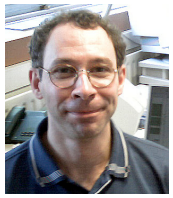
Multimedia information retrieval Modeling

Contrainte de correspondance Document-Document pour la RI

Application à la divergence de Kullback-Leibler



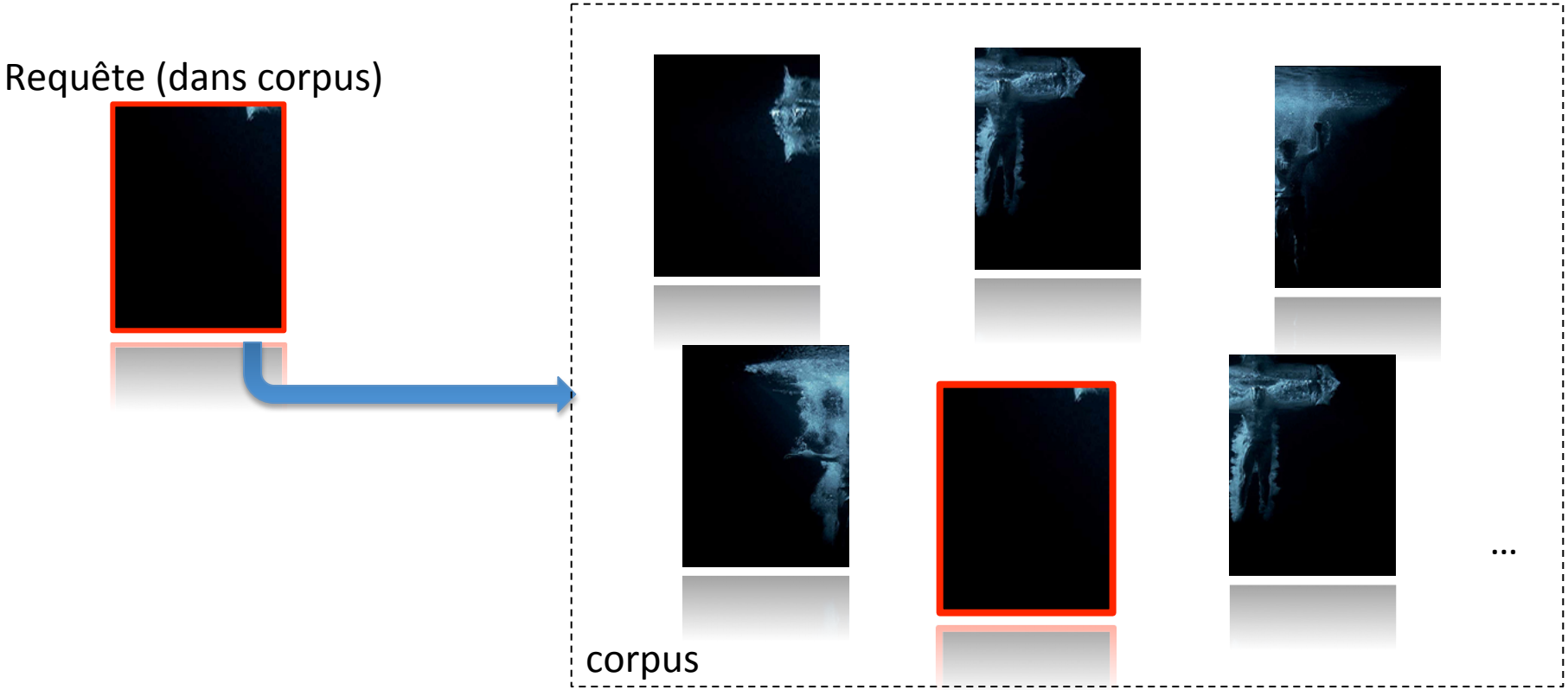
P. Mulhem J.-P. Chevallet

RhôneAlpes Région

Prélude…

De tout temps, les hommes ont voulu retrouver de l'information

- Recherche d'image fixe par l'exemple



Plan

1. Point de départ : comportement inattendu
2. Existant : Contraintes axiomatiques pour la RI
3. Comportement attendu : DDMC
4. Modèles de langues utilisant la divergence de Kullback-Leibler

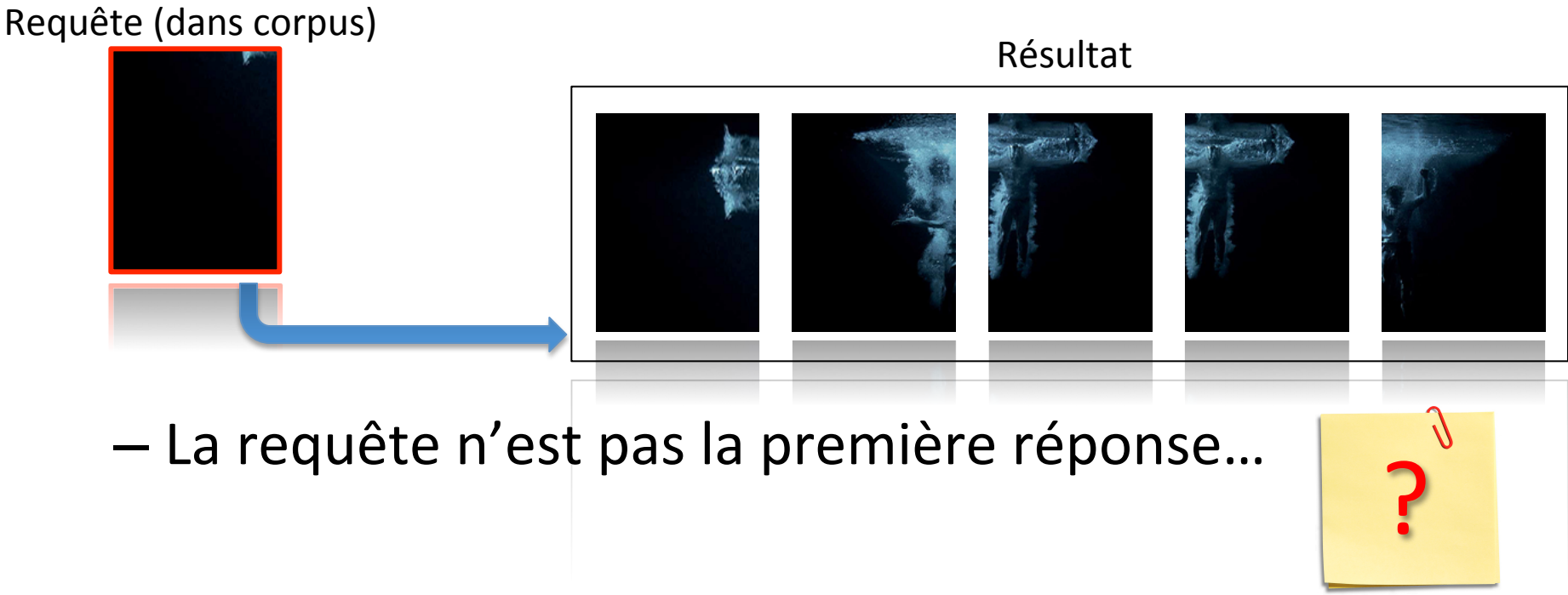
   Modification pour validation de DDMC : nSKL
5. Etude de nSKL

   Théorique

   Expérimentale

1. Point de départ - Comportement «inattendu»

- Recherche d'image fixe
  - sacs de mots, modèle de langue par divergence de Kullback-Leibler, lissage Jelinek-Mercer



  - La requête n'est pas la première réponse…

2. Existant : contraintes axiomatiques pour la RI

- En 2004, <u>révolution</u> par [Fang, Tao & Zhai 2004]
  - Analyse des formules : comportements attendus
  - Exemple :
    - TFC1 (Term Frequency Constraint One) :

> *Soit : Q = {w} une requête avec un terme w.*
> *Posons : $|D_1|=|D_2|$.*
> *si $c(w, D_1) > c(w, D_2)$, alors $f(D_1,Q)>f(D_2,Q)$.*

Forme analytique [Clinchant Gaussier 2010] : $\frac{\partial f}{\partial c} \geq 0$

2. Existant : contraintes axiomatiques pour la RI

- Fang, Tao et Zhai ont défini 6 contraintes « classiques » en 2004 :
  - Tf : TFC1, TFC2
  - Idf : Term Discrimination Constraint (TDC)
  - |D| : Length Normalization Constraint (LNC1, LNC2)
  - Tf-Length Constraint (TF-LNC)
- Evolutions ultérieures [Clinchant & Gaussier, Lv & Zhai, Rahimi, Shakery & Zhai, Cummins & O'Riordan]

On s'inspire de ses travaux, selon un point de vue global par rapport au corpus.

3. Comportement attendu : DDMC

- Proposition :

Document-Document Matching Constraint

La valeur de correspondance entre
- un document *D* du corpus utilisé comme requête, et
- ce même document dans le corpus *C*

est supérieure ou égale à celle entre *D* utilisé comme requête et chacun des autres documents du corpus.

$$\forall D \in C, \nexists D' \in C, D' \neq D, f(D,D') > f(D,D)$$

  - On espère que cette contrainte peut induire un comportement intéressant pour la RI

NOTE : BM25, et modèle de langue classique ne valident pas DDMC.

Cadre général

1. Etudier un modèle « simple » mais reconnu pour sa qualité, pour voir se qui se passe…

   Modèle de langues avec négative de Kullbak-Leibler, Lissage de Jelinek-Mercer

2. Tenter de trouver une modification simple qui valide DDMC, pour l'étudier…

   Lissage de requête

3. Etudier ce modèle modifié
   - DDMC, contraintes « classiques », fichier inverse

4. Modèles de langues utilisant la divergence de Kullback-Leibler

- Modèle de référence (nKL)
  - Négative de la divergence de Kullback-Leibler

$$nKL(Q||D) = -\sum_{t \in Q} P_{ML}(t|Q).\log(\frac{P_{ML}(t|Q)}{P_{\lambda}(t|D)})$$

  - Modèle de requête Q : maximisation de vraisemblance

$$P_{ML}(t|Q) = \frac{c(t,Q)}{|Q|}$$

  - Modèle d'un document D : Lissage de Jelinek-Mercer (adapté aux requêtes longues [Zhai & Lafferty 2001])

$$P_{\lambda}(t|D) = (1-\lambda).P_{ML}(t|D) + \lambda.P_{ML}(t|C)$$

4. Modèles de langues utilisant la divergence de Kullback-Leibler

- Explication de DDMC non validée par nKL :

modèles de documents

$$P_\lambda(t|D) = (1-\lambda).P_{ML}(t|Q) + \lambda.P_{ML}(t|C)$$

$\neq$

modèles de requêtes

$$P_{ML}(t|Q) = \frac{c(t,Q)}{|Q|}$$

- Proposition:
  - utilisation de nKL avec les requêtes lissées, nSKL

    (idem [Louis & Nenkova 2009])

$$nSKL(Q||D) = -\sum_{t \in T} P_\lambda(t|Q).\log\left(\frac{P_\lambda(t|Q)}{P_\lambda(t|D)}\right)$$

5. Etude de nSKL - Théorique

✔ nSKL valide DDMC

$nSKL(D||D) = 0$: valeur maximale .

✔ Validation des contraintes « classiques »

[Fang, Tao & Zhai 2004, Clinchant & Gaussier 2011]

Contraintes	**TFC1**	**TFC2**	**speTDC**	**LNC1**	**LNC2**	**TF-LNC**
nKL	OK	OK	OK	OK	OK	Cond : $c(w,d) < \|D\| . p_{ML}(t\|C)$
nSKL	OK*	OK*	OK*	OK*	OK*	numérique

* Via approximation (cf. papier)

5. Etude de nSKL - Théorique

✔ Compatible avec fichiers inverses [Mulhem & Chevallet 2014]

fichier inverse

$$nSKL(Q||D) \propto_Q \underbrace{\sum_{t \in Q \cap D} (1-\lambda).P_{ML}(t|Q).\log(\frac{P_\lambda(t|D)}{\lambda.P_{ML}(t|C)})}$$

$$\underbrace{+ \sum_{t \in D} P_{ML}(t|C).\log(\frac{P_\lambda(t|D)}{\lambda.P_{ML}(t|C)})}_{K_D}$$

5. Etude de nSKL - Expérimentale

- Images : requêtes par l'exemple
  - ZuBud [Shao et al 2004]
    - 1005 images de 201 bâtiments / 115 requêtes
    - modèle de langue sur mots-clés visuels (2k), SIFT [Lowe 2004, VandeSande et al 2010]

divergence (paramètre)	MAP	MRR	P@10	Taux de Reco.
nKL ($\lambda = 0,2$)	0.6988	0.9072	0.3826	0.8782
nSKLD ($\lambda = 0,1$)	0.6949 (-0,6 %)	0.9029 (-0,5 %)	0.3783 (-1,1 %)	0.8696 (-1,0 %)

- Pas de différence de MAP statistiquement significative (test de Student pairé bilatéral p<=5%)

5. Etude de nSKL - Expérimentale

- Textes
  - TREC-6 ad-hoc [Voorhes & Harman, 2000]
    - 550 000 documents, 50 requêtes
    - Requêtes « longues » : *topic+description+narrative*

divergence (paramètre)	MAP	P@10	P@30
nKL ($\lambda = 0,7$)	0.2646	0.4400	0.3307
nSKLD ($\lambda = 0,4$)	0.2477 (-6,4 %)	0.4360 (-0,9 %)	0.3247 (-1,8 %)

✔ – Pas de différence de MAP statistiquement significative (test de Student pairé bilatéral p<=5%)

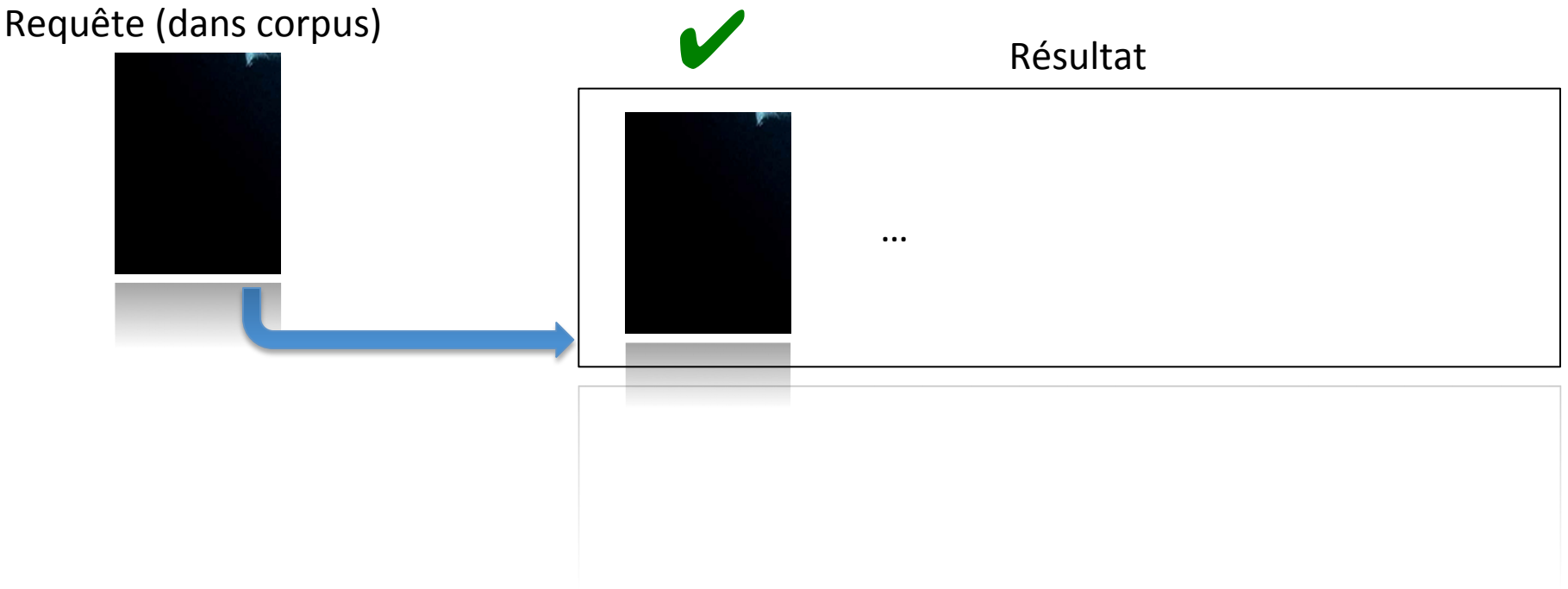
Conclusion

- Proposition d'une contrainte basée sur le corpus, pour des requêtes « documents » : DDMC
- Modification de modèle de langue à base de négative de Divergence de Kullback-Leibler : nSKL
  - Compatibilité avec fichiers inverses
  - Validation partielle des contraintes « classiques »
- Validation expérimentale préliminaire
  - Résultats très proches de nKL

Travaux futurs

- Etude de nSKL plus poussée
  - Validation des contraintes classiques
  - Expérimentations (requêtes/documents longs/ courts)
- Vers une démarche scientifique pour modifier des formules existantes pour les rendre compatibles avec DDMC

- Sur notre exemple initial :



~~5. Etude de nSKL - Théorique~~

✔ TFC1 validée par calcul de dérivée partielle sur les occurrences des termes [Clinchant et Gaussier 2010]

- En notant la variable *c* pour c(w,D)

$$nSKL(Q\|D) = (1-\lambda).\log\left(\frac{1-\lambda}{\lambda}.\frac{c(w,D)}{|D|.P_{ML}(w|C)}+1\right) + \ldots\ldots$$

$$\frac{\partial nSKL(Q\|D)}{\partial c} = \frac{(1-\lambda)^2}{\lambda.|D|.P_{ML}(w|C)+c-\lambda.c} > 0 \quad si \ \lambda \in]0,1[$$

✔ Suivant les mêmes idées, TFC2, TFC3, LNC1, LNC2, speTDC validées partiellement

✔ TF-LNC validée partiellement numériquement

~~5. Etude de nSKL - Théorique~~

- Validation des contraintes de [Fang, Tao & Zhai 2004]
  - Pour les contraintes avec un terme (TFC1, TFC2)

$$nSKL(Q\|D) = (1-\lambda).\log(\frac{1-\lambda}{\lambda}.\frac{c(w,D)}{|D|.P_{ML}(w|C)}+1)$$
$$+\lambda.P_{ML}(w,C).\log(\frac{1-\lambda}{\lambda}.\frac{c(w,D)}{|D|.P_{ML}(w|C)}+1)$$
$$+\lambda.P_{ML}(w',C).\log(\frac{1-\lambda}{\lambda}.\frac{c(w',D)}{|D|.P_{ML}(w'|C)}+1)$$
$$+\sum_{t\in D\setminus\{w,w'\}}\lambda.P_{ML}(t,C).\log(\frac{1-\lambda}{\lambda}.\frac{c(t,D)}{|D|.P_{ML}(t|C)}+1)$$

DDMC pour modèles classiques

- Modèle vectoriel : tf.idf pour D et Q, cosinus
  - DDMC validé : cosinus(D,D)=1

- BM25 : $f_{BM25}(Q,D) = \sum_{t \in D \cap Q} ln \frac{N - df_t + 0.5}{df_t + 0.5} \cdot \frac{(k_1+1).\#t,D}{k_1.((1-b)+b.\frac{|D|}{avdl}) + \#t,D} \cdot \frac{(k_3+1).\#t,Q}{k_3 + \#t,Q}$
  - DDMC non validée sur un exemple

- Modèle de langue
  - Nég. de div. de Kullback-Leibler : $nKL(Q||D) = -\sum_{t \in Q \cap D} P_{ML}(t|Q).\log(\frac{P_{ML}(t|Q)}{P_\lambda(t|D)})$
  - Lissage de Jelinek-Mercer de D : $P_\lambda(t|D) = (1-\lambda).P_{ML}(t|Q) + \lambda.P_{ML}(t|C)$
  - Maximisation de vraisemblance de Q : $P_{ML}(t|Q) = \frac{c(t,Q)}{|Q|}$
  - DDMC non validée sur un exemple (confirmation du cas des images)