], et son aire urbaine la onzième, avec 684 366 habitants[5].

Ses habitants sont appelés les Grenoblois[6].

Sommaire [masquer]

Grenoble


0:07 / 2:40

Raising Cute Pandas: Its Complicated

National Geographic

S'abonner

198 308 vues


METEO A 7 JOURS | EVOLUTION METEO A 14 JOURS

Afficher les détails (précipitations, orage...) | Légende

dimanche 04	lundi 05	mardi 06	mercredi 07	jeudi 08	vendredi 09	samedi 10
12°C / 32°C	16°C / 28°C	13°C / 28°C	12°C / 30°C	13°C / 32°C	14°C / 30°C	14°C / 29°C
Éclaircies	Éclaircies	Ensoleillé	Ensoleillé	Ensoleillé	Éclaircies	Éclaircies
UV 6	UV 5	UV 6	UV 6			
Vent 10 km/h	Vent 10 km/h Raf. 40 km/h	Vent 6 km/h Raf. 40 km/h	Vent < 5 km/h	Vent 6 km/h CONFIANCE 4	Vent 5 km/h CONFIANCE 4	Vent 6 km/h CONFIANCE 3

17h 20h 23h 2h lundi 5h 8h 11h 14h 17h

Les documents

1. Catégories de documents

   – Différents médias : texte, image, vidéo, documents structurés…

   – Différents types d'information par média

   - Texte : livre, article, lettre

   - Image : images par rayons X, photographies, graphiques

2. Classes d'information

   - Méta-Information (information à propos du document) : titre, auteur, date de création, etc.

   - Contenu brut : le document initial (contenu symbolique : information extraite du contenu brut)

Besoin d'information de l'utilisateur

Exprimé sous la forme de requêtes pouvant suivre un langage fixé

- Contraintes sur les méta informations :
    - Attributs : « Article écrit par Alim-Louis Benabid »
    - Structure : « article médical contenant une image de CT-scan »
- Contraintes sur le contenu
    - Contenu brut : « document avec le texte "hépatomégalie" » :
    - Contenu symbolique : « documents portant sur les effets secondaires de l'aspirine »

Besoin d'information de l'utilisateur

- Types de requêtes :
  - Navigation (*site de l'UGA*)
  - Service (*le livreur de pizza le plus proche*)
  - Information (*combien d'heures par jour dort un chat*)
- Complexité :
  - Ambiguïté (UGA : *Université Grenoble Alpes*, *Ouganda*, *club de foot UGA Décines*…)
  - Précision (*restaurant Grenoble* vs *restaurant Grenoble ouvert dimanche végétarien pas cher*)
  - Vocabulaire (*changer huile voiture* vs *vidange*)

Besoin d’information de l’utilisateur

Syntaxe des requêtes :

- Les moteurs de recherche actuels supportent tout type de requête
- Il existe des mots-clés/opérateurs : + @ $ # - " * ..
    - **Site:** univ-grenoble-alpes.fr/
    - **Related:**lemonde.fr
    - chat **OR** panda
    - **Info:** univ-grenoble-alpes.fr/
    - **Cache:** univ-grenoble-alpes.fr/
- https://support.google.com/websearch/answer/2466433?hl=fr&ref_topic=3081620

Satisfaction de l’utilisateur

Le système doit

- être simple à utiliser
- fournir les meilleures réponses possibles, et ces réponses doivent être « pertinentes » pour l’utilisateur

  —> Pertinence système vs pertinence utilisateur
- fournir un nombre raisonnable de réponses
- donner des réponses rapides

  Difficile de satisfaire tous ces points à la fois…

Satisfaction de l’utilisateur

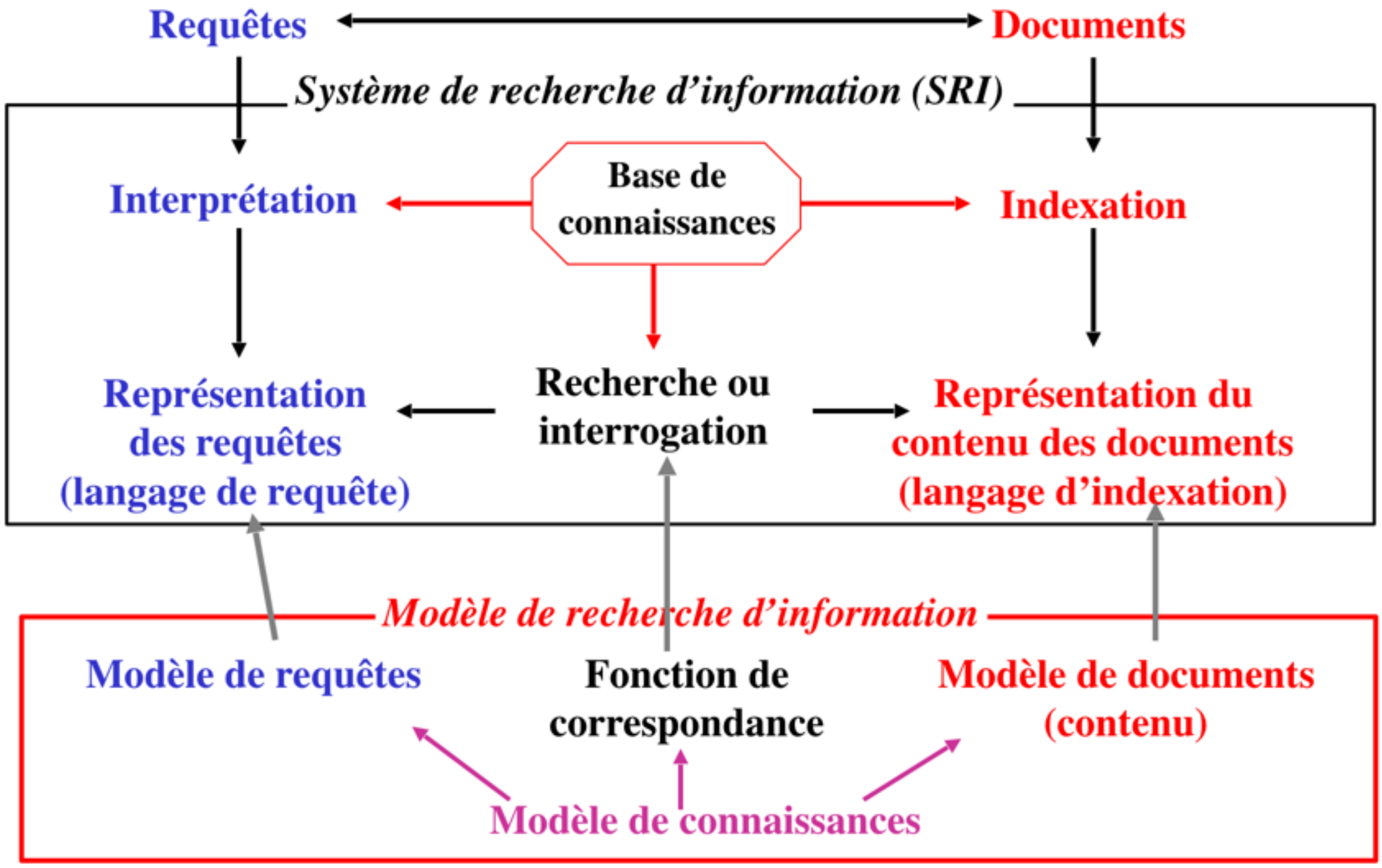
Qu’est-ce qui fait d’un document un bon résultat ?

- Pertinence du thème
- Clarté
- Adaptation aux spécificités de l’utilisateur (lisible, compréhensible…)
- Nouveauté
- …

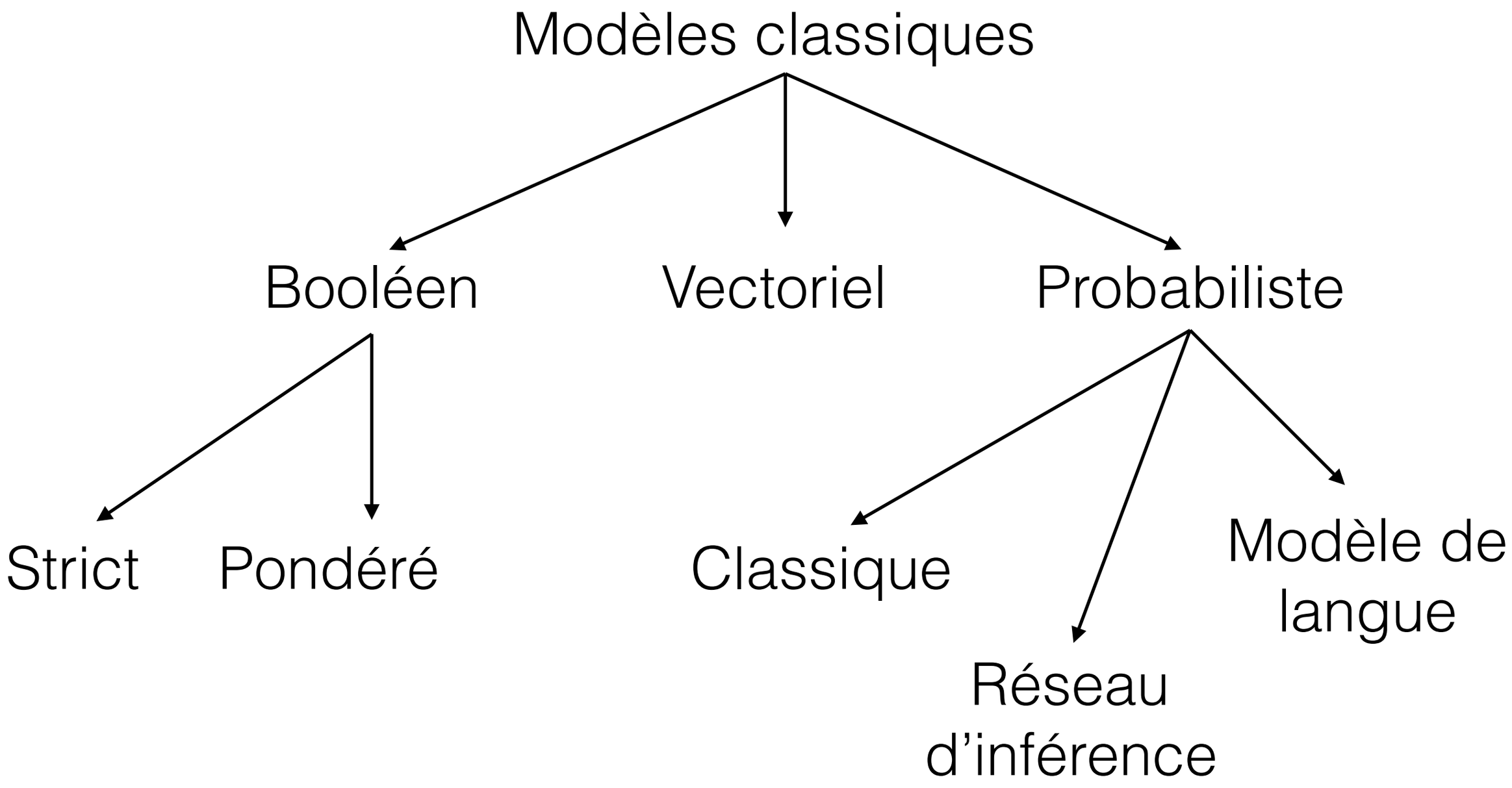
Plan

1. Introduction
2. Elements clés en recherche d'information (RI)
3. Modèles de RI
4. Systèmes de recherche d'information (SRI)
5. Evaluation des SRI

Modèles de RI



Modèles de RI



Le modèle booléen strict

- Modèle de connaissances : $T = \{t_i\}$, $i \in [1, .. N]$
    - Termes t_i qui indexent les documents
- Le modèle de documents (contenu) est une expression booléenne dans la logique des propositions avec les t_i considérés comme des propositions
    - Documents : termes "positifs" pour ceux dont parle le document, sinon "négatifs"
        - Ex. : document D1 est représenté par une formule $\mathcal{D}1 = t_1 \wedge t_3 \wedge t_{250} \wedge t_{254} (\wedge_{j \neq 1,3,250,254} \neg t_j)$
    - Une requête Q est représentée par une formule logique Q quelconque :
        - Ex . : $Q = (t_1 \wedge t_3) \vee (t_{25} \wedge t_{145} \wedge \neg t_{134})$

Le modèle booléen strict

La fonction de correspondance est basée sur l'implication logique en logique des propositions :

- Un document D répond à une requête Q si et seulement si $D \supset Q$
    - Utilisation de déduction par axiomes : $(a \wedge b) \supset a$, $(a \wedge b) \supset b$, $a \supset (a \vee b)$, $b \supset (a \vee b)$, …

- Exemple : $\mathcal{D} = t_1 \wedge t_3 \wedge \neg t_2 \wedge \neg t_4$ et $Q = t_1 \vee t_4$
    - Déduction :

      1. $t_1 \wedge t_3 \wedge \neg t_2 \wedge \neg t_4 \supset t_1$ (équivalent à $\mathcal{D} \supset t_1$)

      2. $t_1 \supset t_1 \vee t_4$ (équivalent à $t_1 \supset Q$)

      Q est donc dérivable à partir de D, donc $D \supset Q$, donc le document répond à la requête.

Le modèle booléen strict

Remarques

- Correspondance stricte : Oui/Non

$Q = t_1 \wedge t_3 \wedge t_4$ $\quad Q \not\subset D_1$

$D_1 = t_1 \wedge t_4 \wedge \neg t_2 \wedge \neg t_3$ $\quad D_1$ pas pertinent, mais contenu proche

- Pas de distinction entre les documents pertinents

$Q = t_1 \wedge t_4$

$D_2 = t_1 \wedge t_4 \wedge \neg t_2 \wedge \neg t_3 \wedge \neg t_5 \wedge \neg t_6 \wedge \neg t_7$

$D_3 = t_1 \wedge t_3 \wedge t_4 \wedge t_5 \wedge t_6 \wedge t_7 \wedge \neg t_2$

$D_2 \supset Q$ et $D_3 \supset Q$

D_2 plus pertinent que D_3

Fréquence des termes ?

Le modèle booléen strict

Remarques (suite)

- Utilisable pour des experts, mais difficile pour des utilisateurs "courants"
    - Une requête $t_1 \wedge t_4 \wedge t_6$ peut donner des centaines de réponses, et la requête $t_1 \wedge t_4 \wedge t_6 \wedge t_{10}$ peut ne donner aucune réponse …
        - Il faut donc savoir jongler avec les $\wedge$ et les $\vee$ ….
- Expression de requêtes complexes
    - $Q = ((t_1 \wedge t_4) \vee t_6) \wedge (t_8 \vee (\neg t_{10} \wedge t_{40}))$ … ???
    - Sens du $\vee$ logique (inclusif) différent du "ou" courant (exclusif)

⇒ Ceci amène à définir des modèles fournissant des résultats sous forme de listes (avec classement)

Le modèle booléen pondéré

- Extension du modèle booléen en intégrant des pondérations (dénotant la représentativité d'un terme pour un document).
- Modèle de connaissances : $T = \{t_i\}$, $i \in [1, .. N]$
  - Termes t_i qui indexent les documents
- Un document D est représenté par :
  - Une formule logique $\mathcal{D}$ (idem modèle booléen)
  - Une fonction $W_{\mathcal{D}} : T \rightarrow [0,1]$, qui pour chaque terme de T donne le poids de ce terme dans $\mathcal{D}$. Le poids vaut 0 pour un terme non présent dans le document.
- Requête : idem booléen strict.

Le modèle booléen pondéré

- Fonction de correspondance non binaire (on se passe des implications logiques) basée sur une similarité notée Sim
  - Version 1
    - Utilisation de logique floue (avec a et b des termes)
      - $Sim_1(\mathcal{D}, (a \wedge b)) = \min [W_{\mathcal{D}}(a), W_{\mathcal{D}}(b)]$
      - $Sim_1(\mathcal{D}, (a \vee b)) = \max [W_{\mathcal{D}}(a), W_{\mathcal{D}}(b)]$
      - $Sim_1(\mathcal{D}, (\neg a)) = 1 - W_{\mathcal{D}}(a)$
      - $Sim_1(\mathcal{D}, (x \wedge y)) = \min [Sim_I(\mathcal{D}, x) , Sim_I(\mathcal{D}, y)]$ (x et y sont des sous-requêtes)
- Limitation : on ne tient pas compte dans la réponse de tous les termes de la requête : ex; min(0.5, min(0.3,0.5))=min(1, min(0.3,1))

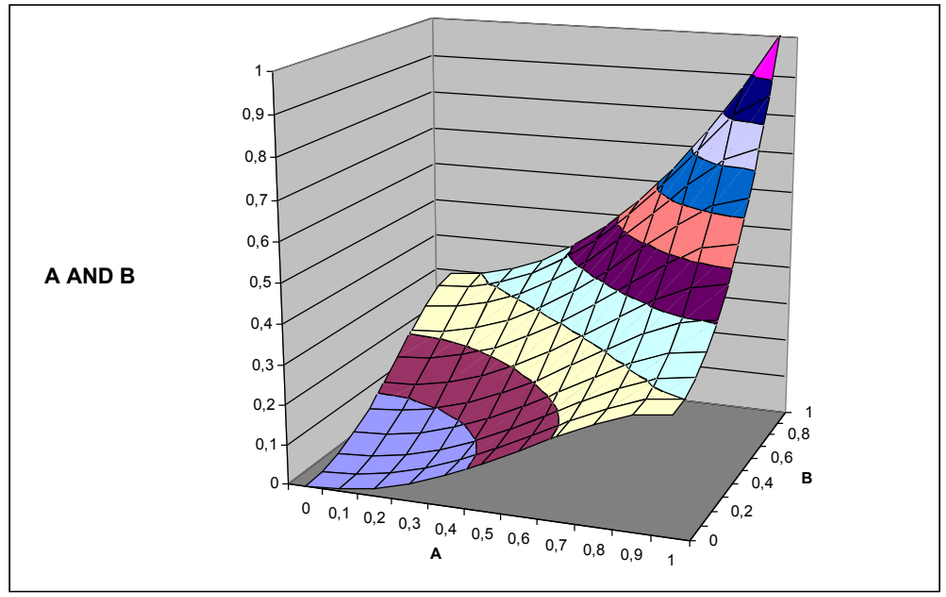
Le modèle booléen pondéré

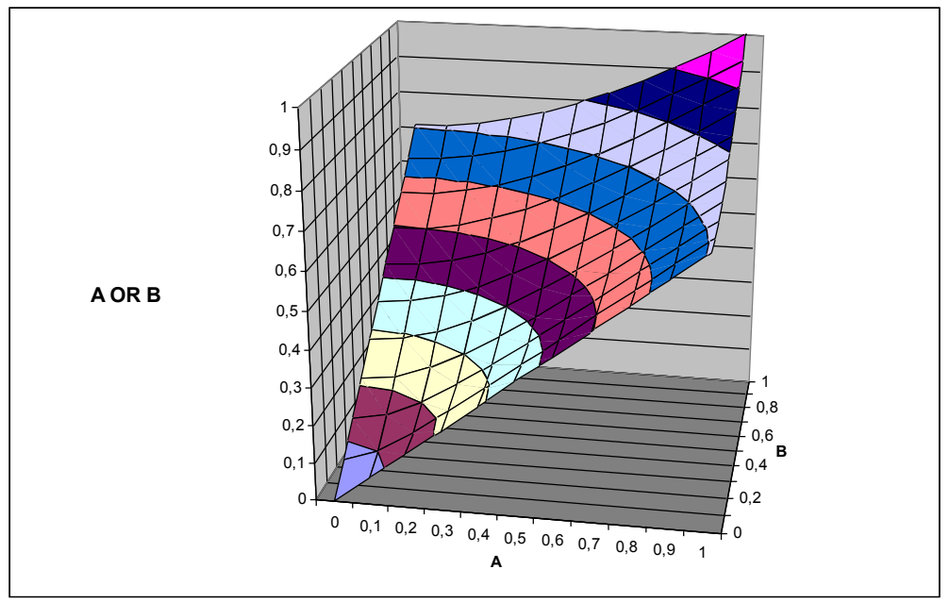
- Version 2
  - Définition d'une mesure de similarité qui tient davantage compte de chacun des termes de la requête

$$Sim_2(D, a \wedge b) = 1 - \sqrt{\frac{(1 - W_D(a))^2 + (1 - W_D(b))^2}{2}}$$

$$Sim_2(D, a \vee b) = \sqrt{\frac{W_D(a)^2 + W_D(b)^2}{2}}$$

- Limitation : pas de formule pour la négation…





Le modèle booléen pondéré

Exemple avec des valeurs binaires

	a	b	Similarité 1		Similarité 2	
Documents			a ∨ b	a ∧ b	a ∨ b	a ∧ b
D1	1	1	1	1	1	1
D2	1	0	1	0	$1/\sqrt{2}$	$1-1/\sqrt{2}$
D3	0	1	1	0	$1/\sqrt{2}$	$1-1/\sqrt{2}$
D4	0	0	0	0	0	0

Le modèle booléen pondéré

Exemple avec des valeurs non binaires

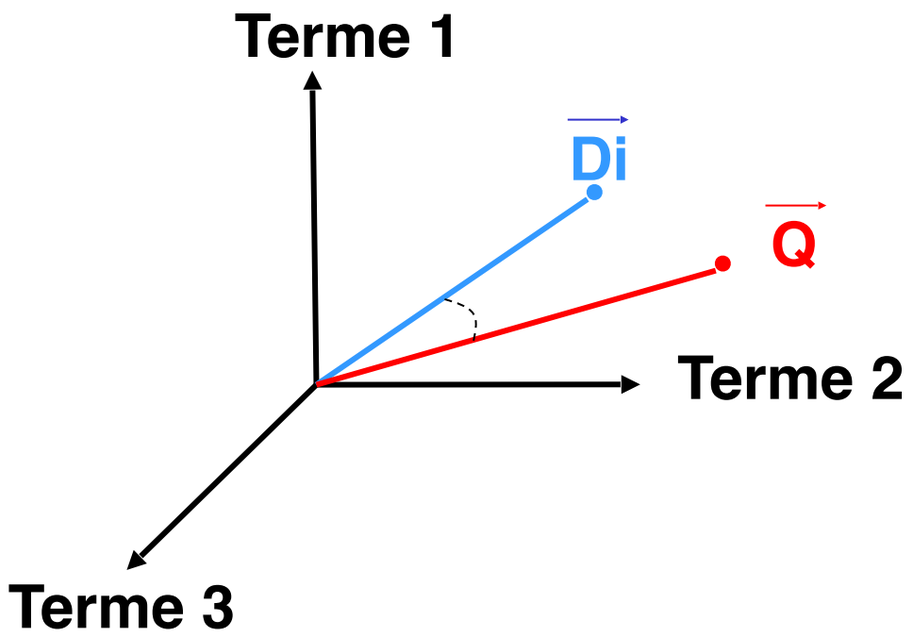
	a	b	Similarité 1		Similarité 2	
Documents			$a \vee b$	$a \wedge b$	$a \vee b$	$a \wedge b$
D1	1	1	1	1	1	1
D2	0.8	1	1	0.8	0.91	0.86
D3	0	0.5	0.5	0	0.35	0.21
D4	0.8	0	0.8	0	0.57	0.28

Le modèle vectoriel

- Modèle de connaissances : $T = \{t_i\}$, $i \in [1, .. N]$
- Tous les documents sont décrits suivant ce vocabulaire
- Un document D_i est représenté par un vecteur D_i décrit dans l'espace vectoriel R^N défini par T :
  - $D_i = (w_{i,1}, w_{i,2}, \ldots, w_{i,j}, \ldots, w_{i,N})$, avec $w_{k,l}$ le poids d'un terme pour un document
- Une requête Q est représentée par un vecteur Q décrit dans l'espace vectoriel R^N défini par T :
  - $Q = (w_{Q,1}, w_{Q,2}, \ldots, w_{Q,j}, \ldots, w_{Q,N})$

Le modèle vectoriel

Plus les vecteurs représentant les documents/ requêtes sont « proches », plus les documents/ requêtes sont similaires :



Le modèle vectoriel

Comment trouver les poids des termes pour les documents :

- Soit le document : « Un violon est issu de bois précieux comme l'érable, palissandre, l'ébène… »

- Pour le représenter, la première idée est de compter les mots les plus fréquents exceptés les termes non significatifs comme *de*, *avec*, *comme*…

  - « Un violon est composé de bois précieux comme l'érable, le palissandre, l'ébène… »

Le modèle vectoriel

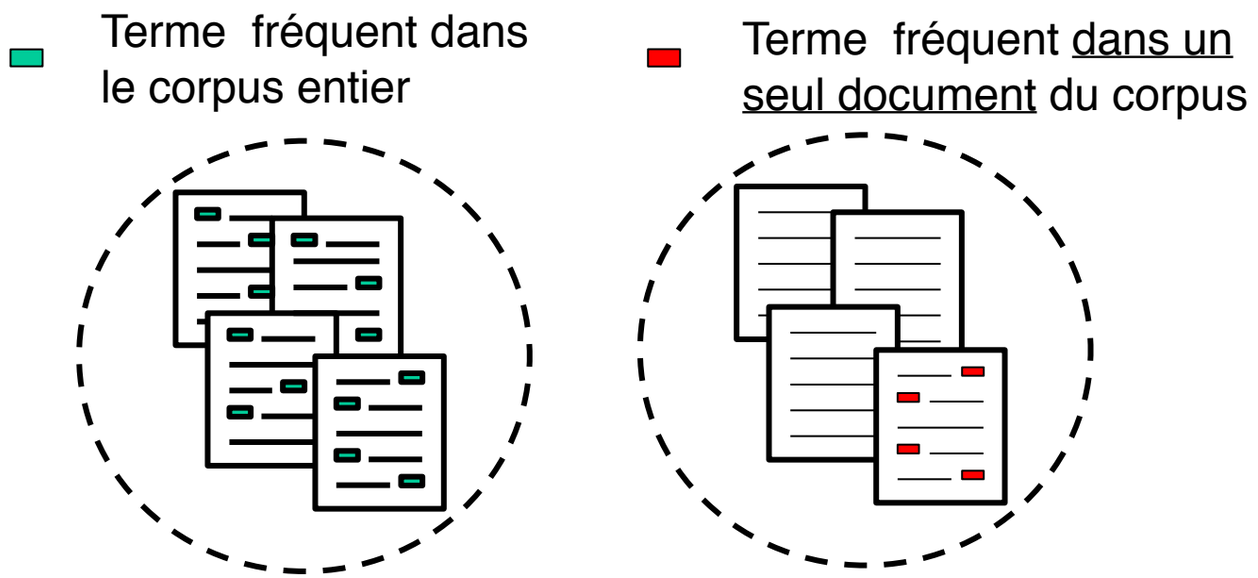
Comment trouver les poids des termes pour les documents :

- On définit la fréquence d'un terme (term frequency)
  - $tf_{i,j}$: la fréquence du terme t_j dans le document D_i est égale au nombre d'occurrences de t_j dans D_i
  - Exemple : si *violon* apparaît 5 fois dans le document D_3, avec violon=t_{23}, alors $tf_{3,23} = 5$

Le modèle vectoriel

Comment trouver les poids des termes pour les documents :

- On tient compte du corpus (base de documents) entier, un terme qui apparaît beaucoup ne discrimine pas nécessairement les documents :



Le modèle vectoriel

Comment trouver les poids des termes pour les documents :

- On définit la **fréquence documentaire d'un terme** df_j
  - df_j : la fréquence dans le corpus du terme t_j est le nombre de documents du corpus où t_j apparaît
- On utilise l'**inverse de la fréquence documentaire**, idf_j :
  - Définition simple : $idf_j = 1 / df_j$
  - Définition la plus utilisée : $idf_j = \log(N_D / df_j)$, avec N_D le nombre de documents du corpus.

Le modèle vectoriel

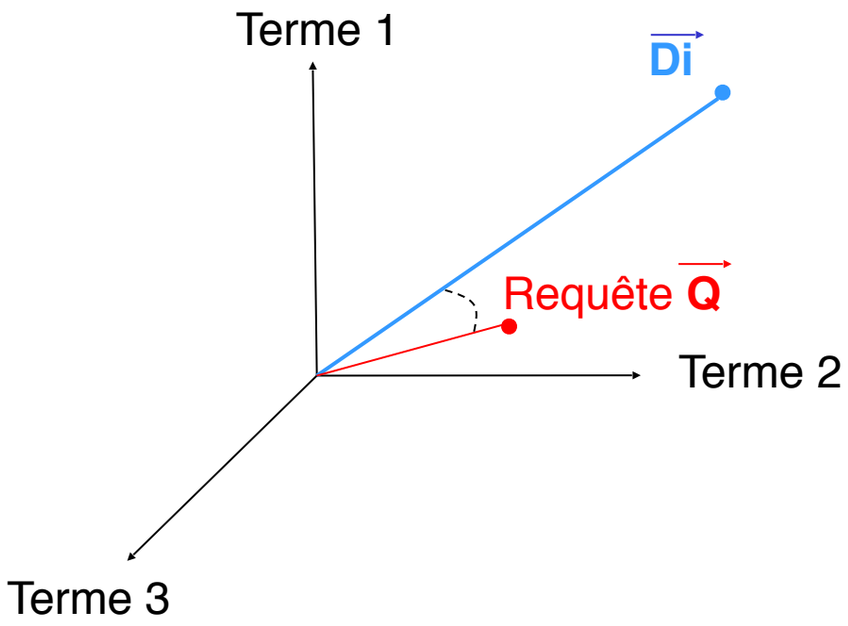
Comment trouver les poids des termes pour les documents :

- Combinaison du tf et de l'idf pour un vecteur document :
  - Le poids d'un terme dénote la capacité du terme à discriminer les documents
  - Exemple le plus courant : $w_{i,j} = tf_{i,j} \cdot idf_j$
- Utilisation du tf ou du tf.idf pour une requête

Le modèle vectoriel

Comment calculer la similarité entre un document et une requêtes :

- Fonction de correspondance : fonction de l'angle entre le vecteur requête Q et le vecteur document D_i



💡Plus l'angle est petit et plus le document correspond à la requête

Le modèle vectoriel

Comment calculer la similarité entre un document et une requêtes :

- Fonctions de correspondance :
  - Cosinus 👍 $Sim(\vec{D}_i, \vec{Q}) = \frac{\sum_{k=1}^{N} (w_{i,k} . w_{q,k})}{\sqrt{\sum_{k=1}^{N} (w_{i,k}{}^2) . \sum_{k=1}^{N} (w_{q,k}{}^2)}} = \frac{\sum_{k=1}^{N} (w_{i,k} . w_{q,k})}{\|\vec{D}i\| . \|\vec{Q}\|}$
  - Jaccard $Sim_{Jaccard}(\vec{D}_i, \vec{Q}) = \frac{\sum_{k=1}^{N} (w_{i,k} . w_{q,k})}{\sum_{k=1}^{N} (w_{i,k}{}^2) + \sum_{k=1}^{N} (w_{q,k}{}^2) - \sum_{k=1}^{N} (w_{i,k} . w_{q,k})}$
  - Dice $Sim_{Dice}(\vec{D}_i, \vec{Q}) = \frac{2 * \sum_{k=1}^{N} (w_{i,k} . w_{q,k})}{\sum_{k=1}^{N} (w_{i,k}{}^2) + \sum_{k=1}^{N} (w_{q,k}{}^2)}$