Mesurer la qualité des systèmes de catégories de blogs

Ivan Garrido-Marquez, Jorge Garcia Flores,
François Lévy, and Adeline Nazarenko

LIPN, CNRS & Université Paris 13 – Sorbonne Paris Cité, 99 Jean-Baptiste Clément, 93430 Villetaneuse, France
{garridomarquez, jgflores, fl, adeline.nazarenko}@lipn.univ-paris13.fr

Résumé : Dans le monde prolifique de la blogosphère, retrouver une information pertinente est un défi. Pour aider leurs lecteurs, les auteurs de blog annotent souvent leurs billets avec des catégories mais ce système de catégories est difficile à maintenir quand le blog évolue au fil du temps. Pour guider la révision des annotations des blogs, nous proposons ici deux métriques, l'équilibre et le coût d'accès. Elles permettent de mesurer la qualité formelle des systèmes de catégories considérés en tant que systèmes d'index. Nos expériences montrent qu'en raisonnant sur les caractéristiques du système d'annotation et sur les mesures d'équilibre et de coût qu'on obtient par calcul, on a une vue synthétique sur le système de catégories à un instant t et que cela donne des indications sur les améliorations qu'on peut y apporter. Ces mesures sont conçues pour être intégrées dans un outil de diagnostic et de révision interactive des annotations de blogs que nous développons.

Mots-clés : Mesure de qualité, Catégorisation, Équilibre, Entropie, Blogs, Annotation

1 Introduction

Les blog sont des sites web utilisés pour la publication d'articles d'actualité présentant des informations ou les réflexions d'un ou plusieurs auteurs sur un sujet donné. Ces « billets », qui sont datés et signés, sont présentés par ordre antéchronologique. On estime que la barre des 150 millions de blogs a été franchie en 2010 (Chapman, 2011) et c'est devenu un moyen de communication et d'influence essentiel pour les individus, les entreprises et les médias.

Dans un monde aussi prolifique, retrouver une information pertinente est un défi. Pour aider leurs lecteurs, les auteurs de blogs cherchent généralement à annoter leurs billets en leur associant des catégories et/ou des « tags » (mots-clefs). L'ensemble des annotations aide le lecteur à retouver l'information pertinente en lui permettant de *naviguer* dans le blog (dans le plan de classement ou par proximité thématique) ou de *formuler des requêtes* et de retrouver directement tous les billets associés à un thème donné.

Avec le temps, le nombre de billets augmente et les thèmes évoluent au gré de l'actualité, si bien que le système de catégories proposé à un instant donné peut être beaucoup moins adéquat quelques mois ou quelques années plus tard. Nous nous intéressons ici au rôle d'index que jouent les systèmes de catégories de blogs : est-ce que le jeu de catégories proposé permet à un lecteur de retrouver efficacement les billets qui l'intéressent dans le blog ?

Dans cette optique, nous proposons des métriques permettant de mesurer la qualité d'un système de catégories de blogs. Il ne s'agit pas de vérifier l'adéquation des annotations au contenu des billets (nous supposons que les annotations, souvent posées par les auteurs des billets, sont localement pertinentes) mais de mesurer la qualité globale du système de catégories

* Ce travail s'inscrit dans le cadre des travaux sur l'accès aux contenus développé dans l'axe « Analyse sémantique computationnelle » du Labex EFL (ANR-10-LABX-0083).

comme système d'indexation. On observe en effet que les auteurs, qui annotent leurs billets un à un, n'ont guère de vision globale sur l'efficacité du système de catégories qu'ils construisent de manière de manière incrémentale. Les métriques permettent de détecter quand le système de catégories perd en efficacité et de suggérer les modifications à y apporter.

Après une revue des travaux analysant la qualité des systèmes d'annotation pour l'accès à l'information (sec. 2), nous présentons FLOG, le corpus de blogs que nous avons constitué (sec. 3). Nous introduisons deux mesures complémentaires pour apprécier l'équilibre du système d'annotation et le coût d'accès à l'information pour le lecteur (sec. 4) et nous présentons les résultats obtenus pour les blogs de FLOG (sec. 5). La discussion (sec. 6) montre pour finir comment nous prévoyons d'utiliser ces mesures pour faire le diagnostic des systèmes de catégories des blogs et proposer des mesures correctives aux annotateurs.

2 Etat de l'art : la qualité des systèmes d'annotation

Dans la littérature, la qualité des systèmes d'annotation est d'abord étudiée sous l'angle de l'adéquation entre l'annotation et la sémantique du texte annoté. Des études de ce type sont réalisées depuis longtemps sur l'indexation de ressources bibliographiques, notamment sur le domaine médical : Funk & Reid (1983) ont testé la cohérence d'indexation de 9 catégories définies à partir des en-têtes, sous-titres et concepts de MESH dans un ensemble d'articles en vue d'améliorer la fiabilité des stratégies de recherche. Leininger (2000) analyse en détail la cohérence inter-indexeur dans la base de données PsycINFO en utilisant deux mesures proposées par Hooper (1965) et Rolling (1981). Wilczynski & Haynes (2009) s'intéressent également à la capacité discriminante du vocabulaire d'indexation pour mesurer la qualité d'un système d'indexation ou d'annotation, tandis que (Cohen, 1960; Mathet *et al.*, 2012) ont évalué la qualité des systèmes d'indexation basé sur un vocabulaire contrôlé (Funk & Reid, 1983; Leininger, 2000; Wilczynski & Haynes, 2009). Nous ne nous intéressons pas directement à la qualité des annotations, que nous supposons bonne [1]. Nous cherchons à mesurer la qualité d'un système d'annotation considéré comme un outil d'accès à l'information. Dans cette perspective, il faut considérer le système d'annotation comme l'association d'un jeu de catégories, d'une collection de documents et de l'ensemble des liens d'annotations qui relient les catégories aux documents.

3 FLOG, un corpus de blogs français

Ce corpus (Garrido-Marquez *et al.*, 2016) contient 20 blogs différents, 25 000 billets et 11 millions des mots. Les blogs relèvent de 4 grands thèmes (cuisine, jeux video, technologie et droit) et le corpus couvre une période de 10 ans. Les billets sont annotés par leurs auteurs avec des catégories et/ou des tags.

On observe sur ce corpus que les habitudes d'annotation varient d'un blog à l'autre. Nous nous intéressons ici spécifiquement aux annotations de type catégories. Le nombre de catégories par blog varie entre 4 et 91 et pas nécessairement en proportion du nombre de billets, puisque le nombre moyen de billets par catégorie va de 2 à 64 (à l'arrondi près).

1. Nous faisons l'hypothèse que les auteurs de billets savent dans quelle(s) catégorie(s) les ranger ou disposent d'outils pour les aider à le faire, à partir de l'analyse du contenu du billet.

Les annotateurs utilisent les systèmes de catégories de différentes manières. Dans les *systèmes mono-catégoriels*, les catégories sont utilisées de manière exclusive et un billet n'est annoté que par une seule catégorie. Il y a 6 blogs de ce type dans FLOG. Dans les *systèmes multi-catégoriels*, un même billet peut être associé à plusieurs catégories. Enfin, les systèmes de catégories peuvent être structurés hiérarchiquement (*systèmes hiérarchiques*). Le corpus FLOG ne contient aucun blog de cet type, même si `technologie2` s'en rapproche [2].

4 Mesurer l'équilibre et le coût d'accès d'un système d'annotation

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

S'agissant d'un blog, la distribution ne comporte qu'un ensemble fini $\mathcal{F}$ d'événements qui sont des unions d'événements élémentaires. Dans l'analyse de Shannon (1948), la quantité d'information $I(e)$ recelée par un événement particulier e est $-\log_b(P(e))$ et l'entropie est la valeur espérée de cette quantité d'information. En notant $\mathcal{A}$ les événements élémentaires de $\mathcal{F}$, on peut mesurer l'entropie H :

$$H = E_{\mathcal{F}}[I(e)] = \sum_{e \in \mathcal{A}} P(e)I(e) = -\sum_{e \in \mathcal{A}} P(e)\log_b(P(e)) \tag{1}$$

Pour les systèmes mono-catégoriels, l'adaptation est directe : un événement élémentaire est une catégorie et l'on utilise la fréquence de la catégorie comme sa probabilité. La mesure d'entropie ne suffit cependant pas pour comparer deux blogs ou deux versions différentes d'un système de catégories, parce que la valeur maximale dépend du nombre de catégories (le meilleur système aurait une catégorie par billet !). L'*équilibre* (Pielou, 1966) rapporte donc l'entropie calculée par l'équation 1 à l'entropie maximum susceptible d'être obtenue avec le même nombre de catégories, soit pour un blog x comportant N billets et n catégories x_i ($i = 1 \ldots n$) :

$$H(x) = -\sum_{i=1}^{n} \frac{|x_i|}{N}\log_b(\frac{|x_i|}{N}) \qquad Equilibre(x) = \frac{H(x)}{\max(H(x))} = \frac{H(x)}{\log_b(n)} \tag{2}$$

Dès lors qu'un billet peut être annoté par plusieurs catégories, ce qui est le cas dans 2/3 des blogs du corpus FLOG, les catégories ne représentent plus des événements élémentaires car certains billets sont décrits par une *combinaison de catégories*. On peut montrer cependant que les événements élémentaires sont calculables à partir des combinaisons de catégories sans négation, ce qui permet d'utiliser la mesure d'équilibre ci-dessus.

Nous cherchons également à apprécier le *coût d'accès* aux billets au sein d'un blog indépendamment des interfaces de navigation proposées aux utilisateurs. Nous considérons pour cela le schéma de base où un lecteur formule une requête composée d'une ou plusieurs catégories et reçoit en retour un ensemble de billets qu'il doit parcourir pour trouver celui qui l'intéresse. Le coût d'accès aux billets du blog b s'exprime comme la somme des coûts de composition de la requête ($cout_{req}$) et de sélection d'un billet dans l'ensemble des billets retournés ($cout_{doc}$) :

$$Cout(b) = cout_{req}(b) + cout_{doc}(b) \tag{3}$$

2. Comme les billets annotés par les catégories feuilles de l'arbre sont parfois aussi annotés par des catégories plus génériques, nous le considérons comme un système multi-catégoriel.

Blog	**Type**	**T**$_{voc}$	**N**$_{doc}$	**Équilibre**	C_{cat}	C_{doc}	**Coût**
cuisine1	Mono	60	452	0,69	60	7,53	67,53
cuisine2	Mono	26	927	0,82	26	35,65	61,65
jeuxvideo1	Multi	43	1422	0,82	140,76	46,85	187,62
technologie1	Multi	56	1423	0,63	59,81	229,88	289,69
technologie5	Multi	16	132	0,73	33,59	34,32	67,92
droit1	Mono	4	243	0,72	4	60,75	64,75
jeuxvideo2	Multi	33	1234	0,86	129,93	6,17	136,11
technologie2	Multi	38	243	0,89	69,87	7,62	77,50
jeuxvideo3	Multi	91	5486	0,76	91	335,04	426,04
jeuxvideo4	Multi	40	1501	0,80	87,01	30,01	117,02
droit2	Multi	48	931	0,64	76,65	121,63	198,29
cuisine3	Mono	50	395	0,83	50	7,90	57,90
technologie3	Multi	41	343	0,84	53,28	24,08	77,37
droit3	Multi	13	283	0,73	17,76	52,66	70,42
cuisine4	Mono	25	1561	0,69	25,00	62,44	87,44
droit4	Multi	15	1572	0,57	43,13	161,21	204,34
technologie4	Mono	12	573	0,74	12,00	47,75	59,75
jeuxvideo5	Multi	37	1134	0,90	105,28	11,30	116,58
technologie6	Multi	16	374	0,85	95,70	16,20	111,91
jeuxvideo6	Multi	18	184	0,93	18,18	11,71	29,90

TABLE 1 – Caractéristiques des blogs du corpus FLOG et mesures d'équilibre et de coûts

Pour un blog ayant un système d'annotation multi-catégoriel (dont le mono-categoriel est pour ce calcul un cas particulier), il faut tenir compte de la taille T_{voc} du vocabulaire de catégories proposé. Formuler une requête r de longueur l revient à choisir successivement l catégories parmi les T_{voc} disponibles, ce qui a un coût $cout_{req}(r) = \sum_{i=0}^{l-1}(T_{voc}(b) - i)$. Le coût $cout_{req}(b)$ de composition des requêtes du blog est l'espérance de ce coût de composition pour une requête. De même, le coût de sélection du document est l'espérance du nombre $Eff_c(r)$ de documents ramenés par r.

$$C_{multi}(b) = E_{r \in Req}\Big(\sum_{i=0}^{l(r)-1} (T_{voc}(b) - i) + Eff_c(r)\Big) \tag{4}$$

Dans un système hiérarchique, les billets sont généralement décrits par une seule catégorie mais le choix de la catégorie est guidé par la structure arborescente. Considérons pour simplifier un arbre complet et équilibré de degré d. La hauteur h de cet arbre est $h = log_d(T_{voc})$. Pour sélectionner une catégorie feuille, il faut choisir h fois une catégorie parmi les d disponibles à chaque niveau. En notant Eff_c l'effectif moyen des catégories feuilles, on a :

$$C_{arbre}(b) = log_d(T_{voc}(b)).d + Eff_c(b) \tag{5}$$

5 Résultats expérimentaux

Le tableau 1 présente les mesures d'équilibre et de coûts pour chacun des 20 blogs du corpus FLOG. L'équilibre, qui varie entre $0,57$ et $0,93$, donne une idée rapide de la distribution des billets dans les catégories des différents blogs : on voit par exemple que les billets de `jeuxvideo6` sont plus uniformément distribués que ceux de `cuisine1` [3] ; on voit aussi que

3. De fait, `cuisine1` présente une catégorie majoritaire associée à près de 25% des billets ; près de 50% des billets se retrouvent dans les 3 principales catégories et 85% des billets se concentrent sur seulement 15 catégories.

les mesures de coûts sont extrêmement variables (de 20 à 420) et qu'on peut avoir un coût d'accès élevé eu égard au nombre de documents même pour un blog relativement équilibré (ex. `technologie2`), preuve que les deux mesures sont complémentaires.

L'évolution de ces mesures dans le temps est également intéressante. Elle montre que les auteurs, mêmes s'ils veillent à l'adéquation des catégories qu'ils posent au contenu des billets (analyse locale), ne se rendent pas toujours compte de la qualité de l'index créé par ces catégories (analyse globale). Dans la majorité des cas, on observe en effet que l'équilibre se dégrade au cours du temps. Les graphiques de la figure 1 le montrent pour les blogs `jeuxvideo3` et `cuisine4` (courbes noires). Dans certains cas, l'ajout de nouvelles catégories limite la dégradation de l'équilibre (ex. `jeuxvideo3`) mais dans d'autres, cela paraît au contraire contre-productif (ex. `cuisine4`)[4]. On fait la même constatation sur les graphes de coût.

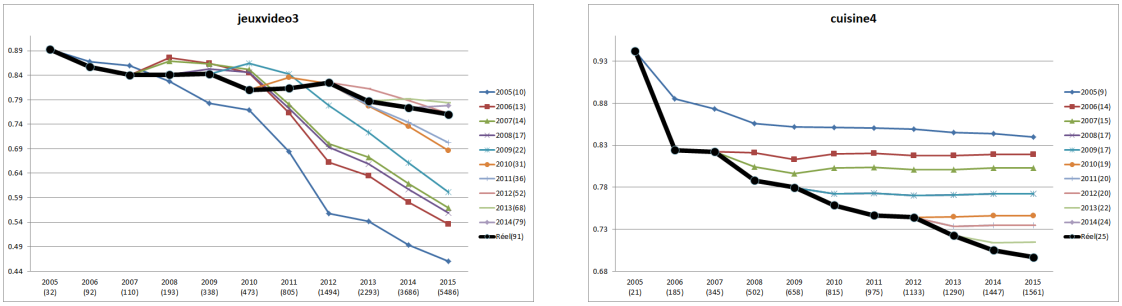

FIGURE 1 – Evolution de l'équilibre (courbes noires) et projections de l'équilibre à jeu de catégories constant (courbes de couleur)

6 Discussion : vers un diagnostic des systèmes de catégories de blogs

L'analyse longitudinale d'un corpus de blogs et les mesures que nous avons introduites pour mesurer la qualité de l'indexation proposée par les systèmes de catégories montrent que cette qualité d'indexation est difficile à contrôler par les auteurs de blogs qui se préoccupent prioritairement de la qualité des annotations qu'ils posent au regard des contenus qu'ils cherchent à annoter et qui n'ont qu'une vision locale des blogs qu'ils annotent.

Nous considérons qu'une plateforme de gestion de blog doit non seulement proposer des fonctionnalités d'annotation – permettre de poser des tags/catégories sur les billets, avec éventuellement des outils d'aide à l'annotation – mais qu'elle doit aussi offrir des outils de diagnostic permettant d'apprécier et d'améliorer la qualité d'un système de catégories en termes d'indexation. Les mesures proposées dans cet article permettent de fonder ce type de diagnostic.

Le suivi des mesures d'équilibre et de coût 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. Il faut ensuite *localiser* les catégories défaillantes et proposer des mesures correctives à l'utilisateur. Les indications dépendent des mesures obtenues. Si l'équilibre est faible, on peut soit décomposer les grosses catégories en sous-catégories soit regrouper des petites catégories. Si le

4. Les courbes en couleur montrent les mesures d'équilibre qu'on aurait obtenues si on avait gardé un jeu de catégories inchangé : les courbes bleu clair retracent ainsi l'évolution de l'équilibre qu'on aurait obtenu en conservant le jeu de catégories de 2005 jusqu'en 2015.

coût d'accès aux documents est élevé, il faut globalement affiner la granularité des catégories, soit en décomposant les catégories existantes, soit en introduisant des catégories indépendantes (le système devient multi-catégoriel) pour réduire le coût d'accès aux documents sans augmenter trop le nombre de catégories. Quand le coût d'accès aux catégories est élevé, il faut une réorganisation globale du système de catégories en système multi-catégoriel ou hiérarchique. Il arrive aussi que certains systèmes de catégories soient inefficaces car redondants – c'est le cas du blog `technologie5` – et on s'en rend compte quand on observe qu'on a un système multi-catégoriel avec des coût d'accès aux catégories et aux documents tous les deux élevés.

Établir la liste des corrections à proposer à l'auteur nécessite cependant de compléter l'analyse globale qui est faite ici par une analyse plus détaillée. Il faut localiser les catégories les plus grosses, les plus petites ou les plus redondantes. Il faut également tenir compte de l'âge des catégories et de leur taux d'activité : il est inutile de regrouper des catégories qui sont en croissance forte mais peut-être urgent, à l'inverse, de décomposer une catégorie importante qui continue à se développer. Il faut enfin prioriser les corrections à faire et tenir compte du coût induit par la correction (combien de billets faut-il réannoter ?).

C'est l'auteur qui choisit ou pas de *réparer* le système de catégories à partir des propositions de correction qui lui sont faites mais on voit que les mesures proposées ici permettent d'établir un diagnostic et de l'éclairer sur la qualité de l'index que constitue son système de catégories.

Références

CHAPMAN C. (2011). A brief history of blogging. `http://www.webdesignerdepot.com/2011/03/a-brief-history-of-blogging/`. [Marketing, Web Design, WordPress · Mar 14, 2011].

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

FUNK M. E. & REID C. A. (1983). Indexing consistency in MEDLINE. *Bulletin of the Medical Library Association*, **71**(2), 176–183.

GARRIDO-MARQUEZ I., AUDIBERT L., GARCÍA-FLORES J., LÉVY F. & NAZARENKO A. (2016). A French Weblog Corpus for New Insights on Blog Post Tagging. In A. M. ORTIZ & C. PÉREZ-HERNÁNDEZ, Eds., *CILC2016. 8th International Conference on Corpus Linguistics*, volume 1 of *EPiC Series in Language and Linguistics*, p. 144–158 : EasyChair.

HOOPER R. S. (1965). *Indexer consistency tests : origin, measurement, results, and utilization*. Rapport interne, IBM Corporation, Bethestda, MD.

LEININGER K. (2000). Interindexer consistency in psycinfo. *Journal of Librarianship and Information Science*, **32**(1), 4–8.

MATHET Y., WIDLÖCHER A., FORT K., FRANÇOIS C., GALIBERT O., GROUIN C., KAHN J., ROSSET S. & ZWEIGENBAUM P. (2012). Manual Corpus Annotation : Giving Meaning to the Evaluation Metrics. In *International Conference on Computational Linguistics*, p. 809–818, Mumbaï, India.

PIELOU E. (1966). The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology*, **13**, 131 – 144.

ROLLING L. (1981). Indexing consistency, quality and efficiency. *Information Processing & Management*, **17**(2), 69–76.

SHANNON C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 379–423, 623–656.

WILCZYNSKI N. L. & HAYNES R. B. (2009). Consistency and accuracy of indexing systematic review articles and meta-analyses in medline. *Health Information & Libraries Journal*, **26**(3), 203–210.