Mesurer la qualité des systèmes de catégories de blogs

Ivan Garrido-Marquez Jorge Garcia Flores François Lévy
Adeline Nazarenko

LIPN, CNRS & Université Paris 13 – Sorbonne Paris Cité

July 6, 2017

Plan

Systèmes d'annotation des blogs

Les billets de blogs sont annotés par leurs auteurs (généralement).
On trouve 2 grands types d'annotations dans les blogs :

- les **catégories**
- les mot-clefs (tags)

Nous nous intéressons ici

- non à la qualité des annotations en tant que telles (hypothèse : les auteurs choisissent les tags/catégories à bon escient par rapport au contenu des billets)
- mais à la **qualité du système d'indexation** que forment les annotations

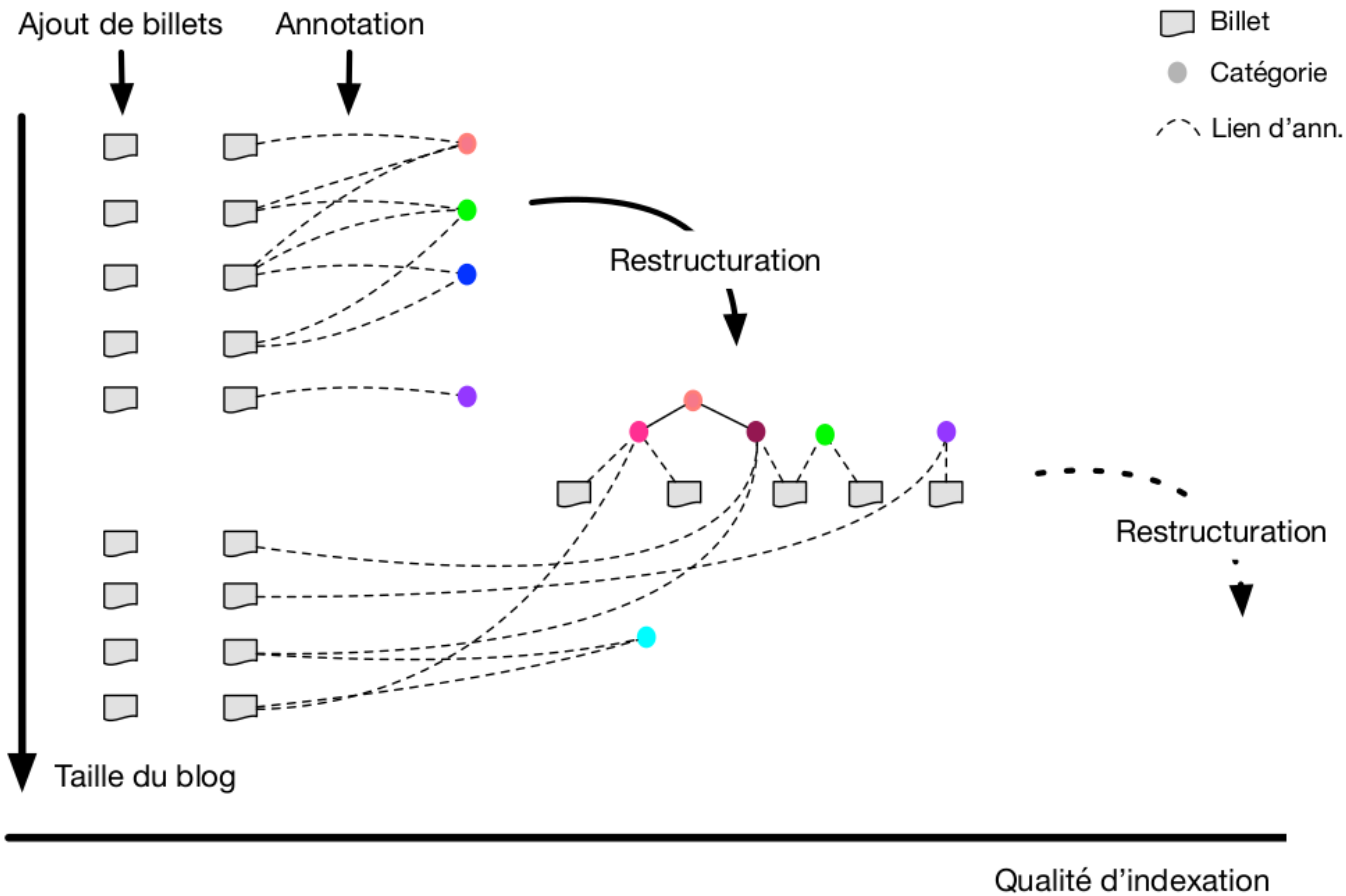
Systèmes d'annotation des blogs

On trouve différents types de systèmes des catégories

- mono-catégoriel : un billet n'est annoté que par une seule catégorie
- multi-catégoriel : les billets peuvent être associés à plusieurs catégories
- hiérarchique : les catégories sont organisées en arbre

Avec le temps, le nombre de billets augmente et les thèmes évoluent, si bien que le système de catégories proposé à un instant donné peut être beaucoup moins adéquat quelques mois ou quelques années plus tard.

Processus d'annotation des blogs



Nous proposons des métriques pour mesurer la qualité d'un système de catégories de blogs pour déterminer quand et comment un système de catégorie doit être restructuré.

Différentes perspectives sur la qualité des annotations et systèmes d'annotation

- Adéquation entre l'annotation et la sémantique du texte annoté
  - annotateur unique : cohérence [6] [1]
  - plusieurs annotateurs : mesures d'accord [3] [5]
  - annotation automatique : précision et rappel
- Capacité discriminante du vocabulaire d'indexation [9]
- Complexité d'une campagne d'annotation [2]
- **Qualité d'accès à l'information**
  - nous proposons deux mesures
    - équilibre
    - coût d'accès

Mesure d'équilibre

L'*entropie* d'un système de catégories caractérise l'information intrinsèquement contenue dans ce système, vue comme une distribution de probabilité sur les catégories.

Quantité d'information (pour un événement e)

$$I(e) = -\log_b(P(e))$$

Entropie : valeur espérée de la quantité d'information

$$H = E[I(e)] = -\sum_{e \in A} P(e) \log_b(P(e)) \quad (1)$$

où A est l'ensemble des événements élémentaires de la distribution (discrète).

Nous nous appuyons sur deux notions pour l'équilibre:

- l'entropie informationnelle classique [8]
- la diversité biologique des espèces dans une population [7]

Mesure d'équilibre (suite)

Equilibre : mesure normalisée de l'entropie du blog x en prenant comme valeur de référence l'entropie maximum qu'on peut obtenir avec le même nombre n de catégories ($\log(n)$)

$$Equilibre(x) = \frac{H(x)}{\max(H(x))} = \frac{H(x)}{\log(n)} \tag{2}$$

- Équilibre des blogs mono-catégoriels

$$H(x) = -\sum_{i=1}^{n} \frac{|x_i|}{N} \log(\frac{|x_i|}{N}) \tag{3}$$

- Équilibre des blogs multi-catégoriels : similaire mais chaque combinaison de catégories utilisée est un événement élementaire

Mesure de coût d'accès

Le **coût d'accès** aux billets d'un blog b donne une idée de la complexité de l'accès. Il dépend

- du coût de composition de la requête ($cout_{req}$)
- du coût de sélection d'un billet dans l'ensemble des billets retournés ($cout_{doc}$)

$$Cout(b) = cout_{req}(b) + cout_{doc}(b) \tag{4}$$

Mesure de coût d'accès (suite)

- Blogs mono-catégoriels
  - $T_{voc}(b)$: nombre de catégories du blog b
  - $Eff(r)$: effectif moyen de documents ramenés par la requête r

$$C_{mono}(b) = T_{voc}(b) + Eff(r) \tag{5}$$

- Blogs multi-catégoriels : formuler une requête r de longueur l revient à choisir successivement l catégories parmi les T_{voc} disponibles.

$$C_{multi}(b) = E_{r \in Req}\Big(\sum_{i=0}^{l(r)-1} (T_{voc}(b) - i) + Eff_c(r)\Big) \tag{6}$$

- Blogs hiérarchiques : le choix d'une catégorie feuille revient à un parcours dans un arbre de hauteur $h = log_d(T_{voc})$ (pour un arbre complet et équilibré de degré d)

$$C_{arbre}(b) = log_d(T_{voc}(b)).d + Eff_c(b) \tag{7}$$

Expérimentation

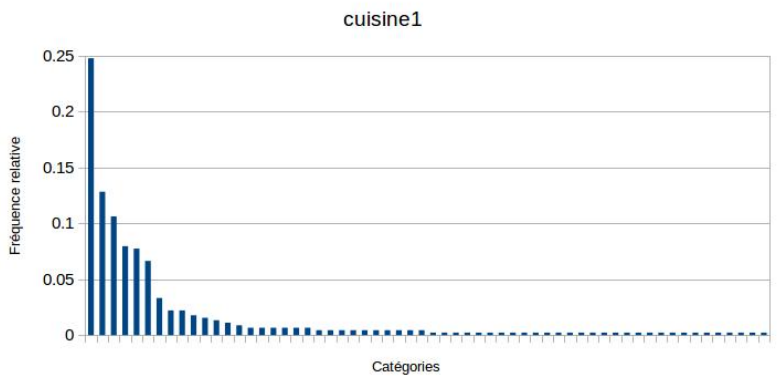
Nous appliquons ces mesures sur les blogs du corpus FLOG [4].

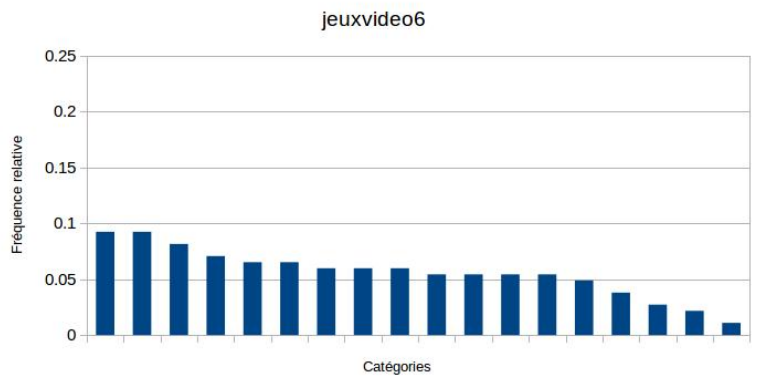
- 20 blogs différents
- 25 000 billets et 11 millions des mots
- billets échelonnés sur une période de 10 ans
- systèmes mono-categoriels et multi-categoriels
- 4 grands thèmes (cuisine, jeux video, technologie et droit)

L'analyse des résultats donne des pistes sur les amélioratons à apporter aux systèmes de catégories.

Une entropie similaire, un équilibre différent

Deux blogs de profils différents

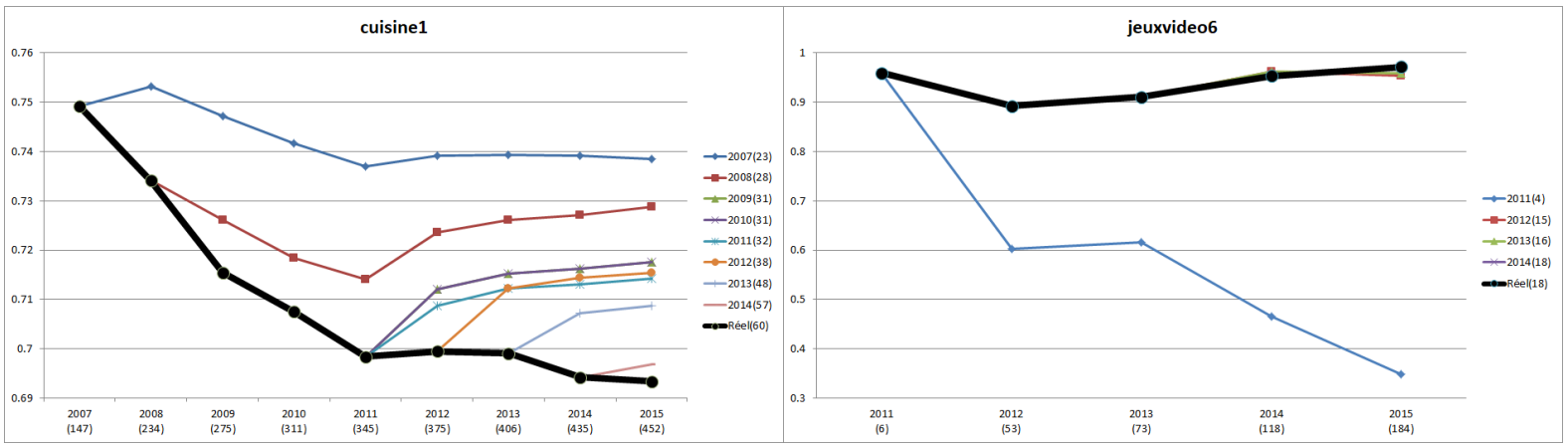




La différence se reflète dans l'équilibre :
cuisine1 - entropie:2,839 entropie max: 4,094 équilibre: 0,693
jeuxvideo6 - entropie:2,809 entropie max: 2,89 équilibre: 0,972

Dégradation de l'équilibre au fil du temps

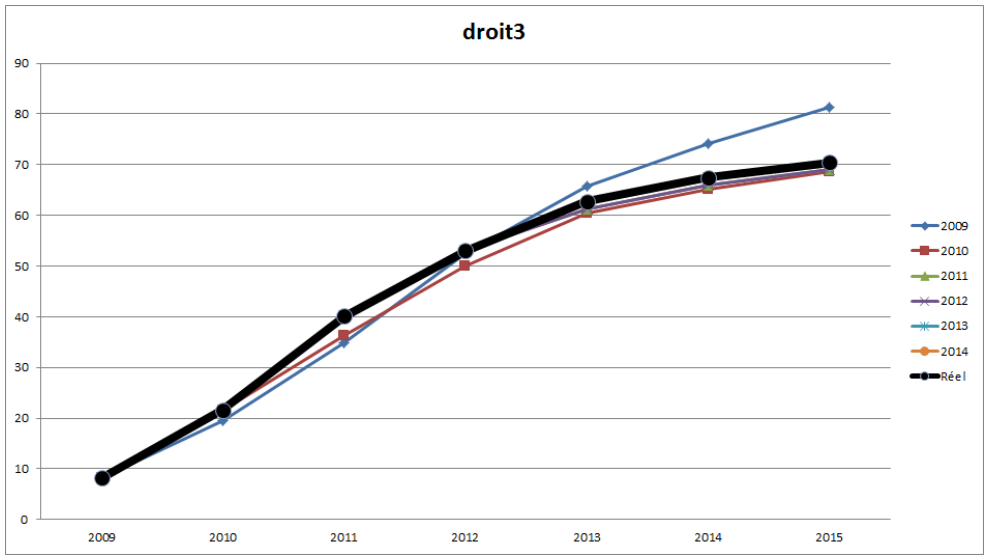
Le suivi des mesures d'équilibre permet de *détecter* une dégradation de la qualité d'indexation qui rend les billets difficiles d'accès pour le lecteur et d'alerter l'auteur du blog.



On observe que pour 17/20 blogs, l'équilibre se dégrade au fil des années.

Augmentation du coût d'accès au fil du temps

Le coût augmente avec l'ajout de nouvelles catégories et de nouveaux billets.



En général

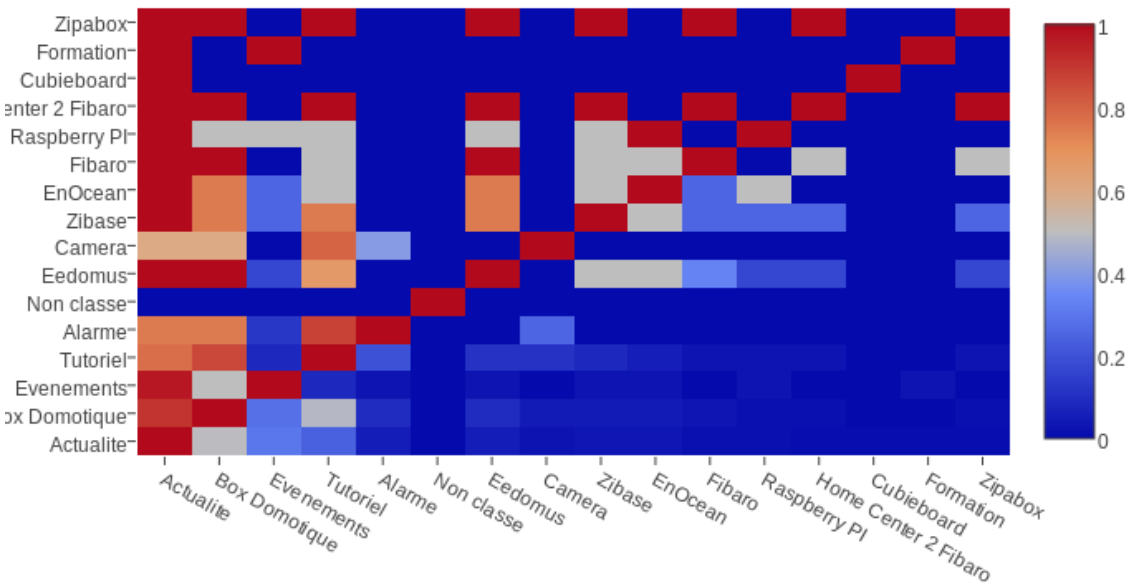
- les systèmes multi-catégoriels sont moins coûteux que leurs correspondants mono-catégoriels
- les systèmes hiérarchiques sont moins coûteux que leurs correspondants multi-catégoriels

Quand le nombre de catégorie devient important, il est pertinent de réorganiser les systèmes en systèmes multi-catégoriels ou hiérarchiques.

Coût d'accès et redondance

Système multi-catégoriels + Coût élévé ⇒ Catégories redondantes ?
Exemple du blog Technologie5

- 88% des billets du blog sont etiquetés "Actualité"
- 9/16 catégories sont incluses dans "Actualité"
- 4 catégories ont une forte intersection avec "Actualité" (de 75% à 97%)
- "Home Center 2" et "Zipabox" sont toujours utilisées ensemble



Une remise à plat du système de catégories du blog s'impose

Vers un diagnostic des systèmes de catégories de blogs

Quand on observe une dégradation, il faut analyser le système de catégories pour décider s'il convient d'intervenir ou non.

Quand l'équilibre est faible, on peut

- décomposer les grosses catégories en sous-catégories
- regrouper des petites catégories

Quand le coût d'accès aux documents est élevé, il faut globalement affiner la granularité

- en décomposant les catégories existantes
- en autorisant les catégories multiples

Quand le coût d'accès aux catégories est élevé, il faut une réorganisation globale du système de catégories en système multi-catégoriel ou hiérarchique.

Conclusion

- Une plateforme de gestion de blog doit proposer des fonctionnalités d'annotation mais aussi offrir des outils de diagnostic permettant d'apprécier et d'améliorer la qualité d'indexation globale offerte par un système de catégories.
- Les mesures proposées dans cet article permettent de fonder ce type de diagnostic. Elles permettent de *détecter* un problème. Il faut ensuite *localiser* les catégories défaillantes et *proposer des mesures correctives* à l'utilisateur.
- C'est l'auteur/annotateur qui choisit ou pas de *réparer* le système de catégories à partir des propositions de correction qui lui sont faites.

Merci!
Des questions ou des commentaires?


Jacob Cohen.
A coefficient of agreement for nominal scales.
Educational and Psychological Measurement, 20(1):37–46, 1960.


Karën Fort, Adeline Nazarenko, and Sophie Rosset.
Modeling the complexity of manual annotation tasks: a grid of analysis.
In *International Conference on Computational Linguistics*, pages 895–910, 2012.


Mark E. Funk and Carolyn Anne Reid.
Indexing consistency in MEDLINE.
Bulletin of the Medical Library Association, 71(2):176–183, 1983.


Ivan Garrido-Marquez, Laurent Audibert, Jorge García-Flores, Fran\c{c}ois Lévy, and Adeline Nazarenko.
A French Weblog Corpus for New Insights on Blog Post Tagging.
In Antonio Moreno Ortiz and Chantal Pérez-Hernández, editors, *CILC2016. 8th International Conference on Corpus Linguistics*,

volume 1 of *EPiC Series in Language and Linguistics*, pages 144–158. EasyChair, 2016.


Kurt Leininger.
Interindexer consistency in psycinfo.
Journal of Librarianship and Information Science, 32(1):4–8, 2000.


Yann Mathet, Antoine Widlöcher, Karën Fort, Claire François, Olivier Galibert, Cyril Grouin, Juliette Kahn, Sophie Rosset, and Pierre Zweigenbaum.
Manual Corpus Annotation: Giving Meaning to the Evaluation Metrics.
In *International Conference on Computational Linguistics*, pages 809–818, Mumbaï, India, December 2012.


E.C. Pielou.
The measurement of diversity in different types of biological collections.
Journal of Theoretical Biology, 13:131 – 144, 1966.


Claude Shannon.

A mathematical theory of communication.
The Bell System Technical Journal, 27:379–423, 623–656, Juillet, Octobre 1948.

Nancy L. Wilczynski and R. Brian Haynes.
Consistency and accuracy of indexing systematic review articles and meta-analyses in medline.
Health Information & Libraries Journal, 26(3):203–210, 2009.