

Recherche d'Information (RI)

Mathias Géry

Mathias.Gery@univ-st-etienne.fr

Laboratoire Hubert Curien, UMR CNRS 5516

Université Jean Monnet Saint-Étienne

Laboratoire Hubert Curien

- Effectifs : 240, dont 130 enseignants-chercheurs / chercheurs.
- Département Optique-Photonique.
- Département Informatique-Télécom-Image :
 - Thématique « Formation, compréhension et analyse de l'image ».
 - Thématique « Systèmes embarqués sécurisés et architectures matérielles ».
 - Thématique « Connected Intelligence ».
 - **Thématique « Data Intelligence » (Amaury Habrard).**



Laboratoire Hubert Curien

- Thématique « Data Intelligence » (Amaury Habrard) :
 - Projet « Machine Learning » (E. Fromont).
 - **Projet « Data Mining / Information Retrieval » (C. Largeton) :**
 - 8 enseignants-chercheurs, 5 doctorants / post-docs.
 - Thèmes de recherche :
 - Fouille et Recherche d'Information (RI) Structurée (XML).
 - RI et réseaux sociaux.
 - Fouille de données.
 - Fouille de réseaux sociaux, de graphes.
 - RI Multimédia (images).
 - Etc.
 - Prototypes, expérimentations, participation à des compétitions internationales.
 - Relations entreprises (projets, thèses CIFRE).

Recherche d'Information (RI)

- Plan du cours « RI » :

Michel Beigbeder

- 5. RI et structure intra-document (ex. XML)
- 6. RI et structure inter-documents (ex. les liens hypertextes)
- 7. Projet : de la théorie à la pratique, mise en œuvre d'un SRI sur des données structurées du Web, participation à une mini-compétition de RIS.

Recherche d'Information (RI)

- Plan du cours « RI Structurée » (RIS) :
 - 5. RI et structure intra-document (ex. XML).
 - Introduction : RI classique.
 - Web (HTML, XML) : Document Structuré ou Hypertexte ?
 - RI Structurée :
 - Langages de requêtes structurés.
 - Indexation en RIS.
 - Passage Retrieval, modèle à proximité.
 - Pondération en RIS : normalisation, types d'éléments, propagation.
 - Évaluation en RI : la compétition INEX.
 - 6. RI et structure inter-documents (ex. les liens hypertextes).
 - 7. Projet : participation à une mini-compétition de RIS.

Recherche d'Information (RI)

- Cours (diapos sur l'ENT).
- TDs : algorithmes, articles de recherche.
- Projet : Moteur de recherche de documents structurés.
 - À réaliser entre les séances,
 - Discussions en séance,
 - Étapes intermédiaires,
 - Rendu : 25 janvier.
 - ➔ note « pratique » du cours de Recherche d'Information.

Recherche d'Information (RI)

- Planning : attention aux modifications de dernière minute !
- ➔ vérifiez ADE !
 - 1) Vendredi 21/10, 14h-16h, salle B10 Carnot
 - 2) Vendredi 18/11, 14h-16h, salle B10 Carnot
 - 3) Mardi 06/12, 13h30-15h30, salle A13 Métare
 - 4) Mardi 13/12, 13h30-15h30, salle A13 Métare
 - 5) Jeudi 5/01/2017, 14h-16h, salle B10 Carnot
 - 6) Mardi 17/01, 13h30-15h30, salle A13 Métare

Cours en ligne : plateforme Claroline

Cours en ligne : RI2016

<https://ead-sciences.univ-st-etienne.fr/claroline/course/index.php?cid=RI2016>

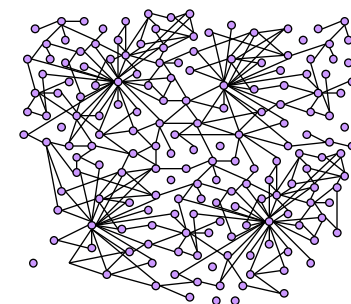
Baaziz	Hamza	Escalle	Dimitri
Bah	Alhassane	Forestier	Fabien
Bendari	Yassine	Gourseaud	Pierre
Berlande	Jacomo	Khalifa	Nizar
Bezrhoud	Ibrahim	Mersel	Dali
Boualem	Oussama	Ramirez	Diana
Bouassida	Haithem tahar	Rouatbi	Nour el houda
Cretel	Lea	Tichit	Ludovic
Da costa vaz	Julien	Wang	Jing
Diallo	Marlyatou	Wang	Zun zun

Recherche d'Information (RI)

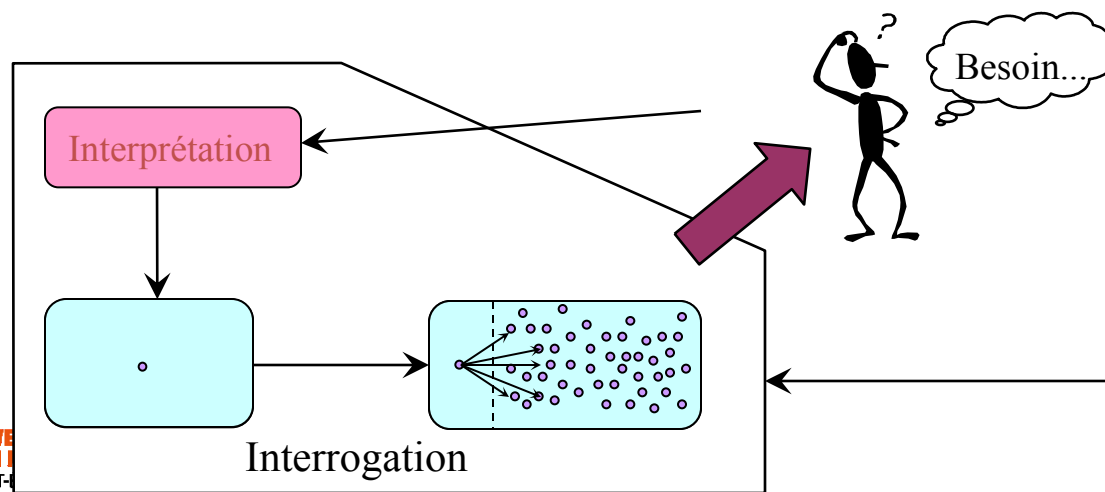
- Plan du cours « RI Structurée » (RIS) :
 - 5. RI et structure intra-document (ex. XML).
 - Introduction : RI classique.
 - Web (HTML, XML) : Document Structuré ou Hypertexte ?
 - RI Structurée :
 - Langages de requêtes structurés.
 - Indexation en RIS.
 - Passage Retrieval, modèle à proximité.
 - Pondération en RIS : normalisation, types d'éléments, propagation.
 - Évaluation en RI : la compétition INEX.
 - 6. RI et structure inter-documents (ex. les liens hypertextes).
 - 7. Projet : participation à une mini-compétition de RIS.

Recherche d'Information « Classique » (rappel)

- « Recherche d'Information », cf. cours M. Beigbeder.
 - Modèles classiques de RI (booléen, vectoriel) :
 - L'unité d'indexation (document) est la page HTML.
 - Correspondance : document \leftrightarrow requête.
- documents atomiques et indépendants.

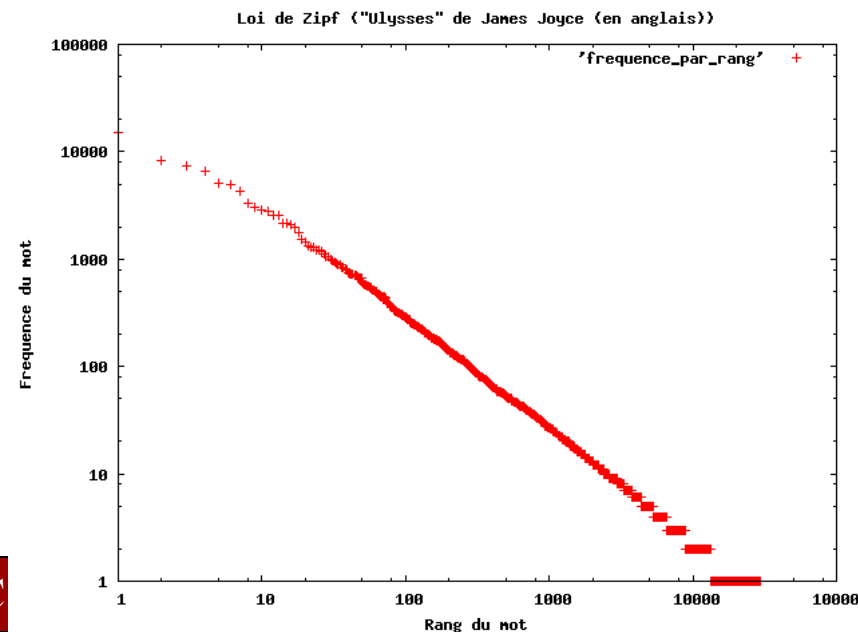


Web



Modèles classiques de RI (rappel)

- 50 ans de recherche en RI !
- Exemple : modèles « bag of words » (« sac de mots »).
 - Documents atomiques et indépendants.
 - Modéliser la distribution des termes :
 - Loi de Zipf (fréquence du terme $f(n) = \frac{K}{n}$, $n = \text{rang du terme}$).
 - Une formule « magique » : pondération $tf.idf$
 - Représentativité.
 - Discriminance.



Pondération en RI classique (rappel)

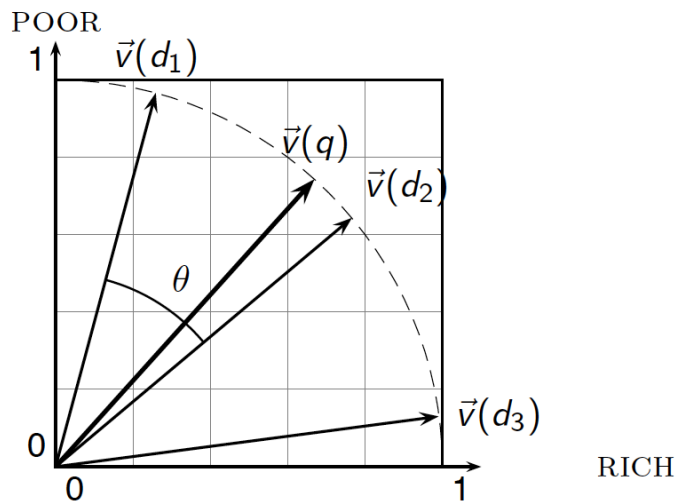
- Pondération *tf.idf* classique ?
- Exemple SMART *ltn* :
 - Poids $w_{t,d}$ du terme t dans le document d :

$$w_{t,d} = TF_{t,d} * IDF_t = (1 + \log(tf_{t,d})) * \log\left(\frac{N}{df_t}\right)$$

- Avec :
 - $TF_{t,d}$ = **représentativité** de t dans d
= par ex. : nombre d'occurrences de t dans d
 - IDF_t = **discriminance** de t dans le corpus
= par ex. : nombre de documents contenant t
 - N = nombre de documents dans le corpus.

Modèles classiques de RI (rappel)

- Exemple : modèle vectoriel [Salton83]
 - Document représenté par un vecteur dans l'espace vectoriel des termes.
 - Index = matrice termes x documents.



$$\vec{d}_1 = (w_{11}, w_{12}, \dots, w_{1j}, \dots, w_{1n})$$

$$\vec{d}_2 = (w_{21}, w_{22}, \dots, w_{2j}, \dots, w_{2n})$$

$$\vec{d}_3 = (w_{31}, w_{32}, \dots, w_{3j}, \dots, w_{3n})$$

$$\vec{q} = (w_{q1}, w_{q2}, \dots, w_{qj}, \dots, w_{qn})$$

d/t	RICH	POOR
d ₁	0.25	0.97
d ₂	0.75	0.63
d ₃	0.98	0.14
q	0.65	0.75

Matrice
termes-documents

Source : P. Nayak, P. Raghavan, Standford

Et la structure dans tout ça ? Quelle structure ?

Recherche d'Information (RI)

- Plan du cours « RI Structurée » (RIS) :
 - 5. RI et structure intra-document (ex. XML).
 - Introduction : RI classique.
 - Web (HTML, XML) : Document Structuré ou Hypertexte ?
 - RI Structurée :
 - Langages de requêtes structurés.
 - Indexation en RIS.
 - Passage Retrieval, modèle à proximité.
 - Pondération en RIS : normalisation, types d'éléments, propagation.
 - Évaluation en RI : la compétition INEX.
 - 6. RI et structure inter-documents (ex. les liens hypertextes).
 - 7. Projet : participation à une mini-compétition de RIS.

Structure du Web et RIS

- **Objectif : RI Structurée (RIS) pour le Web.**
- RI « classique » :
 - Ensemble de documents atomiques et indépendants.
 - ➔ Modèles classiques à repenser pour la RI Structurée.
- Quels documents sur le Web ?
 - On oublie le Web 2.0. Ici : Web documentaire, « Web 1.0 » ou « Web 1.5 ».
 - Client / serveur : 1 requête HTTP ➔ 1 document HTML/XML.
 - Documents Structurés « classiques » (arborescence).
 - Dynamique mais pas trop (CMS, blog, etc.).
- Simplification : Web = Documents Structurés + Graphe.

Structure du Web et RIS

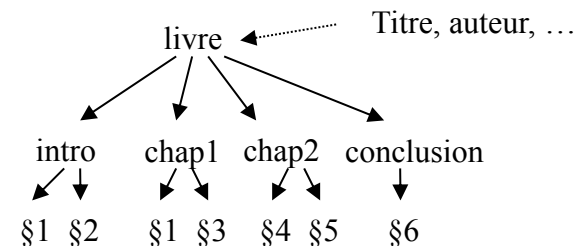
- Qu'est-ce qu'un document sur le Web ?
- Un document structuré arborescent ?
 - HTML, XML, etc.
 - `<h1>`, `<h2>`, `<p>`, `<div>`, ``, `<table>`, etc.
- Un hypertexte ?
 - Page HTML = nœud du graphe.
 - Lien hypertexte = arcs du graphe.
 - ``
 - ``
 - ``
- Une base de données (data-centric) ?

Web : Documents Structurés (DS) ?

- Document Structuré (DS) :
 - contenu + structure + sens de lecture
- Le Web contient des documents structurés :
 - Structure implicite ou formats HTML, XML, MPEG, etc.
 - Un livre est composé de : titre, introduction, chapitre 1, chapitre 2, conclusion, etc.
 - Méta-données (auteur, mots clés, propriétaire, date, etc.)

```
<html>
<head>
<title> Mon titre </title>
<meta name="author"
      content="Mathias">
<meta name="keywords"
      content="AAW Web">
</head>
<body>
```

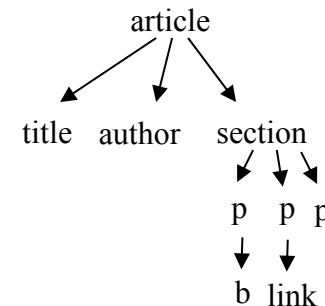
```
<h1>Introduction</h1>
  bla bla
<h1>Chapitre 1</h1>
  bla bla
<h1>Chapitre 2</h1>
  bla bla
<h1>Conclusion</h1>
.....
</body></html>
```



XML : eXtensible Markup Language

- XML :
 - Version simplifiée de SGML.
 - Standard W3C (XML 1.0, 1998).
- Concepts XML :
 - Balises (tags).
 - Balises ouvrantes / fermantes.
 - Eléments :
 - Feuilles ou nœuds.
 - Composants logiques d'un document XML.
 - Attributs.
 - Pas de chevauchement.

```
<article>
  <title>RI Structurée</title>
  <author id="443">Mathias G ry</author>
  <section>
    <p>La RI <b>structur e</b>
      ... bla bla...</p>
    <p>Pour aller <link href="plusloin.xml">
      plus loin</link></p>
    <p>R f rences...</p>
  </section>
</article>
```



XML : eXtensible Markup Language

- Comparable à HTML, mais :

- Extensible (balises / tags).
- DTD (Document Type Definition).
- Les balises doivent être fermées.

- Différents structures :

- Logique,
- physique,
- de mise en forme,
- de navigation, etc.

- XML utilisé par LA compétition de RIS :

- INEX (INitiative for the Evaluation of XML Retrieval), 2002.

```
<!DOCTYPE article [  
  <!ELEMENT article (title, author, section+)>  
  <!ELEMENT title (#PCDATA)>  
  <!ELEMENT author (#PCDATA)>  
  <!ATTLIST author id CDATA #REQUIRED>  
  <!ELEMENT section (p+)>  
  <!ELEMENT p (b, link, #PCDATA)*>  
  <!ELEMENT b (#PCDATA)>  
  <!ELEMENT link (#PCDATA)>  
  <!ATTLIST link href (#CDATA) #REQUIRED>  
>
```

Exemple de document structuré en XML

```
<article xmlns:xlink="http://www.w3.org/1999/xlink">
  <header>
    <title>Handel House Museum</title>
    <id>1707709</id>
  </header>
  <bdy>
    <image width="150px" src="London_Handel_House.jpg" type="thumb">
      <caption>Handel House. Note the <link xlink:href="310649.xml">blue plaque</link>
    </caption>
    </image>
    <p>
      The <b>Handel House Museum</b> at 25 <link xlink:href="2599649.xml">Brook
      Street</link>, in the exclusive central <link xlink:href="17867.xml">London</link>
      district of <link xlink:href="94167.xml">Mayfair</link> was the home of the <link
      xlink:href="11867.xml">German</link> born <link xlink:href="4500.xml">baroque</link>
      composer <link xlink:href="12775.xml">George Frideric Handel</link>
      from 1723 until his death at the house in 1759. He composed works such as
      <it><link xlink:href="149131.xml">The Messiah</link></it>, <it><link
      xlink:href="811987.xml">Zadok the Priest</link></it> and the <it><link
      xlink:href="1246814.xml">Fireworks Music</link></it> there.
    </p>
    <sec>
      <st>The museum</st>
    </p>
    The house has been restored to look as it did during Handel's occupancy. A typical
    early 18th century London terrace house, it comprises a basement, three main storeys and an
    attic, and Handel was the first occupant. The attic was later converted into a fourth full
    floor. The ground floor is now a music and gift shop and the upper floors are leased to a
    charity called the Handel House Trust, and have been open to the public since 8 November
    2001. The interiors have been restored to the somewhat spartan style of Georgian era,
    using mostly architectural elements from elsewhere, as other than the staircase, few of
    the original interior features survived. The Handel House Collection Trust has assembled a
    collection of Handel memorabilia, including the Byrne Collection of several hundred items,
    which was acquired in 1998.
  </p>
</bdy>
</article>
```

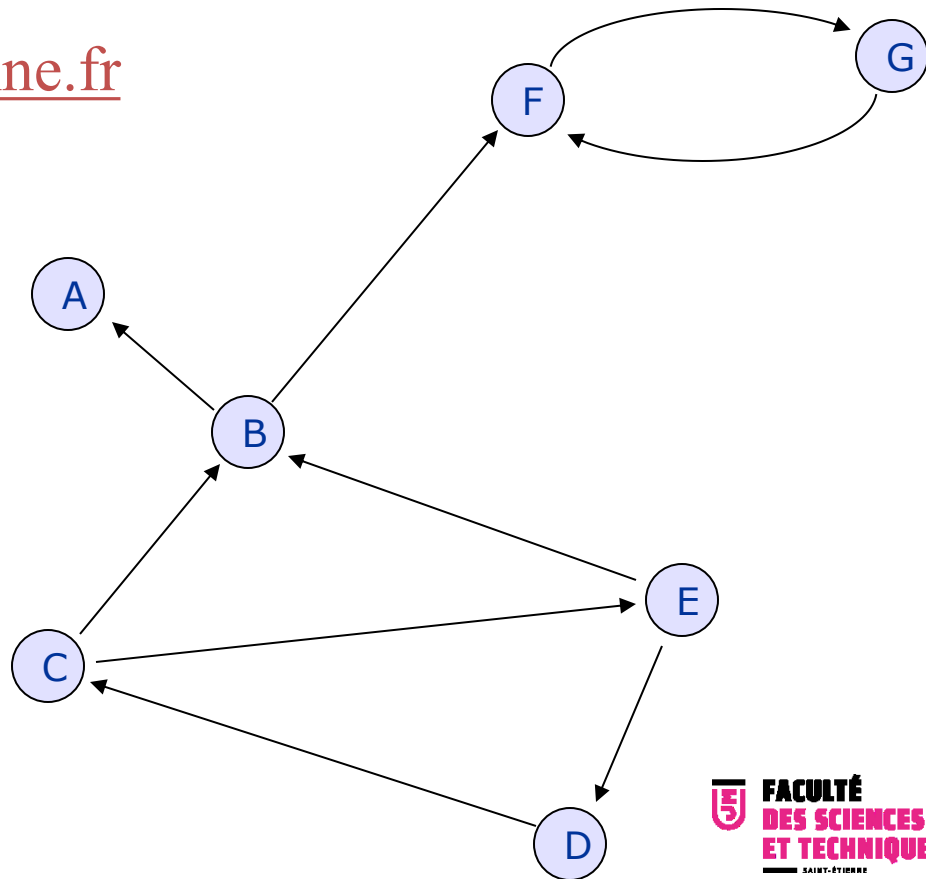
Document « The Haendel House »
(Collection « INEX Wikipédia »)

Compétition de RI
XML « INEX »

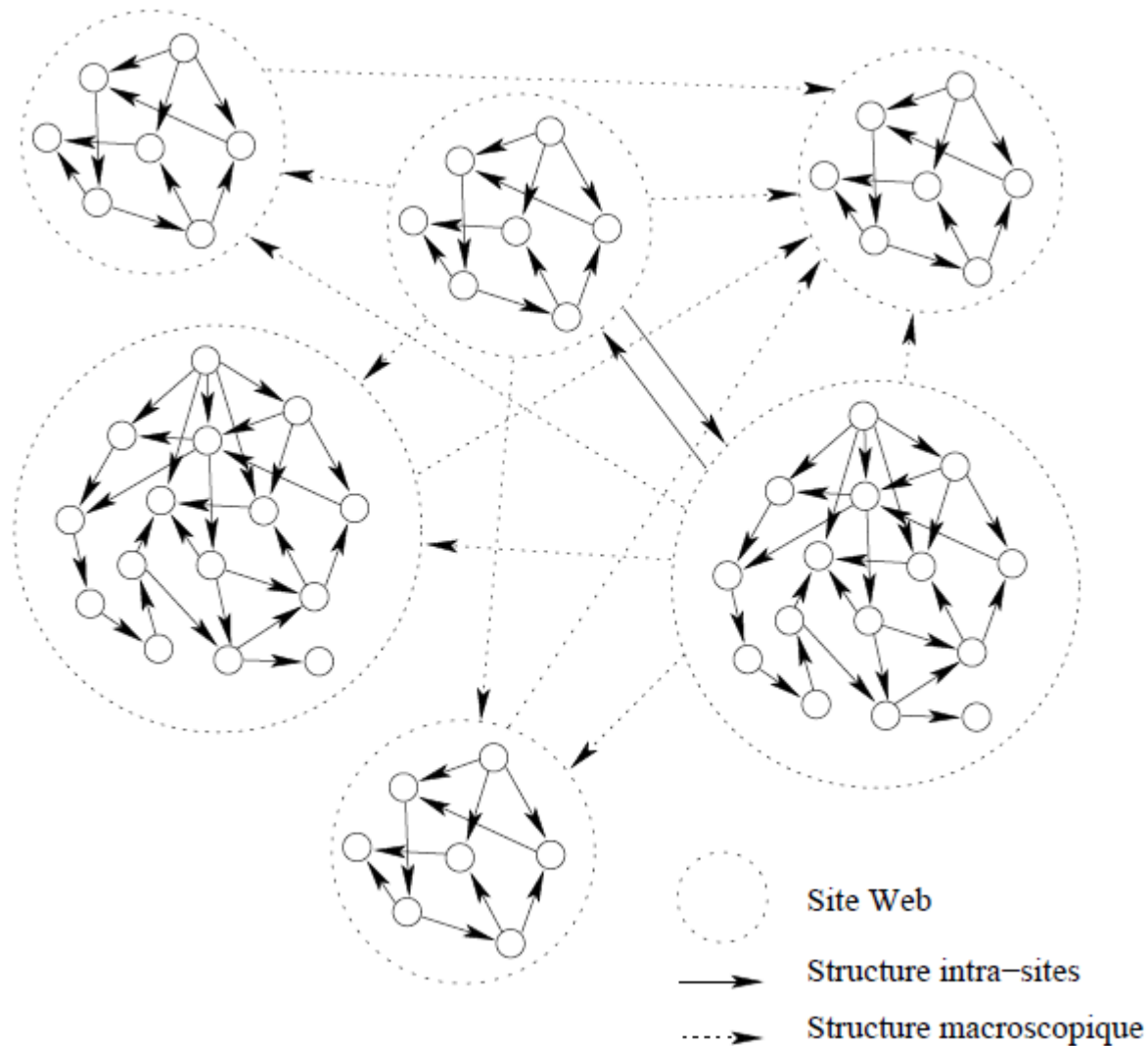
Balises (tags) :
article, bdy,
header, title,
image, caption,
sec, p, etc.

Web : Hypertexte ?

- Hypertexte : représentation non linéaire d'une information textuelle sous la forme d'un graphe de nœuds connectés par des liens.
- URL : <http://www.univ-st-etienne.fr>
- Le Web est un hypertexte :
 - Liens hypertextes, URLs
 - Structure de graphe !
- Graphe interne à un site :
 - Nœuds = pages Web,
 - Arcs = liens hypertextes.
- Graphe inter-sites :
 - Nœuds = sites Web,
 - Arcs = liens hypertextes.

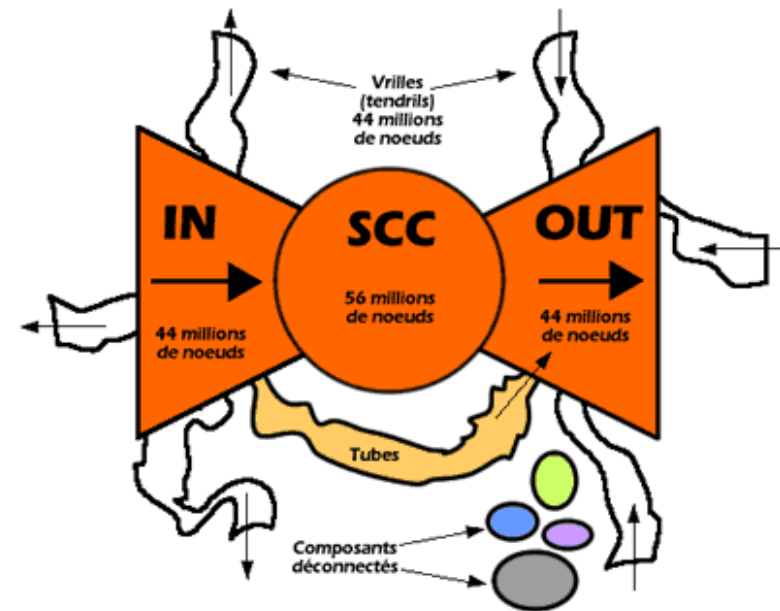


Web : Hypertexte ?



Le graphe du Web

- Graphe orienté :
 - Nœud = pages Web, Arcs = liens hypertexte.
 - Degré entrant/sortant d'un nœud.
 - Chemin : suite de nœuds reliés par des arcs.
- Calculs « classiques » dans le Web vu comme un graphe :
 - Distance : longueur du plus petit chemin (ppc) entre 2 nœuds.
 - Distribution des degrés.
 - Diamètre : moyenne ppc entre toutes les paires de nœuds (19 clics [Albert 99]).
 - Connectivité, ex. : composantes fortement connexes (« nœud papillon » [IBM 99]).



Hypertexte en XML

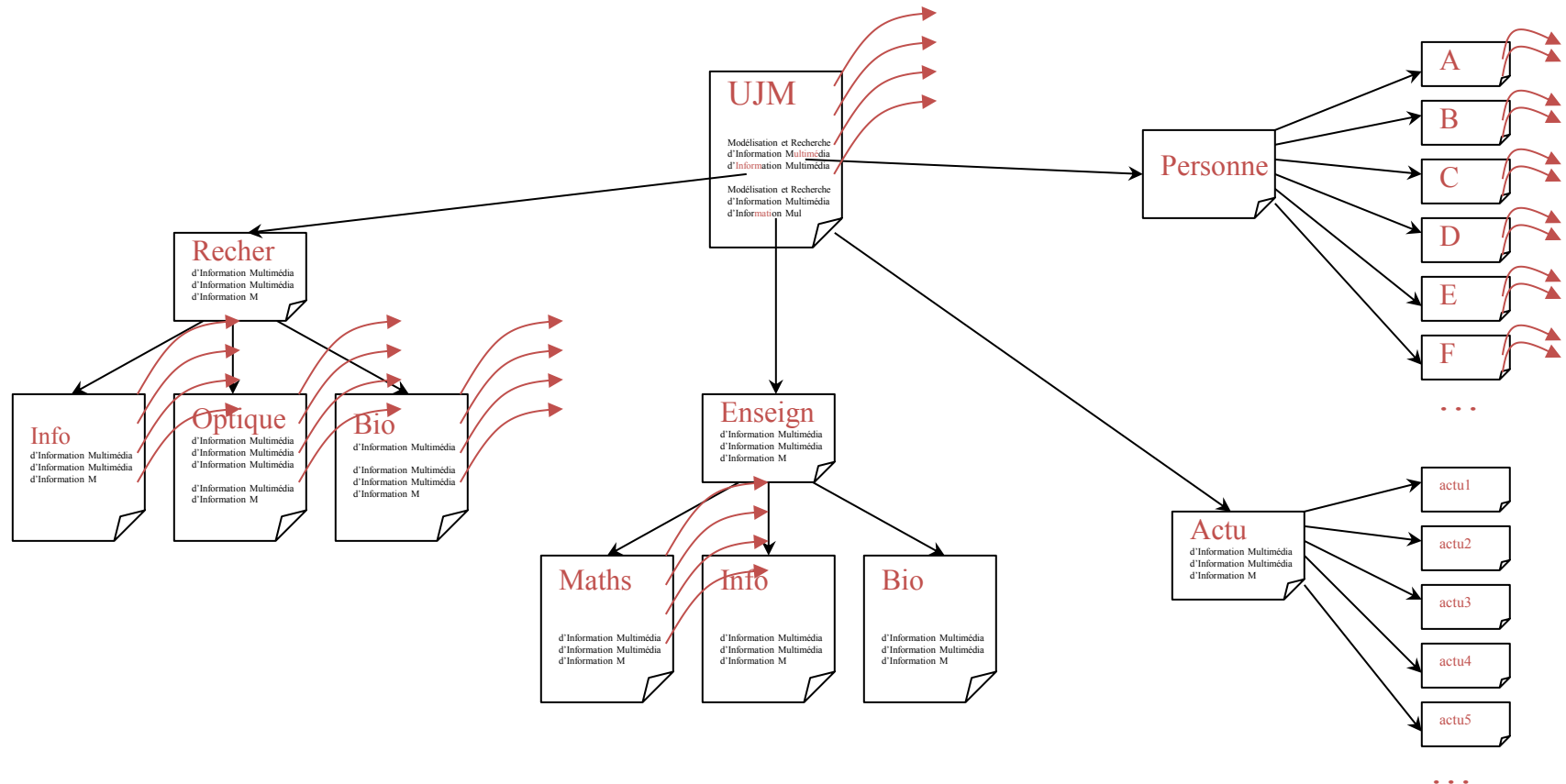
```
<article xmlns:xlink="http://www.w3.org/1999/xlink">
  <header>
    <title>Handel House Museum</title>
    <id>1707709</id>
  </header>
  <bdy>
    <image width="150px" src="London_Handel_House.jpg" type="thumb">
      <caption>Handel House. Note the <link xlink:href="310649.xml">blue plaque</link>
    </caption>
    </image>
    <p>
      The <b>Handel House Museum</b> at 25 <link xlink:href="2599649.xml">Brook
      Street</link>, in the exclusive central <link xlink:href="17867.xml">London</link>
      district of <link xlink:href="94167.xml">Mayfair</link> was the home of the <link
      xlink:href="11867.xml">German</link> born <link xlink:href="4500.xml">baroque</link>
      composer <link xlink:href="12775.xml">George Frideric Handel</link>
      from 1723 until his death at the house in 1759. He composed works such as
      <it><link xlink:href="149131.xml">The Messiah</link></it>, <it><link
      xlink:href="811987.xml">Zadok the Priest</link></it> and the <it><link
      xlink:href="1246814.xml">Fireworks Music</link></it> there.
    </p>
    <sec>
      <st>The museum</st>
    </p>
    The house has been restored to look as it did during Handel's occupancy. A typical
    early 18th century London terrace house, it comprises a basement, three main storeys and an
    attic, and Handel was the first occupant. The attic was later converted into a fourth full
    floor. The ground floor is now a music and gift shop and the upper floors are leased to a
    charity called the Handel House Trust, and have been open to the public since 8 November
    2001. The interiors have been restored to the somewhat spartan style of Georgian era,
    using mostly architectural elements from elsewhere, as other than the staircase, few of
    the original interior features survived. The Handel House Collection Trust has assembled a
    collection of Handel memorabilia, including the Byrne Collection of several hundred items,
    which was acquired in 1998.
  </p>
</sec>
</bdy>
</article>
```

Document « The Haendel House »
(Collection « INEX Wikipédia »)

Document « The
Haendel House »

Balise : link !

DS ou hypertexte ?



→ Relations de référence

→ Relations de composition

Web : base de données ?

- Niveaux de structure fine.
- XML : Data-centric vs Document-centric.
- Import/export BD.
- Exemple data-centric :

```
<article>
  <title>RI Structurée</title>
  <author idAut="443">Mathias Géry</author>
  <editor idEdt="121">Springer</author>
  <nbpages>120</nbpages>
  <prix>30</prix>
</article>
```

- Cf. Web Sémantique...

Documents du Web ?

- En RI, un « document » :
 - Ce qu'on indexe,
 - Ce qu'on retrouve → granularité de la réponse.
- Qu'est-ce qu'un document sur le Web ?
 - Un fichier HTML ? URL : <http://www.univ-st-etienne.fr>
 - Un morceau de fichier HTML ? Paragraphe ? Phrase ?
 - Une suite de fichiers HTML ? Avec les liens ?
 - Un site Web ? Une partie d'un site Web ?
 - Plusieurs sites Web ?
 - Un flux RSS ?
 - Un document virtuel (fragments) personnalisé ?

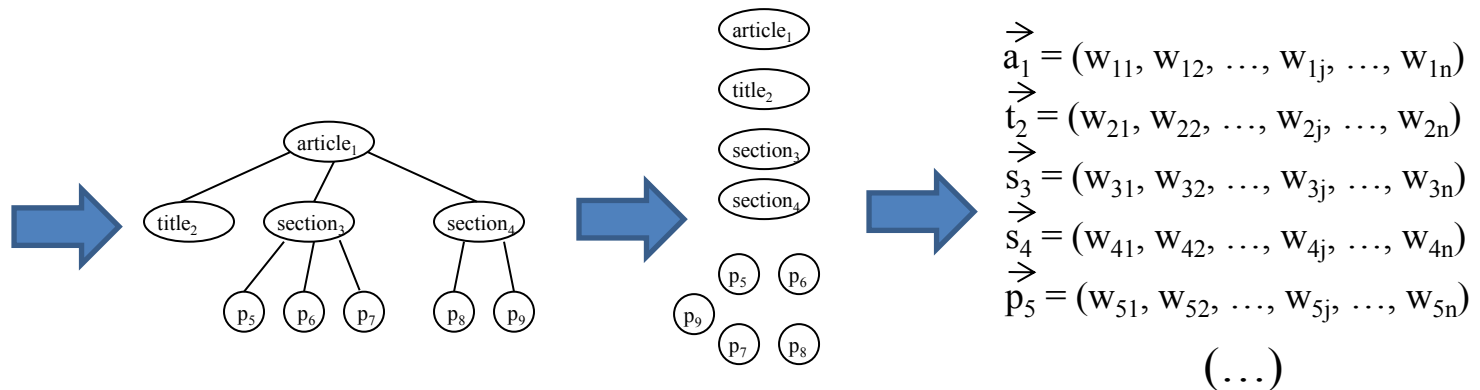
Recherche d'Information (RI)

- Plan du cours « RI Structurée » (RIS) :
 - 5. RI et structure intra-document (ex. XML).
 - Introduction : RI classique.
 - Web (HTML, XML) : Document Structuré ou Hypertexte ?
 - RI Structurée :
 - Langages de requêtes structurés.
 - Indexation en RIS.
 - Passage Retrieval, modèle à proximité.
 - Pondération en RIS : normalisation, types d'éléments, propagation.
 - Évaluation en RI : la compétition INEX.
 - 6. RI et structure inter-documents (ex. les liens hypertextes).
 - 7. Projet : participation à une mini-compétition de RIS.

RIS et modèles classiques

- Modèles classiques (booléens, VSM, probabiliste, ML, etc.) :
 - Documents atomiques et indépendants.
 - Exemple : vecteurs pondérés (modèle vectoriel).

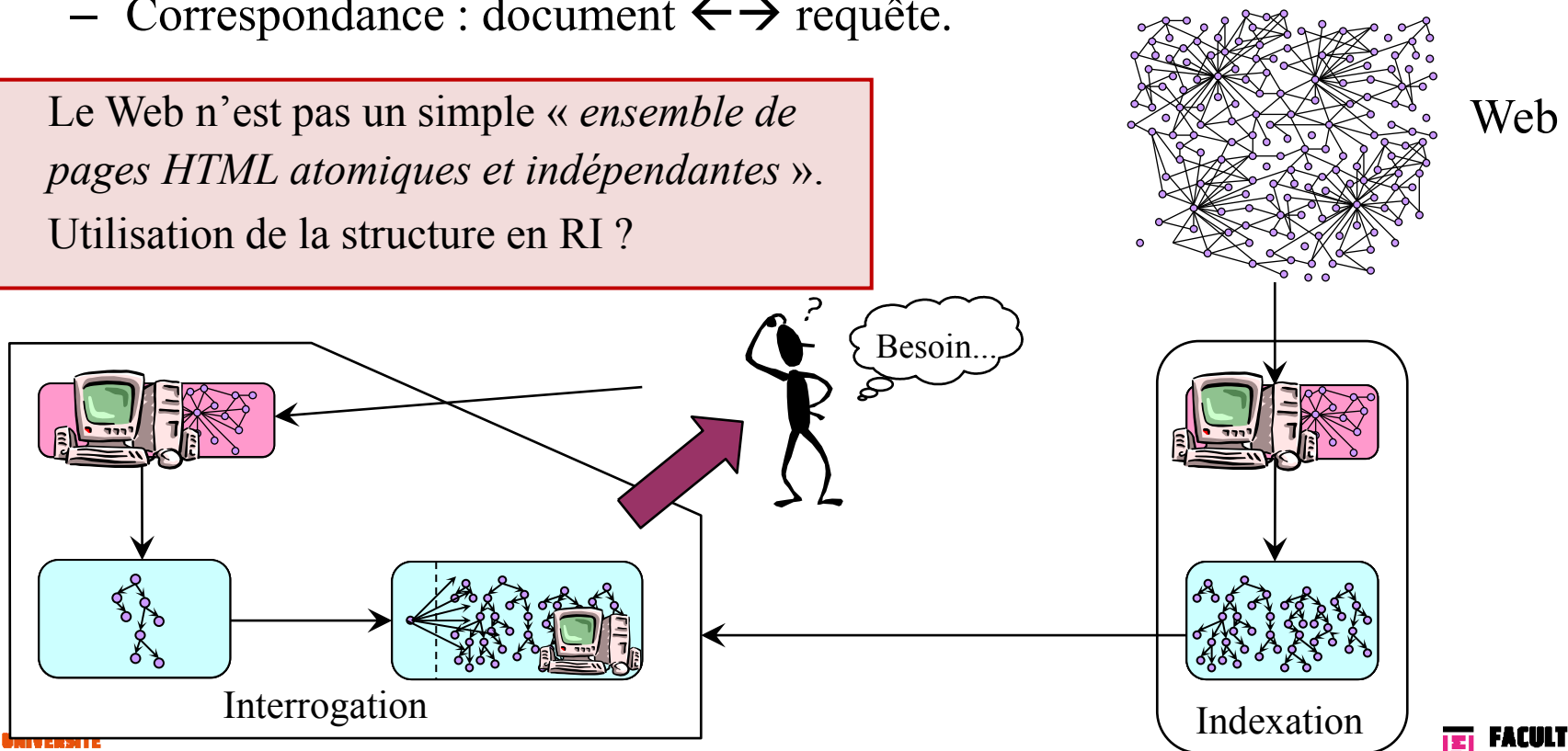
- Ce qu'on peut faire en RIS :



Comment exploiter la structure en RI ?

- Modèles classiques (booléen, vectoriel) :
 - L'unité d'indexation (document) est la page HTML.
 - Correspondance : document \leftrightarrow requête.

- Le Web n'est pas un simple « *ensemble de pages HTML atomiques et indépendantes* ».
- Utilisation de la structure en RI ?



RIS & RIC

- Using the Structure of Documents to Improve the Discovery of Unexpected Information¹

ANALYTICS

In this paper we are interested in what the current mathematical tools have to say about the behavior of the solutions of the Burgers equation in the case of a random initial condition. The Burgers equation is a nonlinear partial differential equation that arises in the study of fluid dynamics and is one of the simplest models of turbulence. The equation is given by

$$u_t + uu_x = \nu u_{xx}$$

where u is the velocity, x is the spatial coordinate, t is time, and ν is the viscosity. The equation is solved on the interval $x \in [0, 1]$ with periodic boundary conditions. The initial condition is a random function $u(x, 0) = u_0(x)$ with a power-law spectrum. The equation is solved numerically using a finite difference method. The results show that the solution develops a singularity at a finite time, which is a characteristic feature of turbulence. The singularity is characterized by a power-law growth of the velocity gradient, which is a universal feature of the Burgers equation.

Keywords: Burgers equation, random initial condition, singularity, turbulence.

1. INTRODUCTION

The Burgers equation is a nonlinear partial differential equation that arises in the study of fluid dynamics and is one of the simplest models of turbulence. The equation is given by

$$u_t + uu_x = \nu u_{xx}$$

where u is the velocity, x is the spatial coordinate, t is time, and ν is the viscosity. The equation is solved on the interval $x \in [0, 1]$ with periodic boundary conditions. The initial condition is a random function $u(x, 0) = u_0(x)$ with a power-law spectrum. The equation is solved numerically using a finite difference method. The results show that the solution develops a singularity at a finite time, which is a characteristic feature of turbulence. The singularity is characterized by a power-law growth of the velocity gradient, which is a universal feature of the Burgers equation.

2. THE BURGERS EQUATION

The Burgers equation is a nonlinear partial differential equation that arises in the study of fluid dynamics and is one of the simplest models of turbulence. The equation is given by

$$u_t + uu_x = \nu u_{xx}$$

where u is the velocity, x is the spatial coordinate, t is time, and ν is the viscosity. The equation is solved on the interval $x \in [0, 1]$ with periodic boundary conditions. The initial condition is a random function $u(x, 0) = u_0(x)$ with a power-law spectrum. The equation is solved numerically using a finite difference method. The results show that the solution develops a singularity at a finite time, which is a characteristic feature of turbulence. The singularity is characterized by a power-law growth of the velocity gradient, which is a universal feature of the Burgers equation.

3. NUMERICAL RESULTS

The Burgers equation is a nonlinear partial differential equation that arises in the study of fluid dynamics and is one of the simplest models of turbulence. The equation is given by

$$u_t + uu_x = \nu u_{xx}$$

where u is the velocity, x is the spatial coordinate, t is time, and ν is the viscosity. The equation is solved on the interval $x \in [0, 1]$ with periodic boundary conditions. The initial condition is a random function $u(x, 0) = u_0(x)$ with a power-law spectrum. The equation is solved numerically using a finite difference method. The results show that the solution develops a singularity at a finite time, which is a characteristic feature of turbulence. The singularity is characterized by a power-law growth of the velocity gradient, which is a universal feature of the Burgers equation.

4. CONCLUSIONS

The Burgers equation is a nonlinear partial differential equation that arises in the study of fluid dynamics and is one of the simplest models of turbulence. The equation is given by

$$u_t + uu_x = \nu u_{xx}$$

where u is the velocity, x is the spatial coordinate, t is time, and ν is the viscosity. The equation is solved on the interval $x \in [0, 1]$ with periodic boundary conditions. The initial condition is a random function $u(x, 0) = u_0(x)$ with a power-law spectrum. The equation is solved numerically using a finite difference method. The results show that the solution develops a singularity at a finite time, which is a characteristic feature of turbulence. The singularity is characterized by a power-law growth of the velocity gradient, which is a universal feature of the Burgers equation.

REFERENCES

[1] Burgers, J. M. (1948). A simplified model for the study of turbulence. *Physica*, 1, 815-824.

[2] Kraichnan, R. H. (1971). The structure of isotropic turbulence at high Reynolds number. *Journal of Fluid Mechanics*, 5, 475-484.

[3] Frisch, U. (1995). *Turbulence: The Legacy of A. N. Kolmogorov*. Cambridge University Press.

[4] Gledzer, M. V., Ilyin, V. V., & Orszag, S. A. (1973). On the existence of singularities in the Burgers equation. *Physica Scripta*, 7, 1-6.

[5] Sulem, P. L., & Sulem, C. (1981). *Nonlinear Evolution Equations in Fluid Mechanics*. Cambridge University Press.

[6] Constantin, P., Foias, D., & Kuvshinov, V. G. (1988). *Navier-Stokes Equations*. Cambridge University Press.

[7] Frisch, U., & Sulem, P. L. (1983). A simple model for the Burgers equation. *Journal of Fluid Mechanics*, 119, 351-369.

[8] Kraichnan, R. H. (1975). The structure of isotropic turbulence at high Reynolds number. *Journal of Fluid Mechanics*, 7, 475-484.

[9] Frisch, U. (1995). *Turbulence: The Legacy of A. N. Kolmogorov*. Cambridge University Press.

[10] Gledzer, M. V., Ilyin, V. V., & Orszag, S. A. (1973). On the existence of singularities in the Burgers equation. *Physica Scripta*, 7, 1-6.

[11] Sulem, P. L., & Sulem, C. (1981). *Nonlinear Evolution Equations in Fluid Mechanics*. Cambridge University Press.

[12] Constantin, P., Foias, D., & Kuvshinov, V. G. (1988). *Navier-Stokes Equations*. Cambridge University Press.

[13] Frisch, U., & Sulem, P. L. (1983). A simple model for the Burgers equation. *Journal of Fluid Mechanics*, 119, 351-369.

[14] Kraichnan, R. H. (1975). The structure of isotropic turbulence at high Reynolds number. *Journal of Fluid Mechanics*, 7, 475-484.

[15] Frisch, U. (1995). *Turbulence: The Legacy of A. N. Kolmogorov*. Cambridge University Press.

[16] Gledzer, M. V., Ilyin, V. V., & Orszag, S. A. (1973). On the existence of singularities in the Burgers equation. *Physica Scripta*, 7, 1-6.

[17] Sulem, P. L., & Sulem, C. (1981). *Nonlinear Evolution Equations in Fluid Mechanics*. Cambridge University Press.

[18] Constantin, P., Foias, D., & Kuvshinov, V. G. (1988). *Navier-Stokes Equations*. Cambridge University Press.

[19] Frisch, U., & Sulem, P. L. (1983). A simple model for the Burgers equation. *Journal of Fluid Mechanics*, 119, 351-369.

[20] Kraichnan, R. H. (1975). The structure of isotropic turbulence at high Reynolds number. *Journal of Fluid Mechanics*, 7, 475-484.

[21] Frisch, U. (1995). *Turbulence: The Legacy of A. N. Kolmogorov*. Cambridge University Press.

[22] Gledzer, M. V., Ilyin, V. V., & Orszag, S. A. (1973). On the existence of singularities in the Burgers equation. *Physica Scripta*, 7, 1-6.

[23] Sulem, P. L., & Sulem, C. (1981). *Nonlinear Evolution Equations in Fluid Mechanics*. Cambridge University Press.

[24] Constantin, P., Foias, D., & Kuvshinov, V. G. (1988). *Navier-Stokes Equations*. Cambridge University Press.

[25] Frisch, U., & Sulem, P. L. (1983). A simple model for the Burgers equation. *Journal of Fluid Mechanics*, 119, 351-369.

[26] Kraichnan, R. H. (1975). The structure of isotropic turbulence at high Reynolds number. *Journal of Fluid Mechanics*, 7, 475-484.

[27] Frisch, U. (1995). *Turbulence: The Legacy of A. N. Kolmogorov*. Cambridge University Press.

[28] Gledzer, M. V., Ilyin, V. V., & Orszag, S. A. (1973). On the existence of singularities in the Burgers equation. *Physica Scripta*, 7, 1-6.

[29] Sulem, P. L., & Sulem, C. (1981). *Nonlinear Evolution Equations in Fluid Mechanics*. Cambridge University Press.

[30] Constantin, P., Foias, D., & Kuvshinov, V. G. (1988). *Navier-Stokes Equations*. Cambridge University Press.

[31] Frisch, U., & Sulem, P. L. (1983). A simple model for the Burgers equation. *Journal of Fluid Mechanics*, 119, 351-369.

[32] Kraichnan, R. H. (1975). The structure of isotropic turbulence at high Reynolds number. *Journal of Fluid Mechanics*, 7, 475-484.

[33] Frisch, U. (1995). *Turbulence: The Legacy of A. N. Kolmogorov*. Cambridge University Press.

[34] Gledzer, M. V., Ilyin, V. V., & Orszag, S. A. (1973). On the existence of singularities in the Burgers equation. *Physica Scripta*, 7, 1-6.

[35] Sulem, P. L., & Sulem, C. (1981). *Nonlinear Evolution Equations in Fluid Mechanics*. Cambridge University Press.

[36] Constantin, P., Foias, D., & Kuvshinov, V. G. (1988). *Navier-Stokes Equations*. Cambridge University Press.

[37] Frisch, U., & Sulem, P. L. (1983). A simple model for the Burgers equation. *Journal of Fluid Mechanics*, 119, 351-369.

[38] Kraichnan, R. H. (1975). The structure of isotropic turbulence at high Reynolds number. *Journal of Fluid Mechanics*, 7, 475-484.

[39] Frisch, U. (1995). *Turbulence: The Legacy of A. N. Kolmogorov*. Cambridge University Press.

[40] Gledzer, M. V., Ilyin, V. V., & Orszag, S. A. (1973). On the existence of singularities in the Burgers equation. *Physica Scripta*, 7, 1-6.

[41] Sulem, P. L., & Sulem, C. (1981). *Nonlinear Evolution Equations in Fluid Mechanics*. Cambridge University Press.

[42] Constantin, P., Foias, D., & Kuvshinov, V. G. (1988). *Navier-Stokes Equations*. Cambridge University Press.

[43] Frisch, U., & Sulem, P. L. (1983). A simple model for the Burgers equation. *Journal of Fluid Mechanics*, 119, 351-369.

[44] Kraichnan, R. H. (1975). The structure of isotropic turbulence at high Reynolds number. *Journal of Fluid Mechanics*, 7, 475-484.

[45] Frisch, U. (1995). *Turbulence: The Legacy of A. N. Kolmogorov*. Cambridge University Press.

[46] Gledzer, M. V., Ilyin, V. V., & Orszag, S. A. (1973). On the existence of singularities in the Burgers equation. *Physica Scripta*, 7, 1-6.

[47] Sulem, P. L., & Sulem, C. (1981). *Nonlinear Evolution Equations in Fluid Mechanics*. Cambridge University Press.

[48] Constantin, P., Foias, D., & Kuvshinov, V. G. (1988). *Navier-Stokes Equations*. Cambridge University Press.

[49] Frisch, U., & Sulem, P. L. (1983). A simple model for the Burgers equation. *Journal of Fluid Mechanics*, 119, 351-369.

[50] Kraichnan, R. H. (1975). The structure of isotropic turbulence at high Reynolds number. *Journal of Fluid Mechanics*, 7, 475-484.

[51] Frisch, U. (1995). *Turbulence: The Legacy of A. N. Kolmogorov*. Cambridge University Press.

[52] Gledzer, M. V., Ilyin, V. V., & Orszag, S. A. (1973). On the existence of singularities in the Burgers equation. *Physica Scripta*, 7, 1-6.

[53] Sulem, P. L., & Sulem, C. (1981). *Nonlinear Evolution Equations in Fluid Mechanics*.

-
- Diagram illustrating a hierarchical tree structure for document classification. The root node is article_1 . It branches into title_2 (highlighted in red), section_3 , and section_4 . section_3 branches into p_5 , p_6 , and p_7 . section_4 branches into p_8 and p_9 . Two curved arrows labeled $t_1 : w_{1,1}$ point from the leaf nodes p_8 and p_9 back to the root node article_1 .

- 
- FACULTÉ
DES SCIENCES
ET TECHNIQUES**
SAINT-ÉTIENNE

Recherche d'Information (RI)

- Plan du cours « RI Structurée » (RIS) :
 - 5. RI et structure intra-document (ex. XML).
 - Introduction : RI classique.
 - Web (HTML, XML) : Document Structuré ou Hypertexte ?
 - RI Structurée :
 - Langages de requêtes structurés.
 - Indexation en RIS.
 - Passage Retrieval, modèle à proximité.
 - Pondération en RIS : normalisation, types d'éléments, propagation.
 - Évaluation en RI : la compétition INEX.
 - 6. RI et structure inter-documents (ex. les liens hypertextes).
 - 7. Projet : participation à une mini-compétition de RIS.

Interrogation : Requêtes structurées

- Exprimer des contraintes sur la structure des documents recherchés.
 - Data-centric (BD) vs **Document-centric (RI)**.
 - BD : correspondance exacte, RI : classement par pertinence.
 - Langages de requêtes, exemples : XML-QL, WebSQL, XQL, XPath, NEXI, XQuery (W3C), etc.
 - Exemples (document-centric) :
 - Un livre contenant un chapitre parlant de « course à pied ».
 - Une section parlant d'évaluation de RI XML, faisant partie d'un chapitre sur l'évaluation en RI.

Langage de requête NEXI

- NEXI (Narrowed Extended XPath I) [Trotman2005] :
 - Simplification de XPath.
 - Compétition INEX.

```
//article[about(../bdy, "information retrieval")]//  
        section[about(.,xml) and about(.,retrieval)]
```

- On cherche une *section* : //article//section.
- Parlant de « xml » et de « retrieval ».
- Contenue dans un article contenant un *bdy* parlant de "information retrieval".

Recherche d'Information (RI)

- Plan du cours « RI Structurée » (RIS) :
 - 5. RI et structure intra-document (ex. XML).
 - Introduction : RI classique.
 - Web (HTML, XML) : Document Structuré ou Hypertexte ?
 - RI Structurée :
 - Langages de requêtes structurés.
 - Indexation en RIS.
 - Passage Retrieval, modèle à proximité.
 - Pondération en RIS : normalisation, types d'éléments, propagation.
 - Évaluation en RI : la compétition INEX.
 - 6. RI et structure inter-documents (ex. les liens hypertextes).
 - 7. Projet : participation à une mini-compétition de RIS.

Indexation en RIS

- Étapes de l'indexation en RI Structurée :
 - Déterminer la/les granularité(s) des réponses :
 - Article, section, paragraphe, etc.
 - Les 3 à la fois ?
 - Déterminer les unités d'indexation.
 - Autoriser le recouvrement ?
 - Conserver la structure dans l'index ?
 - Parser le XML.
 - Étapes habituelles : tokenisation, anti-dico, lemmatisation, pondération, etc.
 - **Pondération** :
 - Prise en compte du type d'élément.
 - Prise en compte de la position de l'élément dans la structure.
 - Prise en compte du contexte d'un élément (ascendants, frères).
 - Prise en compte de la composition (descendants).

Déterminer la granularité / les unités d'indexation

- Placer le curseur entre :
 - Indexation des feuilles « Leaf-Only ».
 - Indexation de tous les nœuds (➔ recouvrement).
 - Autoriser recouvrement à l'indexation ?
- Conserver la structure :
 - Afin de considérer les relations entre les éléments à l'interrogation.
 - ➔ Ranking.
 - ➔ Eliminer recouvrement à l'interrogation.
- Très souvent, on divise l'articles en éléments ➔ on se ramène à un modèle classique (booléen, vectoriel, etc.).

Recherche d'Information (RI)

- Plan du cours « RI Structurée » (RIS) :
 - 5. RI et structure intra-document (ex. XML).
 - Introduction : RI classique.
 - Web (HTML, XML) : Document Structuré ou Hypertexte ?
 - RI Structurée :
 - Langages de requêtes structurés.
 - Indexation en RIS.
 - Passage Retrieval, modèle à proximité.
 - Pondération en RIS : normalisation, types d'éléments, propagation.
 - Évaluation en RI : la compétition INEX.
 - 6. RI et structure inter-documents (ex. les liens hypertextes).
 - 7. Projet : participation à une mini-compétition de RIS.

Indexation : « passage retrieval »

- Cf. cours M. Beigbeder.
- Un cas particulier (et historique) : « Passage Retrieval ».
 - Identifier des « passages pertinents ».
 - [Callan94], [Hearst97], [Salton96].
- Passage :
 - Fenêtre de mots de taille fixe.
 - Sélection de paragraphes.
 - Segmentation thématique, TextTiling [Hearst97].
- Utilisé dans les système de Q/A (Question / Réponse).

Modèle à proximité [Beigbeder2008]

- Approche locale (termes).
- Idée : il est préférable que les termes de la requête apparaissent proches dans les documents.
- Notion d'influence des termes.

- Requêtes booléennes :

$$p_{qOUq'}^d = \max(p_q^d, p_{q'}^d)$$

$$p_{qETq'}^d = \min(p_q^d, p_{q'}^d)$$

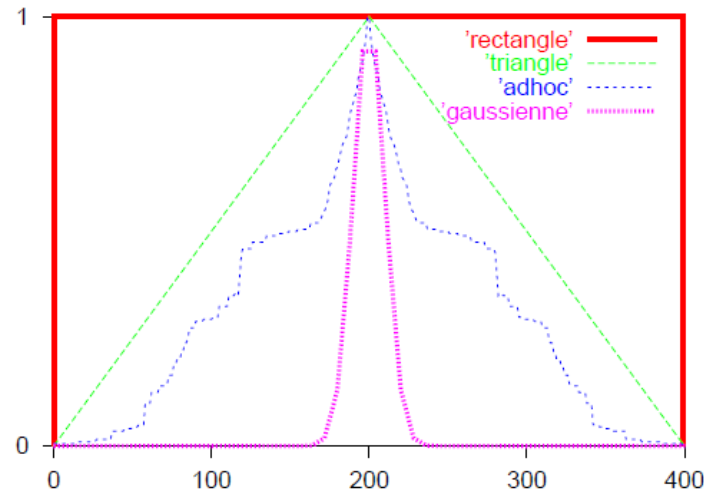
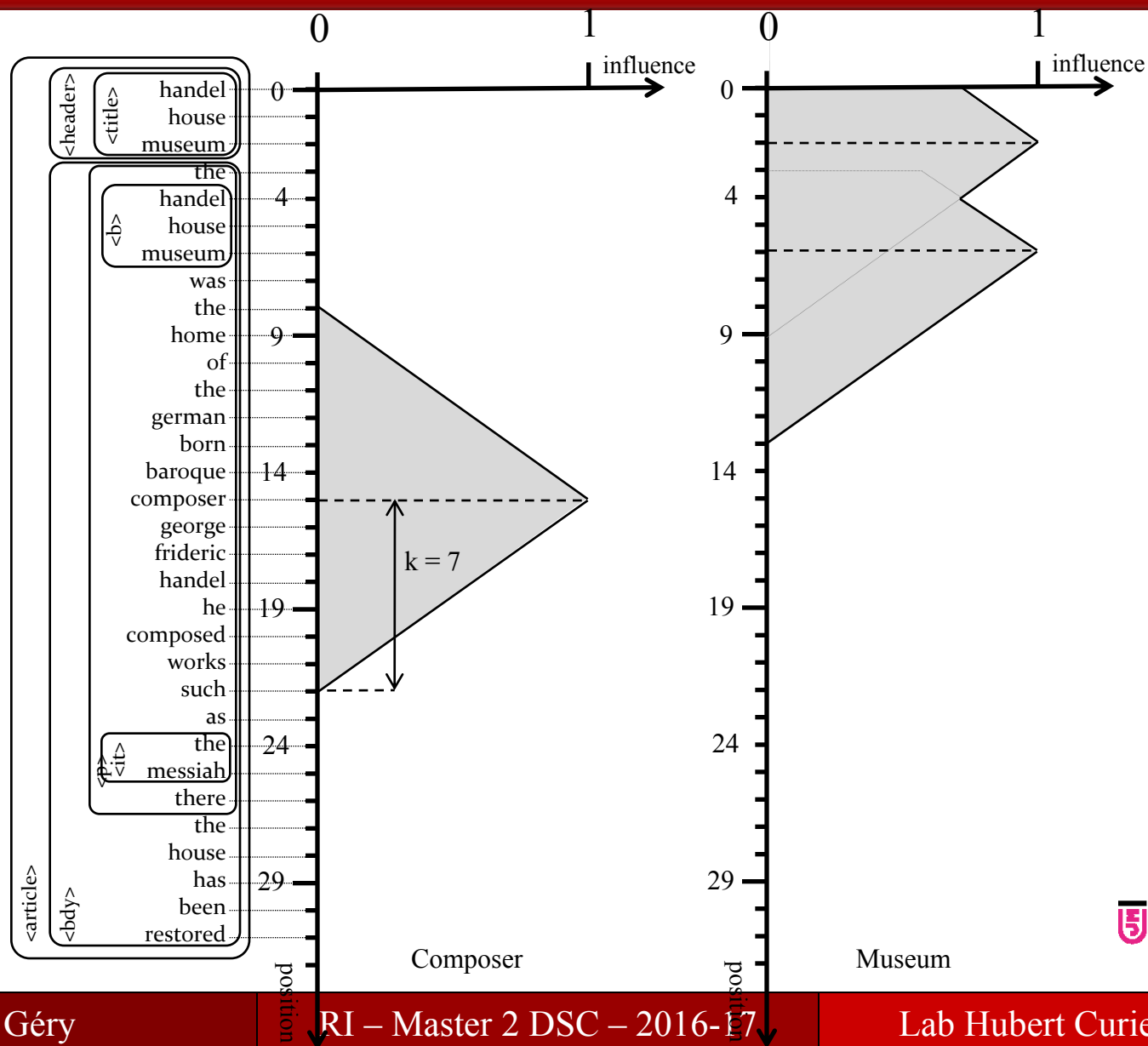


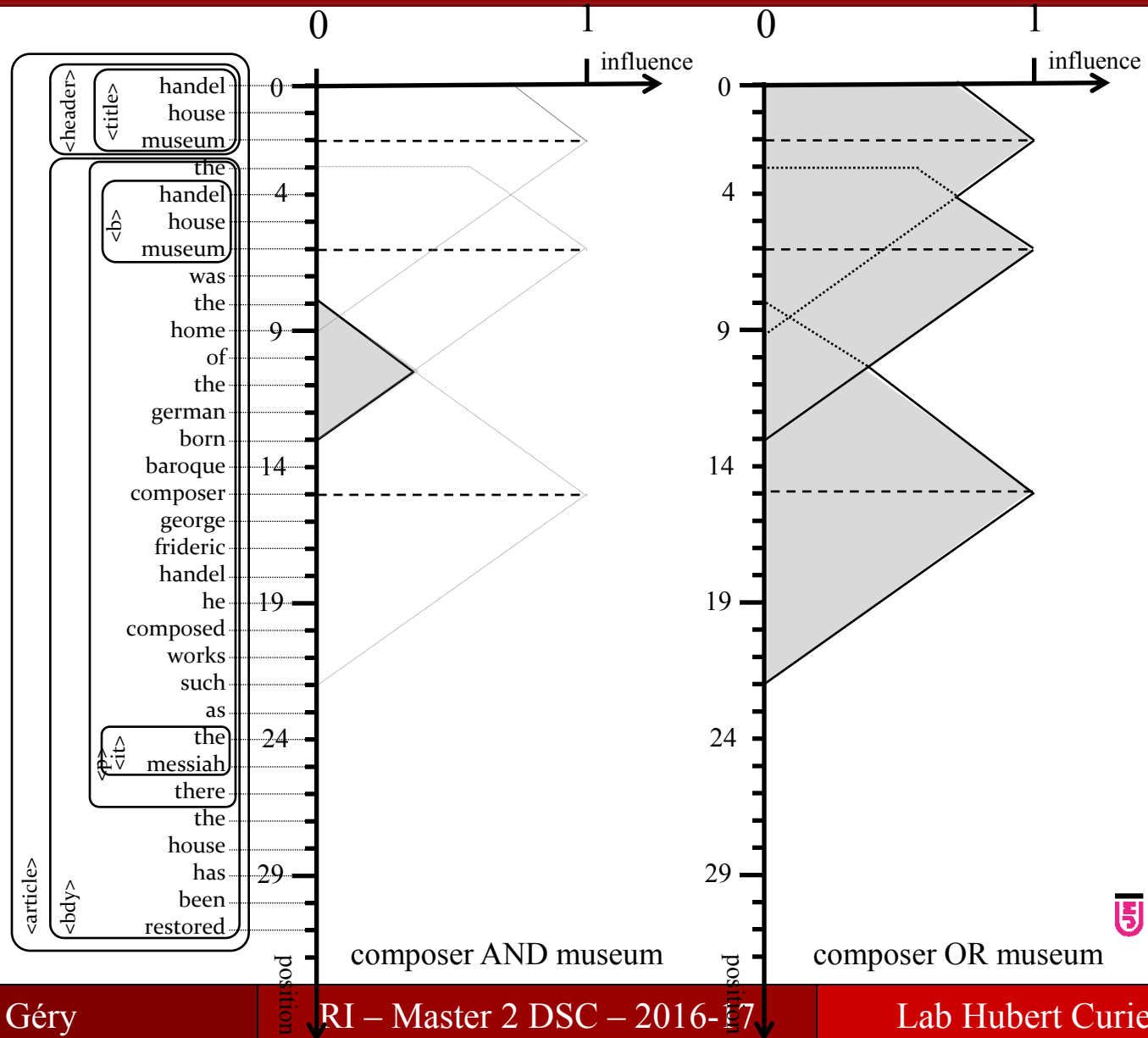
FIG. – Fonctions d'influence (rectangle, triangle, gaussienne, adhoc)

Source : M. Beigbeder

Modèle à proximité : influence



Modèle à proximité : requête AND / OR



Recherche d'Information (RI)

- Plan du cours « RI Structurée » (RIS) :
 - 5. RI et structure intra-document (ex. XML).
 - Introduction : RI classique.
 - Web (HTML, XML) : Document Structuré ou Hypertexte ?
 - RI Structurée :
 - Langages de requêtes structurés.
 - Indexation en RIS.
 - Passage Retrieval, modèle à proximité.
 - Pondération en RIS : normalisation, types d'éléments, propagation.
 - Évaluation en RI : la compétition INEX.
 - 6. RI et structure inter-documents (ex. les liens hypertextes).
 - 7. Projet : participation à une mini-compétition de RIS.

Pondération en RI classique (rappel)

- Pondération *tf.idf* classique ?
- Exemple SMART *ltn* :
 - Poids $w_{t,d}$ du terme t dans le document d :

$$w_{t,d} = TF_{t,d} * IDF_t = (1 + \log(tf_{t,d})) * \log\left(\frac{N}{df_t}\right)$$

– Avec :

- $TF_{t,d}$ = **représentativité** de t dans d
= par ex. : nombre d'occurrences de t dans d
- IDF_t = **discriminance** de t dans le corpus
= par ex. : nombre de documents contenant t
- N = nombre de documents dans le corpus.

$$\begin{aligned}\vec{d}_1 &= (w_{11}, w_{12}, \dots, w_{1j}, \dots, w_{1n}) \\ \vec{d}_2 &= (w_{21}, w_{22}, \dots, w_{2j}, \dots, w_{2n}) \\ \vec{d}_3 &= (w_{31}, w_{32}, \dots, w_{3j}, \dots, w_{3n})\end{aligned}$$

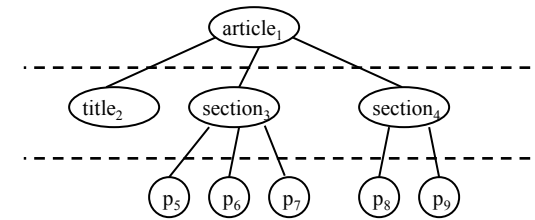
$$\vec{q} = (w_{q1}, w_{q2}, \dots, w_{qj}, \dots, w_{qn})$$

d/t	RICH	POOR
d_1	0.25	0.97
d_2	0.75	0.63
d_3	0.98	0.14
q	0.65	0.75

Matrice
termes-documents

Pondération en RIS : Problèmes ?

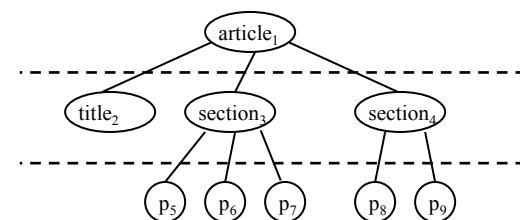
- Indexer des « documents » (éléments) en RIS :
 - Taille (très) variables,
 - Granularité : rôle / « importance » (très) variables,
 - Structure hiérarchique (inclusions / sous-parties) :
 - Composition : un article est composé de plusieurs champs.
 - Exemple : titre, abstract, section.
 - Champs répétables :
 - Exemple : section₁, section₂, etc.
 - Champs hiérarchiques :
 - Exemple : article composé de section₁ composée de p₁, p₂, p₃.
 - Dépendances (liens).



Pondération en RIS : Problèmes ?

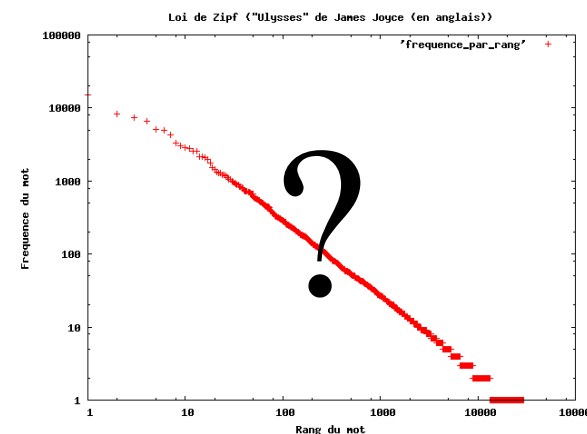
- Distribution des termes :

- Dans le document.
- Dans la collection.
- ➔ $TF_{t,d}$ et IDF_t



- Représentativité d'un terme pour un document :

- Un « document » ? Taille ?
- Une unité d'indexation ?
- Dans un titre / un résumé / un paragraphe ?



- Discriminance d'un terme :

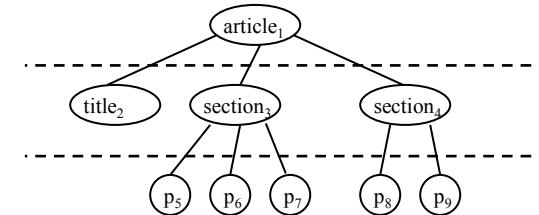
- Discriminer : documents ? Eléments ? Passages ?

Ranking en RIS : Problèmes ?

Classer

- Indexer des « documents » (éléments) en RIS :

- Taille (très) variables,
- Granularité : rôle / « importance » variables,
- Structure hiérarchique (et sous-parties) :
 - Composés de plusieurs champs.
 - Exemple : article composé de section₁, section₂, etc.
 - Exemple : article composé de section₁ composée de p₁, p₂, p₃.
- Dépendances (liens).



Rebelote

Pondération et ranking en RIS

- Quelques solutions :
 - Normalisation par la taille,
 - Types d'éléments,
 - Propagation :
 - des termes,
 - des poids des termes,
 - des scores de pertinence.

Pondération en RIS : normalisation (taille des documents)

- SMART *ltc* :
 - Idée : normaliser par rapport à la taille du document.
 - (racine du carré des composantes du vecteur).
 - T = nombre de termes du corpus.

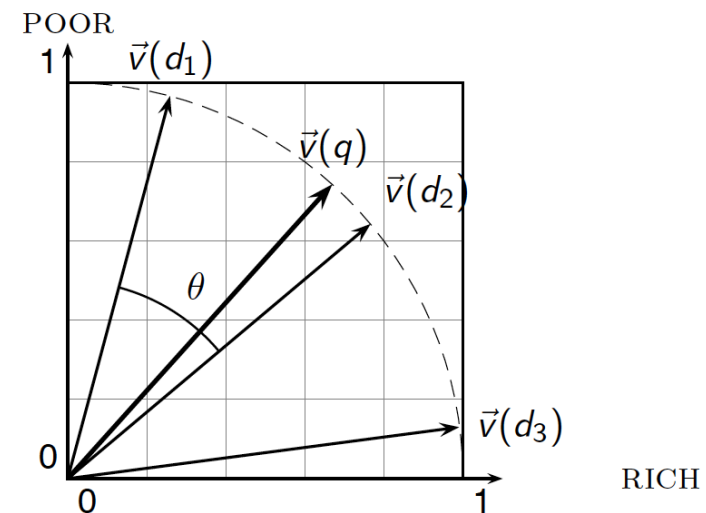
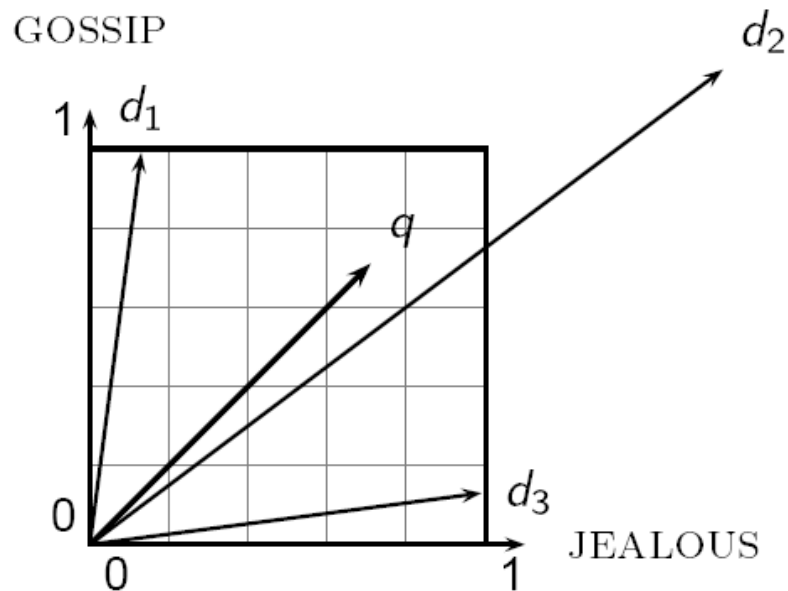
$$w_{t,d} = \frac{(1 + \log(tf_{t,d}))}{\sqrt{\sum_{i=1}^T w_{i,d}^2}} * \log\left(\frac{N}{df_t}\right)$$

- Remarque : c'est la **normalisation du cosinus** !

$$\cos(\vec{q}, \vec{d}) = \frac{\sum_{i=1}^T w_{i,d} * q_i}{\sqrt{\sum_{i=1}^T w_{i,d}^2} * \sqrt{\sum_{i=1}^T q_i^2}}$$

Pondération en RIS : normalisation (taille des documents)

- SMART *ltc* (normalisation cosinus) :



Source : P. Nayak, P. Raghavan, Stanford

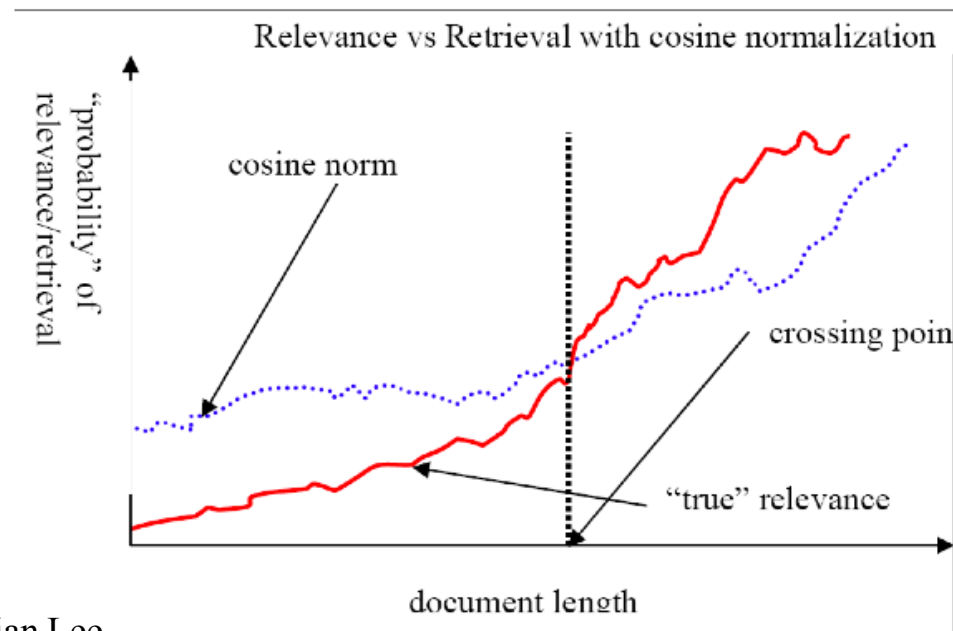
Autres fonctions de pondération : notation SMART

Term frequency		Document frequency		Normalization	
n (natural)	$tf_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1 + \log(tf_{t,d})$	t (idf)	$\log \frac{N}{df_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2 + w_2^2 + \dots + w_M^2}}$
a (augmented)	$0.5 + \frac{0.5 \times tf_{t,d}}{\max_t(tf_{t,d})}$	p (prob idf)	$\max\{0, \log \frac{N - df_t}{df_t}\}$	u (pivoted unique)	$1/u$
b (boolean)	$\begin{cases} 1 & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$			b (byte size)	$1/CharLength^\alpha$, $\alpha < 1$
L (log ave)	$\frac{1 + \log(tf_{t,d})}{1 + \log(\text{ave}_{t \in d}(tf_{t,d}))}$				

Source : P. Nayak, P. Raghavan, Stanford

Pondération en RIS : normalisation (taille des documents)

- Est-ce que la pertinence et la longueur des documents sont indépendants ?
 - Étude pertinence système vs pertinence utilisateur.
 - Constat : la normalisation du cosinus avantage trop les documents courts, et pénalise trop les documents longs.



Source : Lillian Lee

 **FACULTÉ
DES SCIENCES
ET TECHNIQUES**
Source : [Manning08]

Pondération en RIS : normalisation (taille des documents)

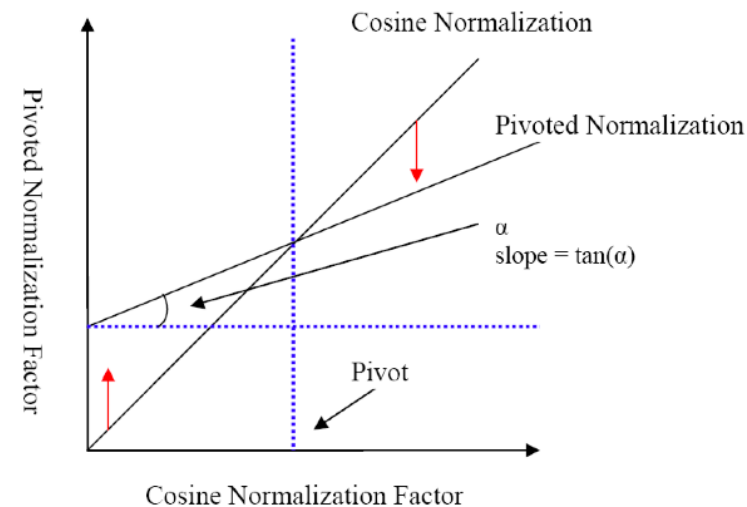
- Idée : un ajustement de la normalisation par la taille.
- SMART *lnu* : Pivoted length normalization [Singhal96] :

$$w_{t,d} = \frac{\frac{1 + \log(tf_{t,d})}{1 + \log(\frac{dl}{avgdl})}}{(1 - slope) * pivot + slope * nt_d}$$

– Avec :

- Avec dl = taille de d ,
- $avgdl$ = taille moyenne des documents,
- nt_d = #termes distincts dans d .
- $pivot$ = moyenne des nt_d .
- $slope$: importance de la normalisation.

Pivot normalization



Source : Lillian Lee

Pondération en RIS : normalisation (taille des documents)

- BM25 :

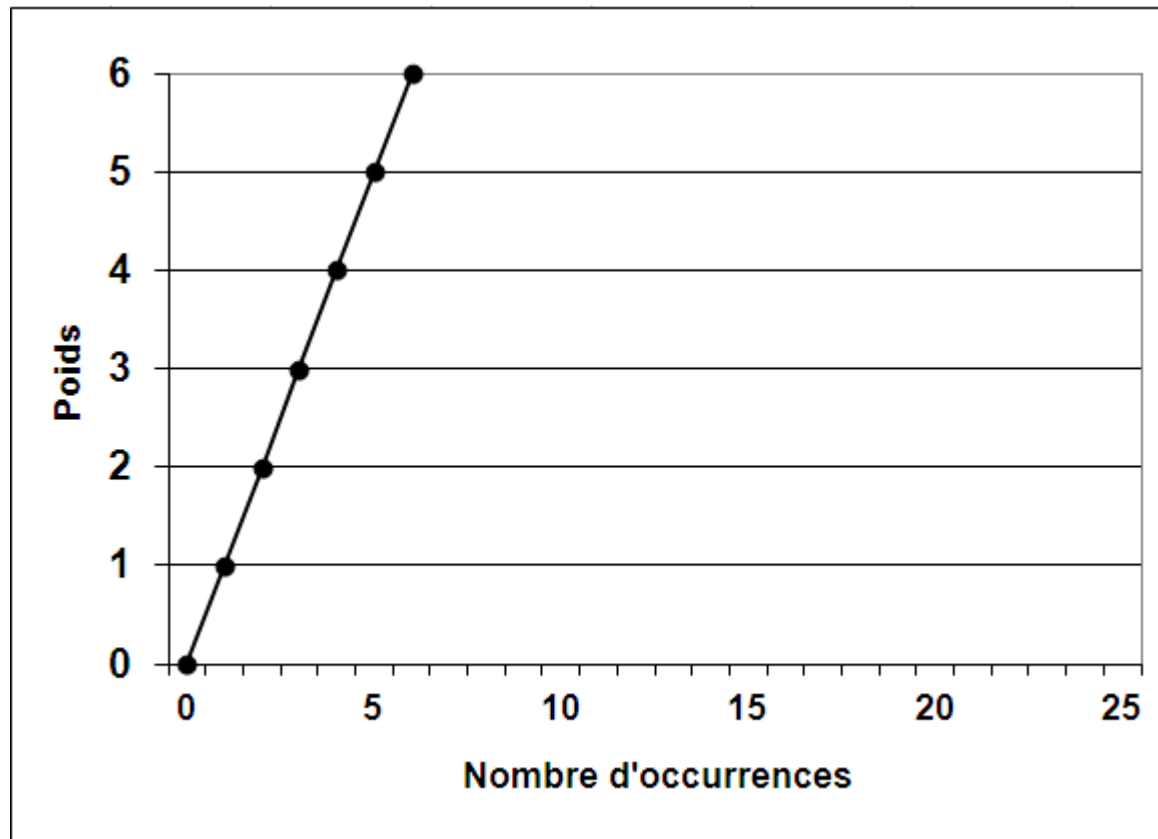
$$w_{t,d} = \frac{tf_{t,d} * (k_1 + 1)}{k_1 * \left((1 - b) + b * \frac{dl}{avdl} \right) + tf_{t,d}} * \log \left(\frac{N - df_t + 0,5}{df_t + 0,5} \right)$$

- Avec dl = taille de d ,
- $avdl$ = taille moyenne des documents,
- b = importance de la normalisation par la taille,
- k_1 = importance de la saturation du tf .

BM25 & normalisation

- Pas de normalisation par la taille ($b=0$).
- Pas de saturation du tf ($k=+\infty$).

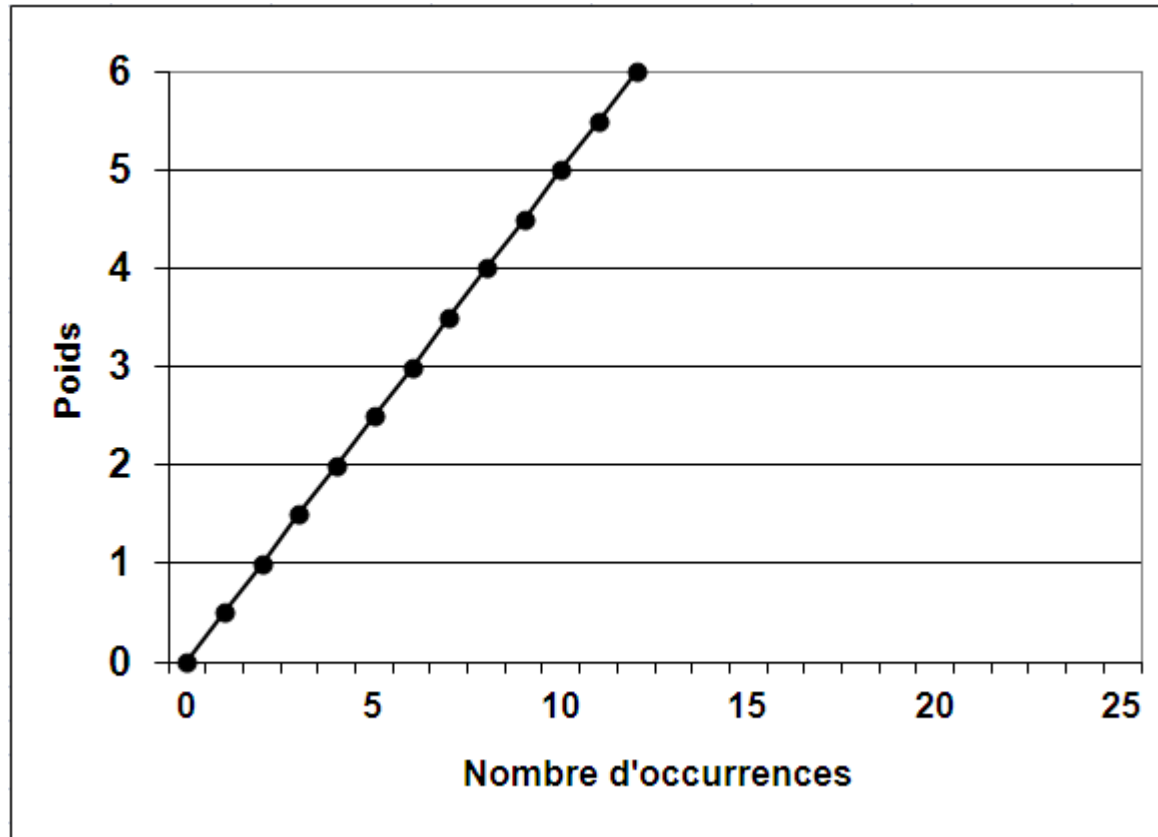
$$\begin{aligned} k_1 &= +\infty \\ b &= 0 \\ dl &= 3 * avdl \end{aligned}$$



BM25 & normalisation

- Légère normalisation par la taille ($b=0,5$).
- Pas de saturation du tf ($k=+\infty$).

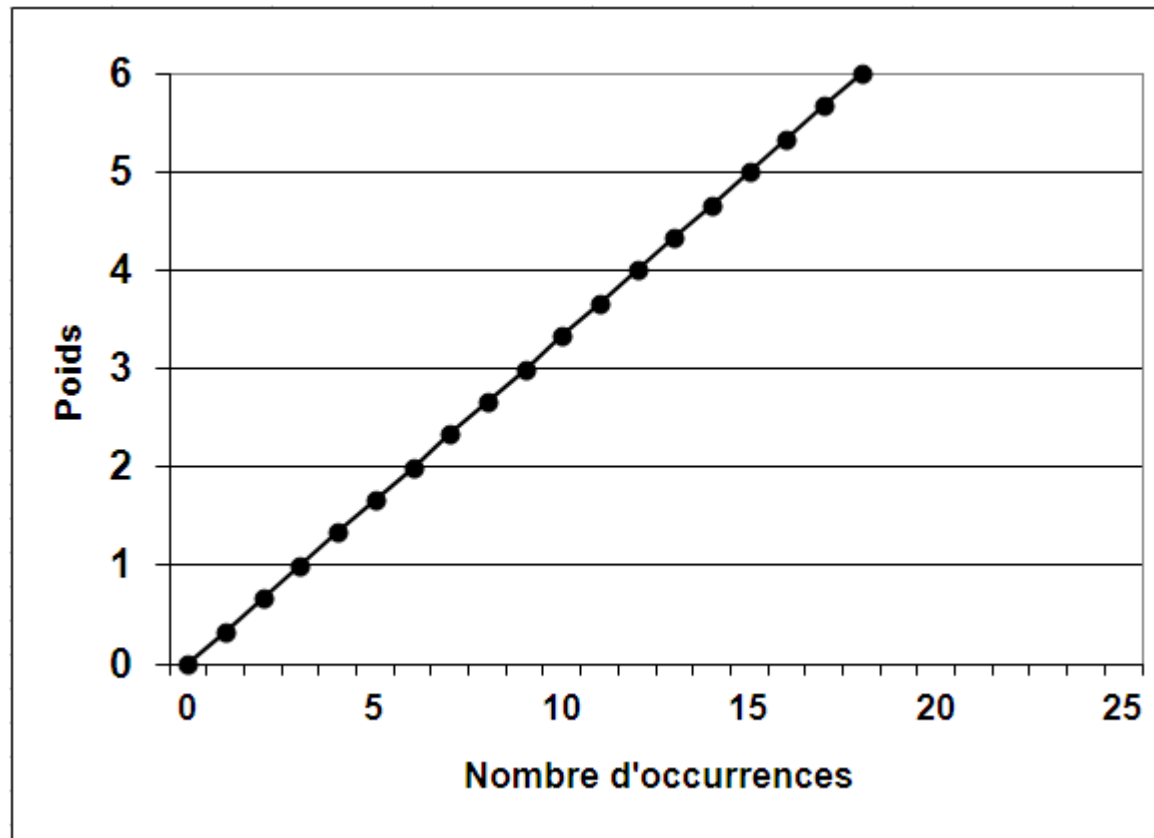
$$\begin{aligned} k_1 &= +\infty \\ b &= 0,5 \\ dl &= 3 * avdl \end{aligned}$$



BM25 & normalisation

- Normalisation par la taille ($b=1$).
- Pas de saturation du tf ($k=+\infty$).

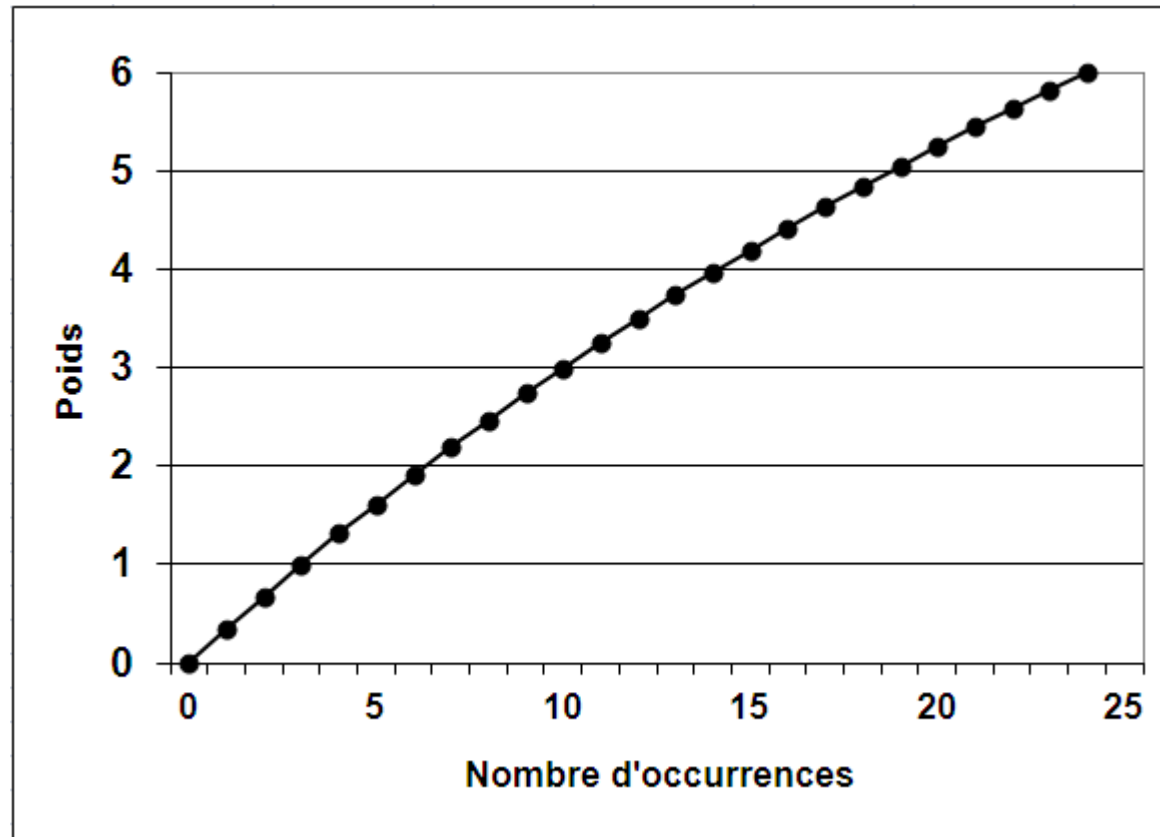
$$\begin{aligned} k_1 &= +\infty \\ b &= 1 \\ dl &= 3 * avdl \end{aligned}$$



BM25 & normalisation

- Normalisation par la taille ($b=1$).
- Très légère saturation du tf ($k=20$).

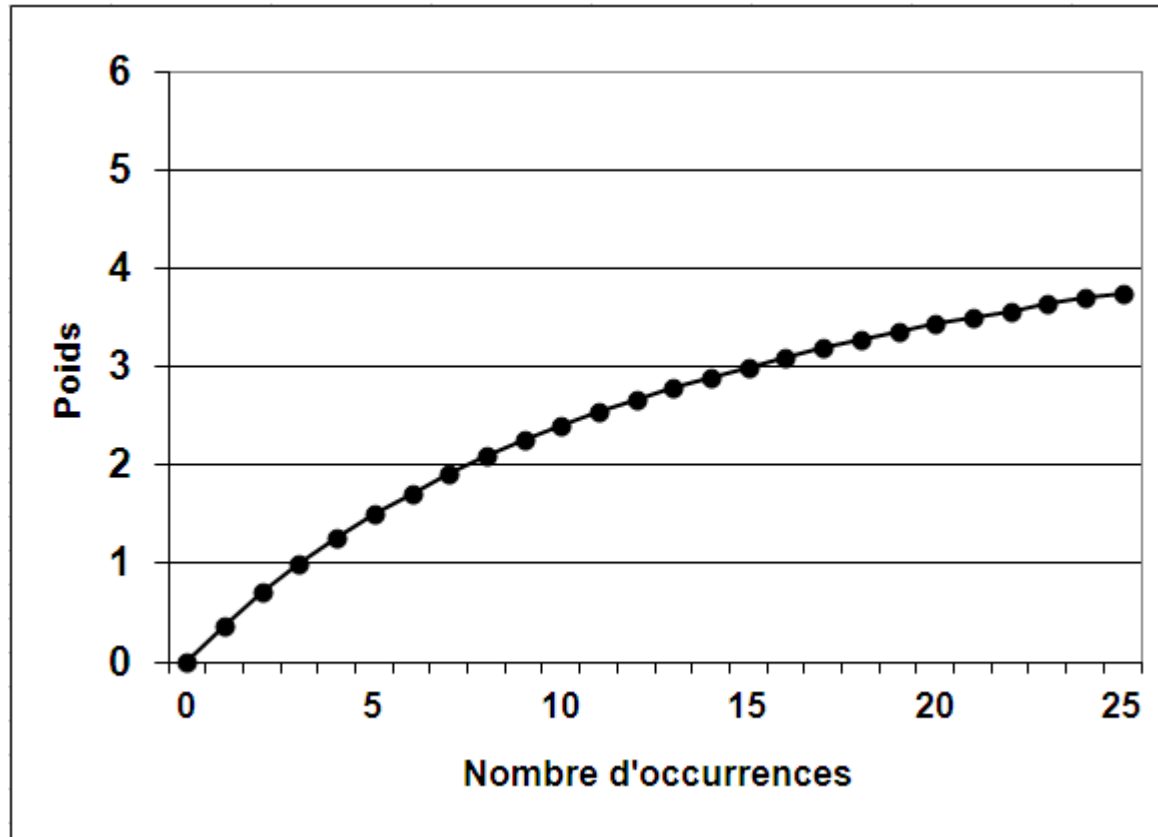
$k_1 = 20$
 $b = 1$
 $dl = 3 * avdl$



BM25 & normalisation

- Normalisation par la taille ($b=1$).
- Légère saturation du tf ($k=5$).

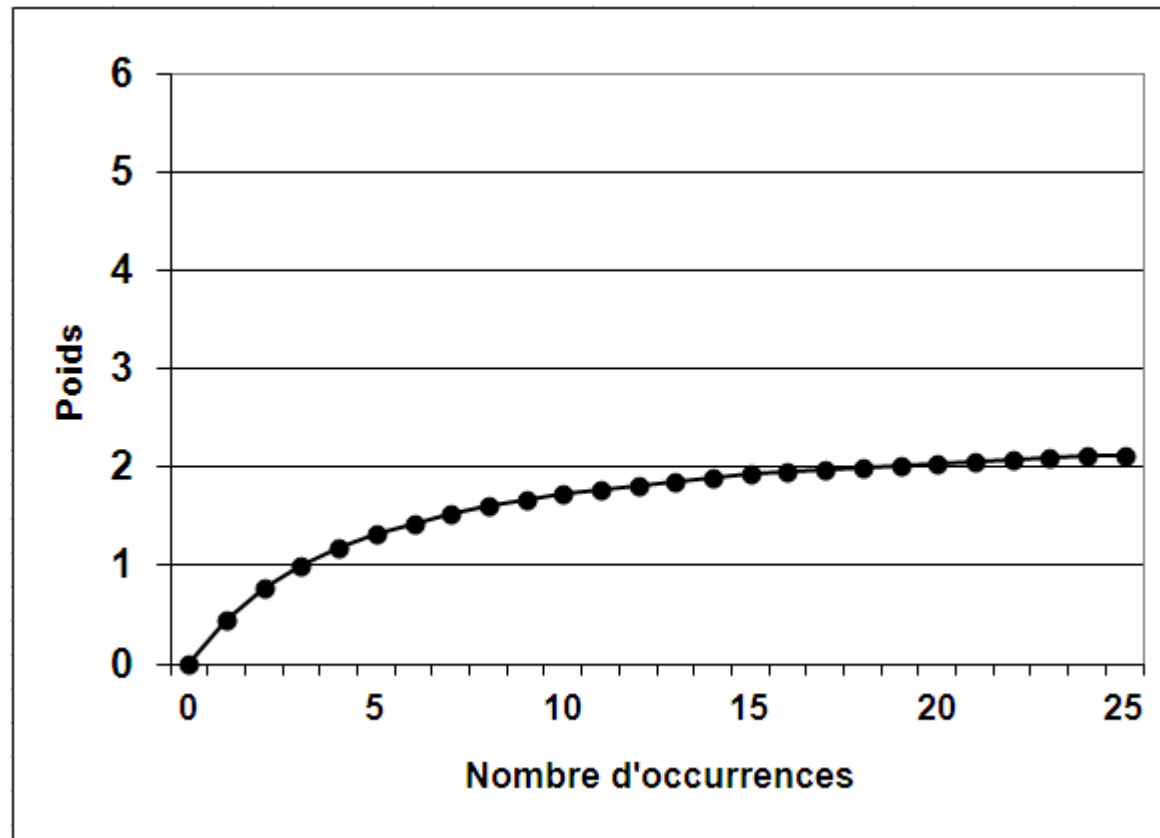
$$\begin{aligned} k_1 &= 5 \\ b &= 1 \\ dl &= 3 * avdl \end{aligned}$$



BM25 & normalisation

- Normalisation par la taille ($b=1$).
- Saturation forte du tf ($k=1,5$).

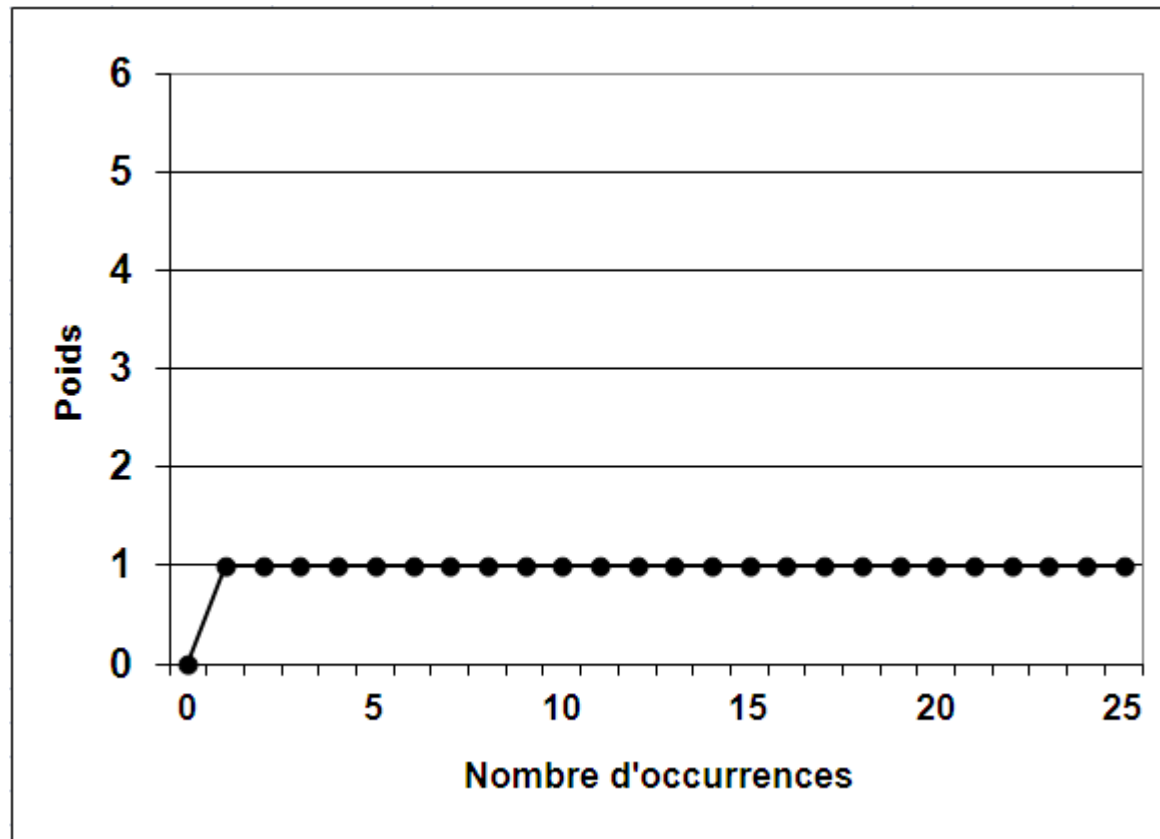
$$\begin{aligned} k_1 &= 1,5 \\ b &= 1 \\ dl &= 3 * avdl \end{aligned}$$



BM25 & normalisation

- Normalisation par la taille ($b=1$).
- Saturation extrême du tf ($k=0$).

$$\begin{aligned} k_1 &= 0 \\ b &= 1 \\ dl &= 3 * avdl \end{aligned}$$



TD

- TD : Q1 \rightarrow Q6

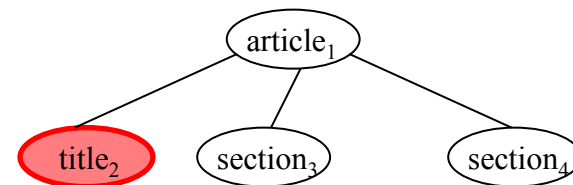
Pondération et ranking en RIS

- Quelques solutions :
 - Normalisation par la taille,
 - Types d'éléments,
 - Propagation :
 - des termes,
 - des poids des termes,
 - des scores de pertinence.

Pondération en RIS : Type des éléments

- Types d'éléments :
 - balises article, section, title, p, etc.
- Paramétrage des poids α_{title} , α_{section} , etc. :
 - Combinaison linéaire des poids [Wilkinson94].
 - ça revient à combiner les scores des éléments.
 - Poids du terme t_4 dans l'élément article_1 :

$$w_{t_4, \text{article}_1} = \alpha_{\text{title}} * w_{t_4, \text{title}_2} + \alpha_{\text{section}} * w_{t_4, \text{section}_3} + \alpha_{\text{section}} * w_{t_4, \text{section}_4}$$



- Calcul de w_{t_4, title_2} , etc. : fonction de pondération quelconque.
- Résultat : une liste **d'articles**.

Using the Structure of Documents to Improve the Discovery of Unexpected Information¹

Philippe Jacquot
University of Lyon, France
Philippe.Jacquot@univ-lyon1.fr

Christine Laguerre
University of Lyon, France
Christine.Laguerre@univ-lyon1.fr

ABSTRACT

In this paper we use document structure to improve the discovery of unexpected information. We use a tree structure to represent the structure of a document. We use a tree structure to represent the structure of a document. We use a tree structure to represent the structure of a document.

KEYWORDS

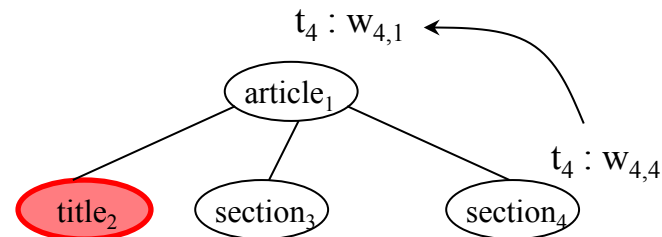
Information, Discovery, Tree, Mining, Unexpected Information, etc.

1. INTRODUCTION

In this paper we use document structure to improve the discovery of unexpected information. We use a tree structure to represent the structure of a document. We use a tree structure to represent the structure of a document. We use a tree structure to represent the structure of a document.

Pondération en RIS : Type des éléments

$$W_{t_4, \text{article}_1} = \alpha_{\text{title}} * W_{t_4, \text{title}_2} + \alpha_{\text{section}} * W_{t_4, \text{section}_3} + \alpha_{\text{section}} * W_{t_4, \text{section}_4}$$



- Remarque : c'est une propagation des poids des termes !
 - Ici : des feuilles vers la racine (agrégation)...
 - ... et pourquoi pas de la racine vers les feuilles (« contexte ») ?
 - Cf. [Wilkinson94].
- Remarque : si on utilise BM25 :
 - ➔ quel calcul de *df* ? Au niveau des articles, des éléments, des titres ?
 - ➔ quel calcul de *avdl* ?
- À suivre...

Pondération en RIS : Type des éléments

- Intégration des α dans BM25 : BM25F [Robertson04] :
 - « f » = fields (title, body, anchor),
 - Deux méthodes :
 - Approche « a posteriori » : combinaison des scores (cf. [Wilkinson94]).

$$w_{t,d} = \alpha_{\text{title}} * w_{t,\text{title}} + \alpha_{\text{sec}} * w_{t,\text{sec}_1} + \alpha_{\text{sec}} * w_{t,\text{sec}_2}$$

- Approche « a priori » ou « impact précoce » : combinaison des tf .

$$tf'_{t,d} = \alpha_{\text{title}} * tf_{t,\text{title}} + \alpha_{\text{sec}} * tf_{t,\text{sec}_1} + \alpha_{\text{sec}} * tf_{t,\text{sec}_2}$$

$$w_{t,d} = \frac{tf'_{t,d} * (k_1 + 1)}{k_1 * \left((1 - b) + b * \frac{dl}{avdl} \right) + tf'_{t,d}} * \log \left(\frac{N - df_t + 0,5}{df_t + 0,5} \right)$$

- Pourquoi ?
 - Préservation des propriétés de BM25 : Saturation du tf (avec k_1).
- Résultat : toujours une liste d'articles.

« a priori » vs « a posteriori »

- Approche « a posteriori » [Wilkinson94] :

$$w_{t,d} = \alpha_{\text{title}} * w_{t,\text{title}} + \alpha_{\text{sec}} * w_{t,\text{sec}}$$

<article>

<title>t₄</title>

<sec>t₄ t₄</sec>

</article>

- Approche « a priori » [Robertson04] :

$$tf'_{t,d} = \alpha_{\text{title}} * tf_{t,\text{title}} + \alpha_{\text{sec}} * tf_{t,\text{sec}}$$

$$w_{t,d} = \frac{tf'_{t,d} * (k_1 + 1)}{k_1 * \left((1-b) + b * \frac{dl}{avdl} \right) + tf'_{t,d}} * \log \left(\frac{N - df_t + 0,5}{df_t + 0,5} \right)$$

- Calcul ?
 - Avec : $\alpha_{\text{title}} = 2$, $\alpha_{\text{sec}} = 1$

- Approche « sans poids » : $w_{t_4,d}$?
- Approche « a posteriori » : $w_{t_4,d}$?
- Approche « a priori » : $w_{t_4,d}$?

#occs	Poids
0	0
1	1
2	1,4
3	1,8
4	1,9

(Pour simplifier)
(BM25)

k₁ = 1,2
b = 0,75
dl = avdl

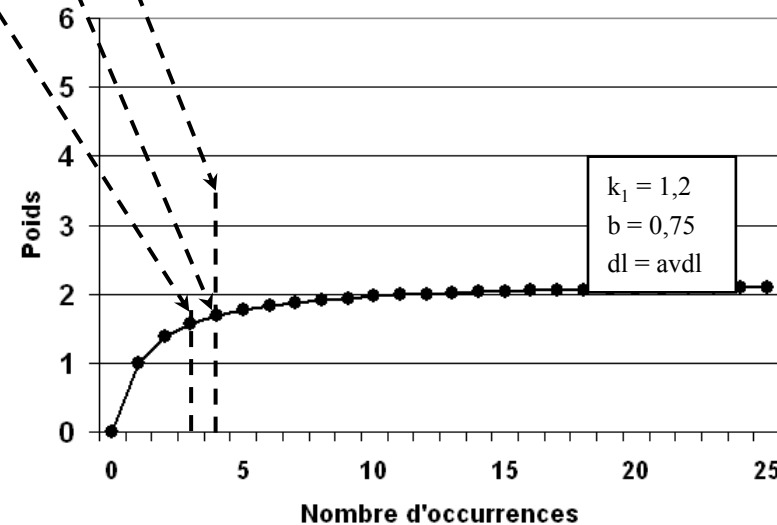
« a priori » vs « a posteriori »

- Sans α : $tf' = 1+2 \rightarrow w = 1,8$
- « a posteriori » : $w = 2*1+1*1,4 = 3,4$
- « a priori » : $tf' = 2*1+1*2 \rightarrow w = 1,9$

<article>
 <title>t₄</title>
 <sec>t₄ t₄</sec>
 </article>

#occs	Poids
0	0
1	1
2	1,4
3	1,8
4	1,9

- « a posteriori » [Wilkinson94] :
 - Un doc contenant un terme t_j dans plusieurs champs aura un score trop élevé par rapport à un doc contenant plusieurs fois t_j dans un seul champ.
 - k_1 ne joue plus son rôle (contrôle de la saturation du tf).
- « a priori » [Robertson04] :
 - Préservation des propriétés de BM25 (tf non linéaire)



TD

- TD :
 - Lire [Robertson2004].
 - Exos sur l'article : Q1->Q20.
 - Exos sur les pondérations : Q7 et Q8.
 - Exos sur l'article : Q21->Q26.

Propagation des poids : optimisation des paramètres

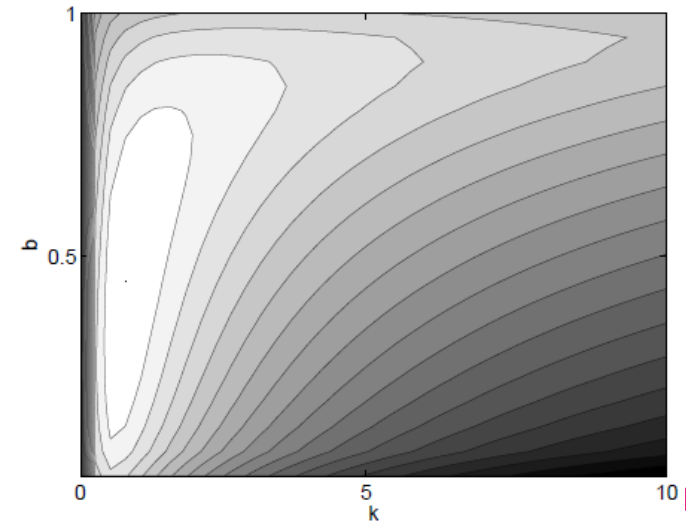
- Comment fixer / optimiser :
 - Poids des champs : α_{title} , α_{section} , etc.
 - ... et tous les autres paramètres du calcul de $w_{t,d}$, ex. : b_{title} , b_{section}
- « a posteriori » avec BM25 comme fonction de pondération :
 - K « champs » (ici, $K = 2$: title, sec).
 - Il faut optimiser $2 * K + (K-1) = 3 * K - 1$ paramètres :
 - On optimise b et k_1 de BM25 pour chaque champ (K optimisations 2D).
 - On fixe α_{title} , on optimise α_{section} (1 optimisation K-1D).
- Ex. Collection INEX : Quels types d'éléments XML ?
 - DTD non connue a priori → liste des balises non contrôlée,
 - Sémantique des balises inconnue,
 - → Grand nombre de types de balises :
 - en 2008 : > 1'000,
 - en 2010 : > 30'000.
- Fixer α_{section} , b_{title} , b_{section} , etc. par optimisation : très coûteux.

BM25F : optimisation des paramètres

- Optimiser : « seulement » $2 + (K-1)$ paramètres ([Robertson04]) :
 - Optimiser b, k_1 avec $\alpha_i = 1$ (1 optimisation 2D).
 - Optimiser α_i pour chaque champ (1 optimisation K-1D).
 - Avec :
$$newb = b \quad newk_1 = k_1 * \frac{atf_{weighted}}{atf_{unweighted}}$$
- Variante BM25F : optimiser b_i pour chaque champ [Zaragoza04] :
 - Optimiser k_1 et b_i pour chaque champ (K optimisations 2D).
 - Avec ces b_i et $\alpha_i = 1$: optimiser k_1 (1 optimisation 1D).
 - Avec ces b_i et k_1 : optimiser les α_i (1 optimisations (K-1)D).
- Fixer $\alpha_{section}, b_{title}, b_{section}$, etc. par optimisation : très coûteux.

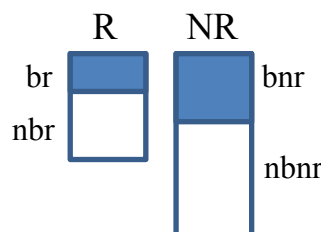
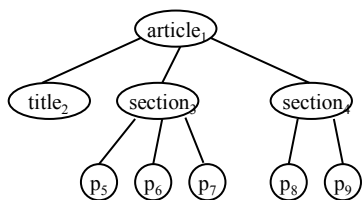
Optimisation

- Paramètres : b , k , α , ... pour chaque champ ?
- Empiriquement [Rapela01].
- Algorithmes génétiques [Trotman05].
- Recuit simulé [Boyan96].
- Optimisation par méthodes de descente de gradient [Taylor06].
 - BM25 : b , k :
 - \rightarrow 2 paramètres.
 - BM25F : k fixé, b et α pour 4 champs :
 - \rightarrow 9 paramètres.
 - BM25F + plein d'autres paramètres :
 - \rightarrow 375 paramètres.



Pondération en RIS : Type des éléments

- Apprentissage du poids α des types d'éléments.
- Modèle probabiliste BM25T [Géry2009] :
 - Calcul basé sur la probabilité de trouver un terme t pertinent (R) dans une balise b .
 - Pour une balise (un type d'élément) b :



$$\alpha_{b,t} = \frac{\frac{br}{nbr}}{\frac{bnr}{nbnr}}$$

Calcul du poids des types d'éléments : modèle probabiliste

- Idée : modèle de RI probabiliste [Robertson76]
- Estimation du score d'un doc. d_j :

$$odd(d_j) = \frac{P(d_j / R)}{P(d_j / NR)}$$

- Avec :

$$\hat{odd}(d_j) = \sum_{t_i \in d_j \cap Q} \log \left(\frac{\frac{r_i}{R - r_i}}{\frac{n_i - r_i}{N - R - n_i + r_i}} \right)$$

	Pertinent	Non pertinent	
#docs avec t_i	r_i	$n_i - r_i$	N_i
#docs sans t_i	$R - r_i$	$N - R - n_i + r_i$	$N - n_i$
	R	$N - R$	N

- Évaluer la capacité d'une balise b à distinguer les termes pertinents.
- Apprentissage d'un poids α_b
- Nombre de termes pert./non pert., marqués/non marqués par b :

$$\alpha_b = \sum_{t_i} \left(\frac{\frac{r_i}{R - r_i}}{\frac{n_i - r_i}{N - R - n_i + r_i}} \right)$$

	Pert.	Non pert.	
#termes t_i marqués par b	r_i	$n_i - r_i$	n_i
#termes t_i non marqués par b	$R - r_i$	$N - R - n_i + r_i$	$N - n_i$
	R	$N - R$	N

Impact précoce du poids des types d'éléments (BM25T)

- Rappel : BM25F préserve la non-linéarité de BM25.

$$tf'_{t,d} = \alpha_{\text{title}} * tf_{t,\text{title}} + \alpha_{\text{sec}} * tf_{t,\text{sec}}$$

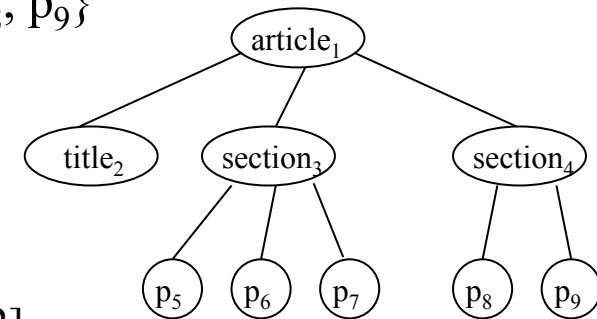
$$w_{t,d} = \frac{tf'_{t,d} * (k_1 + 1)}{k_1 * \left((1-b) + b * \frac{dl}{avdl} \right) + tf'_{t,d}} * \log \left(\frac{N - df_t + 0,5}{df_t + 0,5} \right)$$

- BM25T : même principe :
 - ⇒ Impact « précoce » du poids des types d'éléments.
 - α = moyenne des poids α_b des types d'éléments *englobant* le terme t.
- Dans BM25 : tf est multiplié par α :

$$w_{t,d} = \frac{\alpha * tf_{t,d} * (k_1 + 1)}{k_1 * \left((1-b) + b * \frac{dl}{avdl} \right) + \alpha * tf_{t,d}} * \log \left(\frac{N - df_t + 0,5}{df_t + 0,5} \right)$$

Pondération en RIS au niveau des éléments

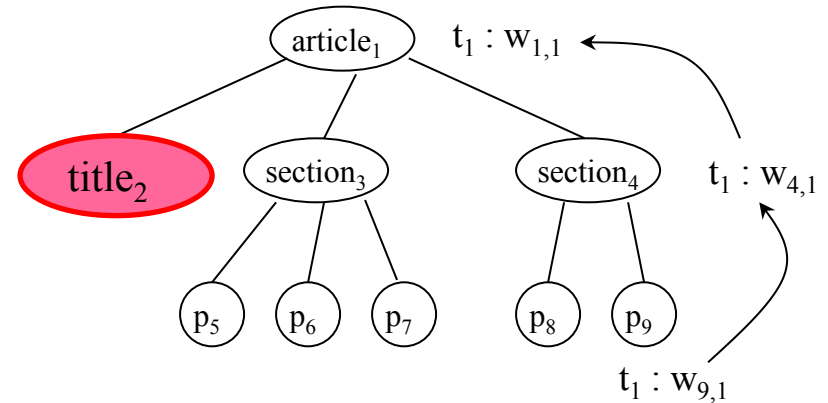
- Adapter *tf.idf* au niveau de granularité des éléments :
 - $E = \{\text{article}_1, \text{title}_2, \text{section}_3, \text{section}_4, p_5, p_6, p_7, p_8, p_9\}$
 - Adapter le *tf* :
 - *etf*, Element Term Frequency [Sigurbjornsson03].
 - Adapter l'*idf* :
 - *ief*, Inverse Element Frequency [Wolff00] [Grabs02].
 - *itf* pour chaque « tag » (type d'élément) [Theobald2006].
 - *itdf* : pour chaque tag dans chaque document [Zargayouna04].
 - Problème : éléments imbriqués (nested elements) :
 - ➔ index redondant !
 - Terme apparait au niveau n ➔ indexé n fois ! ➔ n contributions au *tf* / *idf* !
 - Résultat : une liste d'éléments !



Pondération et ranking en RIS

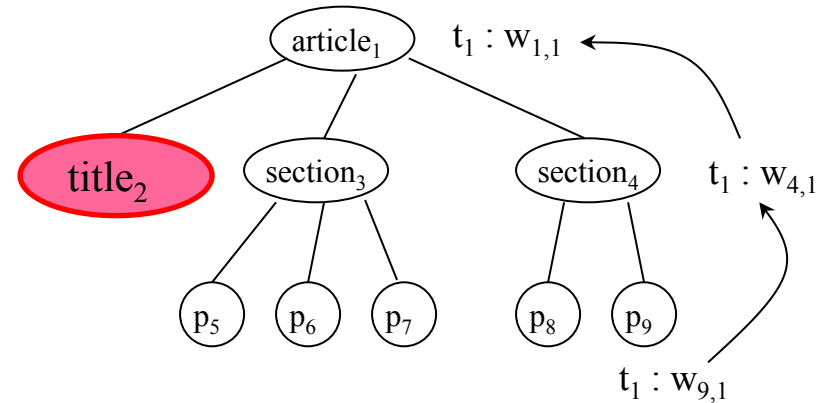
- Quelques solutions :
 - Normalisation par la taille,
 - Types d'éléments,
 - Propagation :
 - des termes,
 - des poids des termes,
 - des scores de pertinence.

Pondération en RIS : propagation...



- Propagation des poids des termes :
 - Agrégation (bas \rightarrow haut), déjà vu : [Wilkinson94].
- Propagation des scores de pertinence :
 - Indexation des feuilles, puis propagation scores (bas \rightarrow haut) [Geva05] :
 - \rightarrow index allégé mais calcul pertinence plus coûteux.
 - Imbrication des éléments : *ief* calculé au niveau des feuilles.

Pondération en RIS : propagation...

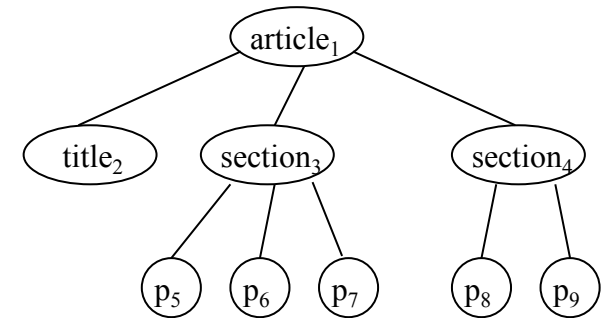


- Propagation des scores de pertinence [Sauvagnat05] :

$$score(e, q) = \rho * m * \sum_{e_l} \alpha^{d(e, e_l)^{-1}} * \beta(e_l) * score(e_l, q) + (1 - \rho) * score(root, q)$$

- Agrégation (bas \rightarrow haut),
- $m = |\{e_l \text{ descendants de } e / score(e_l) > 0\}|$
- ρ : influence du contexte (haut \rightarrow bas).
- α : atténuation en fonction de la distance.
- $\beta(e_l)$: renforcer le rôle des petits fragments.

Pondération en RIS au niveau des éléments



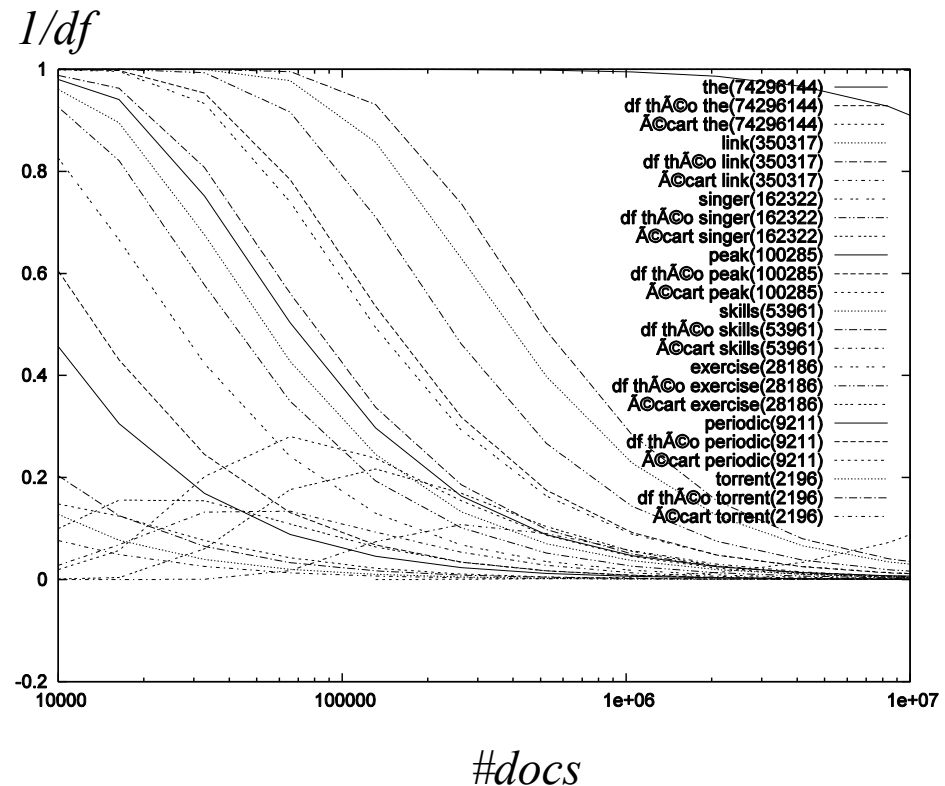
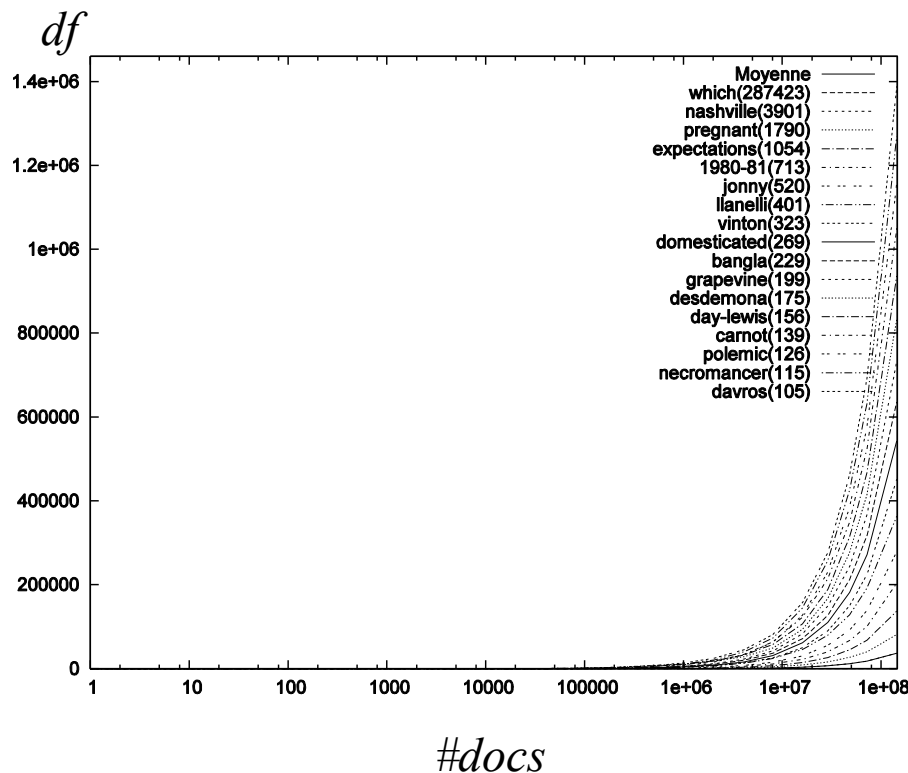
- Adapter BM25F au niveau des éléments.
- BM25E [Lu05] :
 - Un champ = élément + éléments qui contribuent (fils et parfois ancêtres).
 - Hypothèses :
 - Propagation de termes haut → bas : *title*, *abstract* et *section_title*.
 - Propagation de termes bas → haut : tous.
 - Optimiser b , k et calculer *avdl* au niveau des éléments (*avel*) :
 - Trop coûteux !
 - ➔ calculés au niveau des articles ➔ identiques pour tous les éléments.
 - Résultat : une liste d'éléments !
 - Mais... seulement 5 types d'éléments.
 - α optimisé pour 3 champs (*title*, *abstract*, *section_title*).
 - Fixer α_{title} , α_{section} , etc. par optimisation : très coûteux.

Calcul du *df* en RIS

- Distribution des termes dans la collection INEX :
 - Collection de documents issus de Wikipédia (2009).
 - 50 Go de XML.
 - 2,6 millions articles Wikipedia, 400 millions éléments XML.
- Découpage artificiel de la collection en 2ⁿ « documents » :
 - 2 « documents » de 730957517 mots.
 - 4 « documents » de 365478759 mots.
 - 8 « documents » de 182739380 mots.
 - 16 « documents » de 91369690 mots.
 - (...)
 - 1024 « documents » de 1427652 mots.
 - (...)
 - 243652506 « documents » de 6 mots.
 - 487305012 « documents » de 3 mots.
 - 730957517 « documents » de 2 mots.

Calcul du df en RIS

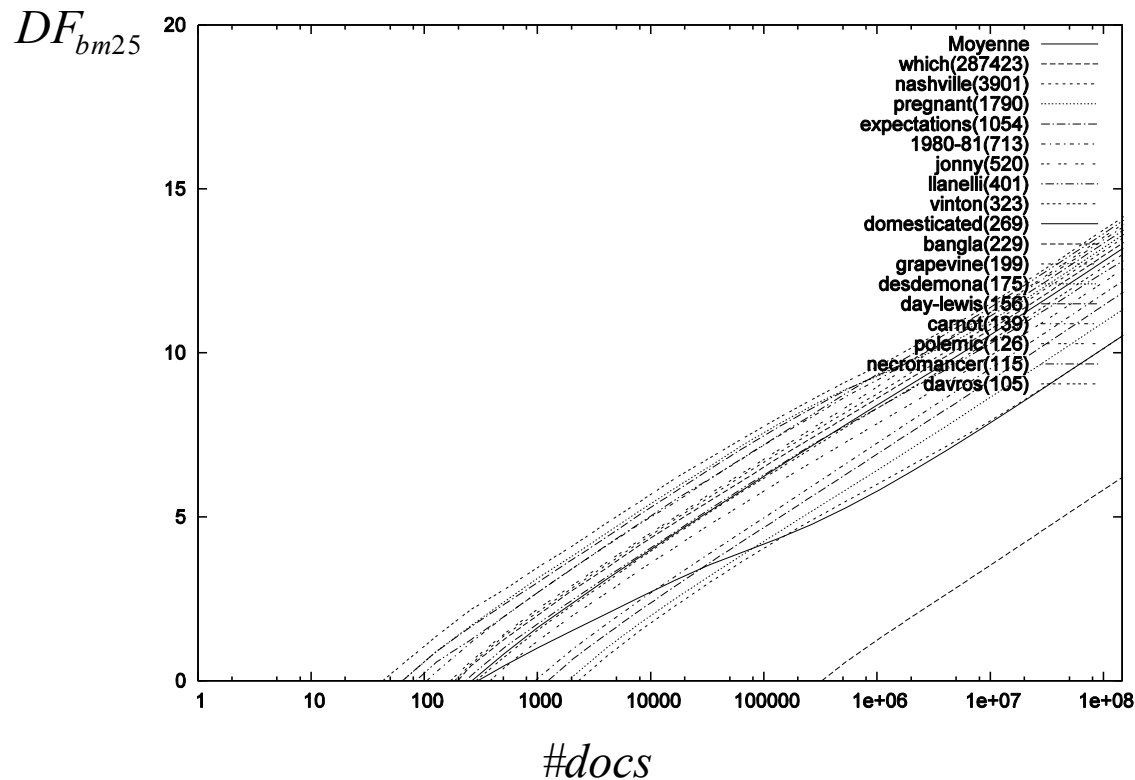
- Évolution du df de quelques termes :



Remarque : $\lim_{taille(doc) \rightarrow \infty} (df) = \sum_{docs} tf$

Calcul du df en RIS

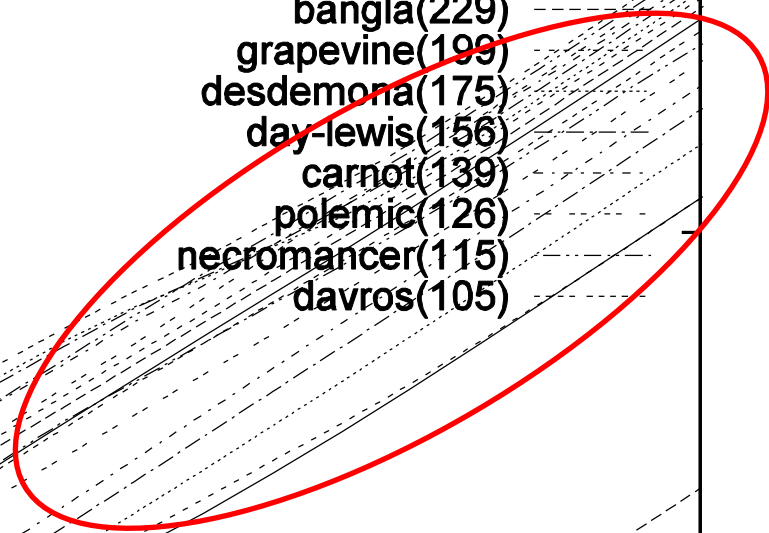
- Évolution du df de quelques termes : $DF_{bm25} = \log \left(\frac{N - df_t + 0,5}{df_t + 0,5} \right)$



La granularité ne change rien au df ?

Calcul du *df* en RIS

Moyenne	————
which(287423)	-----
nashville(3901)	-----
pregnant(1790)	-----
expectations(1054)	-----
1980-81(713)	-----
jonny(520)	-----
llanelli(401)	-----
vinton(323)	-----
domesticated(269)	-----
bangla(229)	-----
grapevine(199)	-----
desdemona(175)	-----
day-lewis(156)	-----
carnot(139)	-----
polemic(126)	-----
necromancer(115)	-----
davros(105)	-----



Recherche d'Information (RI)

- Plan du cours « RI Structurée » (RIS) :
 - 5. RI et structure intra-document (ex. XML).
 - Introduction : RI classique.
 - Web (HTML, XML) : Document Structuré ou Hypertexte ?
 - RI Structurée :
 - Langages de requêtes structurés.
 - Indexation en RIS.
 - Passage Retrieval, modèle à proximité.
 - Pondération en RIS : normalisation, types d'éléments, propagation.
 - Évaluation en RI : la compétition INEX.
 - 6. RI et structure inter-documents (ex. les liens hypertextes).
 - 7. Projet : participation à une mini-compétition de RIS.

Evaluation de la RIS : la compétition INEX

- Évaluation en RI : cf. cours M. Beigbeder
- Initiative for the Evaluation of XML Retrieval.
- 1^{ère} compétition en RIS (RI XML), 2002.
- Collection de test :
 - Collection de documents.
 - Tâches de RI.
 - Jeu de requêtes.
 - Jugements de pertinence.
 - Mesures d'évaluation.
- Adaptés à la RIS !



Compétition INEX : collection de documents

- Plusieurs collections depuis 2002.
- Collection de documents issus de Wikipédia (2006) :
 - 659'388 articles Wikipedia, 52 millions éléments XML,
 - 1'244 balises différentes,
 - 4.5 Go dont 1,6 Go de texte,
 - 114 requêtes (2008) avec jugements de pertinence.
- Collection de documents issus de Wikipédia (2009) :
 - 2,6 millions articles Wikipedia, 400 millions éléments XML,
 - >30'000 balises différentes,
 - 50 Go de XML,
 - 68 + 52 requêtes (2009+2010) avec jugements de pertinence.

Compétition INEX : jeu de requêtes

- Exemple de requête (« topic » n°14, INEX 2010) :

```
<topic id="2010014" ct_no="329">  
  <title>composer museum</title>  
  <description>Documents or parts of documents that describe or identify a  
    museum dedicated or which has a significant section dedicated to a  
    composer</description>  
  <narrative>I want to know the museum that are dedicated to or which have  
    a significant section dedicated to a composer. Other music related museum  
    which are not dedicated to a composer are not relevant (for instance a  
    museum of musical instruments)</narrative>  
</topic>
```

- Seule l'utilisation du champ « *title* » est autorisée.

Compétition INEX : tâches de RI

- Exemple : « track » (piste) RI « ad-hoc » :
 - RI XML, avec ou sans contraintes de structure.
 - Résultat = liste classée d'éléments ou de passages pertinents, à la fois spécifiques et exhaustifs.
- Exemples de « sous-tâches » de RI Ad-Hoc :
 - Relevant in Context (RIC) :
 - Liste d'éléments (ou de passages) groupés par article.
 - Au sein d'un même article : pas forcément dans l'ordre d'origine.
 - Pas de recouvrement.

Cf. spécification INEX 2010 : <http://www.inex.otago.ac.nz/tracks/adhoc/runsubmission.asp?action=specification>

Compétition INEX : tâches de RI (suite)

- Exemples de « sous-tâches » de RI Ad-Hoc :
 - Relevant in Context (RIC).
 - Restricted RIC : idem, maxi 500 caractères / articles (smartphones).
 - Restricted Focused :
 - Liste d'éléments (ou de passages), non groupés.
 - Pas de recouvrement.
 - Maxi 1000 caractères / requête.
 - Efficiency : 15, 150 ou 1500 éléments par requête, en précisant : temps de calcul, coût d'E/S, caractéristiques des machines, etc.

Compétition INEX : tâches de RI

- Mais aussi d'autres « pistes » (tracks) :
 - Book,
 - Data-centric.
 - Interactive.
 - Link-the-wiki.
 - Q/A.
 - Relevance Feedback.
 - Web Service Discovery.
 - XML Mining.

INEX : tâche RIC

- Consulter :
 - <http://www.inex.otago.ac.nz/tracks/adhoc/runsubmission.asp?action=specification>

1) Relevant in Context Task

Motivation for the Task

The scenario underlying the Relevant in Context Task is to return the relevant information (captured by a set of elements or passages) within the context of the full article. The task makes a number of assumptions: Display results will be grouped per article, ~~in their original document order~~, providing access through further navigational means; Users consider the article as the most natural unit, and prefer an overview of relevance in their context.

Evaluation

Task is as before, but viewed as a form of snippet retrieval. We use as main measure the proposal from the Univ. Tampere that takes the reading length into account -- probably the T2I(300) measure from <http://dx.doi.org/10.1007/s10791-010-9133-9> which strongly penalizes the retrieval of non-relevant text. The new measure takes the suggested reading order within each article into account, but almost all earlier RiC submissions contained this ranking.

Results to Return

The aim of the Relevant in Context Task is to first identify relevant articles (the fetching phase), and then to identify the relevant results within the fetched articles (the browsing phase). The /article[1] element itself need not be returned, but is implied by any result from a given article. Since the content of an element is fully contained in its parent element and ascendants, the set may not contain overlapping elements. Also passage results may not be overlapping.

Summarizing: The Relevant in Context Task requires a ranked list of articles, and for each article a ranked list results covering the relevant material in the article. Overlap is not permitted.

Compétition INEX : « runs »

- « Run » : liste de résultats retournée par un SRI :
 - Par défaut : maxi 1500 éléments XML par requête.
 - Grouper par n° de requête, et pour chaque requête : trier par ordre décroissant de score.
 - Selon les tâches, restrictions de taille, d'ordre des éléments, de recouvrement, etc.
- Exemple (avec 7 requêtes) : 1 run = 1 fichier = $7 * 1500$ lignes.
- Chaque participant peut soumettre un nombre limité de runs à chaque tâche.

Format des « runs » INEX

2010001 Q0 364275 12 0.9765 Emse514 /article[1]/bdy[1]/sec[6]/p[1]

- Colonnes séparées par des espaces :
 - Colonne 1 : requête n°2010001.
 - Colonne 2 : Q0 (inutilisé)
 - Colonne 3 : article n° 364275 de la collection INEX.
 - Colonne 4 : cet élément est classé au 12^{ème} rang dans ce run.
 - Colonne 5 : le score de cet élément est 0.9765.
 - Colonne 6 : nom de l'équipe.
 - Colonne 7 : le chemin de l'élément dans l'article 364275.


Remarque : ce chemin peut être : /article[1]

Compétition INEX : jugements de pertinence

- Collection de test construite collaborativement :
 - 1) Participants : plusieurs dizaines de labos.
 - 2) Requêtes proposées par les participants.
 - 3) Organisateurs sélectionnent les requêtes.
 - 4) Participants produisent 1 ou plusieurs « runs ».
 - 5) Organisateurs compilent les runs.
 - 6) Jugements de pertinence :
 - 2 ou 3 requêtes à « évaluer » par participant.
 - Consulter le petit millier d'articles retournés par au moins un système.
 - Signaler les passages pertinents.
 - 7) Organisateurs compilent les jugements de pertinence
 - 8) Classements des différents SRI.

Compétition INEX : jugements de pertinence

GPX




File: 1707709 Show Pool Hide Keywor... Set Keywords Keywords: Help Topic % assessed 7 %

Handel House Museum

1707709 237680213 2008-09-11T10:03:06Z Pageturners 7302682

Houses in Westminster
Music **museums** in the United Kingdom
Biographical **museums** in London
Grade I listed buildings in London
Grade I listed houses
Buildings with blue plaques
Buildings and structures in Westminster



Handel House. Note the blue plaque

The Handel House **Museum** at 25 Brook Street , in the exclusive central London district of Mayfair was the home of the German born baroque **composer** George Frideric Handel from 1723 until his death at the house in 1759. He **composed** works such as The Messiah , Zadok the Priest and the Fireworks Music there.

The **museum**

The house has been restored to look as it did during Handel's occupancy. A typical early 18th century London terrace house, it comprises a basement, three main storeys and an attic, and Handel was the first occupant. The attic was later converted into a fourth full floor. The ground floor is now a music and gift shop and the upper floors are leased to a charity called the Handel House Trust, and have been open to the public since 8 November 2001. The interiors have been restored to the somewhat spartan style of Georgian era, using mostly architectural elements from elsewhere, as other than the staircase, few of the original interior features survived. The Handel House Collection Trust has assembled a collection of Handel memorabilia, including the Byrne Collection of several hundred items, which was acquired in 1998.

OOPS!

Next Document

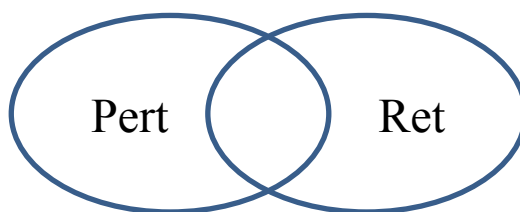
Completely Irrelevant

EXIT

☐ DEBU...

Compétition INEX : mesures d'évaluation

- R/P : Rappel / Précision classique :
 - Pertinence binaire : document pertinent / non pertinent.



$$P = \frac{|Pert \cap Ret|}{|Ret|} \quad R = \frac{|Pert \cap Ret|}{|Pert|}$$

- R/P adapté à la RIS : MAiP, MAgP, etc. [Kamps07]
 - Ici : Generalized P/R (c'est-à-dire : MAgP).
 - Une proportion (nombre de caractères) de chaque article est pertinente.
 - Pour chaque article a retrouvé :

$$P(a) = \frac{|\text{rel}(a) \cap \text{ret}(a)|}{|\text{ret}(a)|} \quad R(a) = \frac{|\text{rel}(a) \cap \text{ret}(a)|}{|\text{rel}(a)|}$$

NB : calculé en considérant tous les éléments de a !

Compétition INEX : mesures d'évaluation

- R/P adapté à la RIS : Generalized P/R [Kamps07] :
 - Pour chaque article a retrouvé : F-score $F(a) \in [0..1]$:

$$F(a) = \frac{2 \cdot P(a) \cdot R(a)}{P(a) + R(a)}$$

- Augmenter l'importance de la précision :
 - $\beta = 1/4 \rightarrow$ précision 4 fois plus importante que le rappel.

$$F_{\beta} = \frac{(1 + \beta^2) \cdot Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall}$$

- Generalized Precision à un rang r :

$$gP(r_{art}) = \frac{\sum_{a=1..r_{art}} F(a)}{r_{art}} \text{ and } gR(r_{art}) = \frac{\sum_{a=1..r_{art}} relart(a)}{Numrel}$$

Compétition INEX : mesures d'évaluation

- R/P adapté à la RIS : Generalized P/R [Kamps07] :
 - AgP : Average Generalized Precision :
 - Moyenne des gP à chaque rang où gR augmente.
 - MAgP : Mean Average Generalized Precision :
 - Moyenne des AgP sur toutes les requêtes.

INEX 2010 : Reading Effort

- Intégration dans gP & co d'un « effort de lecture ».
 - Idée : considérer la tâche comme une forme de « Snippet Retrieval ».
 - ➔ on renforce un peu plus l'importance de la précision.
- Mesure T2I(300) :
 - "Tolerance to Irrelevance" (T2I) score par document.
 - Scénario :
 - L'utilisateur lit les éléments/passages dans le sens de lecture de l'article.
 - Il stoppe sa lecture quand son seuil de tolérance à la non-pertinence est atteint. Par défaut : 300 caractères non pertinents.

Résultat d'un run

- Évaluation globale du run :

num_q	all	52
num_ret	all	72247
num_rel	all	5471
num_rel_ret	all	4608
ret_size	all	618141678
rel_size	all	17641119
rel_ret_size	all	8582907
MAgP	all	0.18345875829022795

ircl_prn.0,00	all	0.5522557322086471
ircl_prn.0,10	all	0.37371163625571857
ircl_prn.0,20	all	0.3080953360973854
ircl_prn.0,30	all	0.2664436343136546
ircl_prn.0,40	all	0.22242334341508974
ircl_prn.0,50	all	0.17615068314583998
ircl_prn.0,60	all	0.14406704300104026
ircl_prn.0,70	all	0.11079644146935459
ircl_prn.0,80	all	0.07262041942612983
ircl_prn.0,90	all	0.036439539380988734
ircl_prn.1,00	all	0.019798090625945834

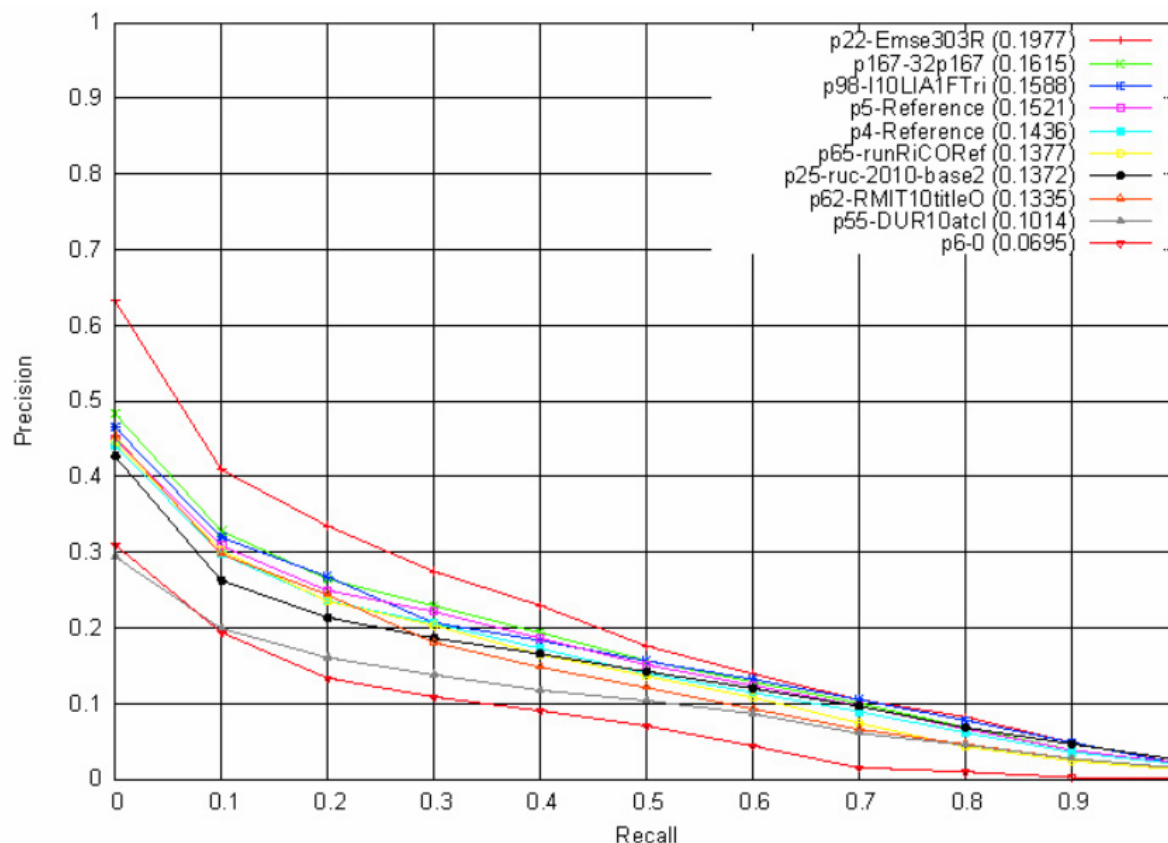
gP[1]	all	0.3802866730047784
gR[1]	all	0.01969663412898154
gP[2]	all	0.41576805344031104
gR[2]	all	0.04169196314697102
gP[3]	all	0.37286232822415644
gR[3]	all	0.05584582917846841
gP[5]	all	0.33684284356258554
gR[5]	all	0.08467317541705312
gP[10]	all	0.3024890560910691
gR[10]	all	0.14339326578042996
gP[25]	all	0.22906001995264594
gR[25]	all	0.25350799302679905
gP[50]	all	0.189295556126249
gR[50]	all	0.34753518690968066

Résultat d'un run

- Évaluation détaillée pour chaque requête :

num_ret	2010003 1498
num_rel	2010003 231
num_rel_ret	2010003 198
ret_size	2010003 13741521
rel_size	2010003 1252105
rel_ret_size	2010003 1019455
gP[1]	2010003 0.0
gR[1]	2010003 0.004329004329004329
gP[2]	2010003 0.4056603773584906
gR[2]	2010003 0.008658008658008658
gP[3]	2010003 0.27044025157232704
gR[3]	2010003 0.012987012987012988
gP[5]	2010003 0.16226415094339625
gR[5]	2010003 0.021645021645021644
gP[10]	2010003 0.1742843238168248
gR[10]	2010003 0.03463203463203463
gP[25]	2010003 0.19619351831310972
gR[25]	2010003 0.07792207792207792
gP[50]	2010003 0.20242714352114968
gR[50]	2010003 0.14285714285714285
AgP	2010003 0.12295704616335873

Compétition INEX : courbes de R/P



#	MAgP	Institute	Run
1	0.1977	ENSM-SE	Emse303R
2	0.1967	ENSM-SE	Emse301R
3	0.1615	Peking University	32p167
4	0.1615	Peking University	36p167
5	0.1598	Peking University	31p167
6	0.1598	Peking University	37p167
7	0.1588	LIA - University of Avignon	I10LIA1FTri

Recherche d'Information (RI)

- Plan du cours « RI Structurée » (RIS) :
 - 5. RI et structure intra-document (ex. XML).
 - Introduction : RI classique.
 - Web (HTML, XML) : Document Structuré ou Hypertexte ?
 - RI Structurée :
 - Langages de requêtes structurés.
 - Indexation en RIS.
 - Passage Retrieval, modèle à proximité.
 - Pondération en RIS : normalisation, types d'éléments, propagation.
 - Évaluation en RI : la compétition INEX.
 - 6. RI et structure inter-documents (ex. les liens hypertextes).
 - 7. Projet : participation à une mini-compétition de RIS.