Controlling the Drift of Semantic Indexing Systems

Ivan Garrido Marquez, Jorge Garcia Flores, François Lévy and Adeline Nazarenko
LIPN, Paris University 13 – Sorbonne Paris Cité & CNRS, Av. J.-B. Clément, Villetaneuse, France

Keywords: Quality Measures, Document Classification, Blog Posts Categories.

Abstract: Document classification is often meant to serve as semantic indexing to help readers finding documents related to a given topic. However, the quality of indexing often deteriorates with time: some categories are misused or forgotten by indexers, others become obsolete or too general to be useful. This paper proposes measures to assess the quality of an indexing system and an algorithm that guides indexers in restructuring their indexes. Focus is put on the reader's rather than on the annotator's point of view (Does the classification really help accessing information? *vs.* Is a category adequate with the content of the document?). The whole approach is illustrated on a corpus of 20 blogs which posts are associated with categories. We show that indexers have difficulties to adapt the blogs indexing systems when the number of posts increases and we show that our approach can significantly improve the quality of these indexing systems, by simulating blog restructuring.

1 INTRODUCTION

For a long time, librarians have been trying to organize documentary collections according to classification plans to facilitate access to information. In addition to the full-text search functionalities, people continue to categorize documents, often using automatic classification tools. This classification can be considered as a semantic indexing: classifying newspaper articles or blog posts allows journalists or readers to quickly find documents that have been published in the past in relation to a given topic.

However, quite often, these indexing systems drift over time, either because the relative importance of the different covered topics evolves (the number of documents related to a topic fluctuates depending on the news), or because people in charge of indexing (indexers) change their way of indexing (*e.g.* one category replaces another, some categories are "forgotten"). Indeed, a document is always indexed on a "local" basis, when it is published. The indexer usually does not have a global view of the indexing system, so more that he does not know in advance which categories will be more or less prevalent in the future.

There is a need for a tool assessing the quality of the indexing system and helping to restructure it if necessary. Two distinct user roles are considered here: the indexer (information provider), who needs to be assisted in his/her indexing or annotation work, and the final user or reader (information consumer), who needs an efficient access to the documents.

This paper proposes measures to assess the quality of an indexing system for its readers, and an algorithm that possibly guides the indexer's restructuring work. It is not an indexing system, in the sense that it does not consider the adequacy of a category to its content, but a system intended to control the overall quality of an existing indexing system. The focus is therefore put in the readers' point of view.

We are interested here in document flows and we consider a set of blogs over a period of 10 years. The analysis of the existing posts classifications confirms that indexing is difficult to master even for small document collections, but we propose a restructuring algorithm that significantly improves the quality of the indexing systems created over time by indexers.

Section 2 recalls work related to knowledge-based document classification in a broad sense, far beyond the indexing perspective which is ours. Section 3 presents in more details category-based indexing systems and shows the heterogeneity of classification practices on our blog corpus. Sections 4 and 5 present the quality measure that are used for auditing the quality of indexes and the algorithm that help indexers to restructure their indexes when necessary. Simulations

This work has been partially funded by the French Ministry of National Education, Higher Education, Training and Scientific Research and is supported by the French National Research Agency (ANR-10-LABX-0083) in the context of the Labex EFL.

Marquez, I., Flores, J., Lévy, F. and Nazarenko, A.
Controlling the Drift of Semantic Indexing Systems.
DOI: 10.5220/0006926501990206
In *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018) - Volume 2: KEOD*, pages 199-206
ISBN: 978-989-758-330-8

show the impact of the suggested restructuring on the quality of indexes of our blog corpus.

2 RELATED WORKS

2.1 Document Classification

Document classification is expected to help document search and navigation in document collections. For a long time, librarians used document subjects, classes and categories to organize the resources in a library, based on their content. The development of full text search as well as the advent of folksonomies and the exploitation of web query log have popularized alternative approaches based on users' requests. Document classification has also become a widespread practice. Bloggers, for instance, often organize their posts into categories that are displayed in the blog interface to help readers explore blog content.

Automatic document classification has received a lot of attention (Sebastiani, 2002) and various level of supervision (supervised, unsupervised or semi-supervised) have been explored. In the literature, the quality of document classification was first studied from the point of view of the adequacy between the categories and the content of the indexed document. Since there is not a single good way to index content, emphasis has been put on indexing consistency. Studies have been carried out for a long time on the indexation of bibliographic resources, in particular for the medical field (Funk and Reid, 1983; Leininger, 2000). (Cohen, 1960; Mathet et al., 2012) have evaluated the quality of indexing or annotation systems based on the agreement between indexers.

(Blei et al., 2003) proposed a fully automated method, which computes both the classes and the classification from a document collection. Classes are characterized as sets of topics and topics are infered from the words cooccurring in documents. The evaluation compares the learning performances of the topic model with that of other models on the same task.

2.2 Index Properties

Evaluating the quality of a document classification as a formal indexing system is a different issue. The basic relevant analysis is Shanon's theory of information (Shannon, 1948). Within a given a system, it says that dominant category is less informative than the others, as its selection brings little information about the indexed documents. On the opposite, a rare category contains much more information when it occurs but, being seldomly selected, it almost never convey its information and it is not informative either. This vision of information is expressed through the notion of entropy (Shannon, 1948; Battail, 1997).

The *informativeness* of a category system only partially accounts for the process of accessing documents because it does not take into accout the size of the classification. Considering annotations as a tool to help readers of electronic media, (Jan et al., 2016) points out the fact that limiting the number of available annotations[1] is cognitively more effective: while a small number of annotations improves the reader's attention, their mutiplication hinders understanding.

Since we are considering category structures as tools to facilitate document access, we are also relying on classical algorithmic results regarding the complexity of the access to data structures (such as search trees) to measure the document access costs for potential users (Cormen et al., 2009).

3 CATEGORY-BASED INDEXING SYSTEMS

To evaluate the quality of an indexing system, one must consider how its potential users would use it to access information. We illustrate this in the case of readers searching for blog posts. We show that the efficiency of a blog indexing system varies greatly from one blog to another and usually tends to decrease over time.

3.1 Category-based Document Access

The quality of the category-based indexing system is considered from a formal point of view, on the basis of the elementary operations that must be done to find the document that meets one's requirements[2]. We evaluate the quality of the indexing system by estimating the average searching time based on the elementary operations performed by users.

Accessing a document is a two-step process for readers (Fig. 1). They have to select the category or categories that best match their information requirement and to browse the documents retrieved by the system until they have read the whole set of documents or

[1] Those annotations are free-text annotations more similar to readers' comments than to well structured metadata used to facilitate information access.

[2] Thus disregarding any other means of information access, such as keyword search, or the ergonomics of the interfaces that may be proposed to readers.

found a relevant document[3].

We evaluate the quality of indexing systems by estimating the average number of elementary operations (querying/category selection and document browsing) performed by users searching for documents.

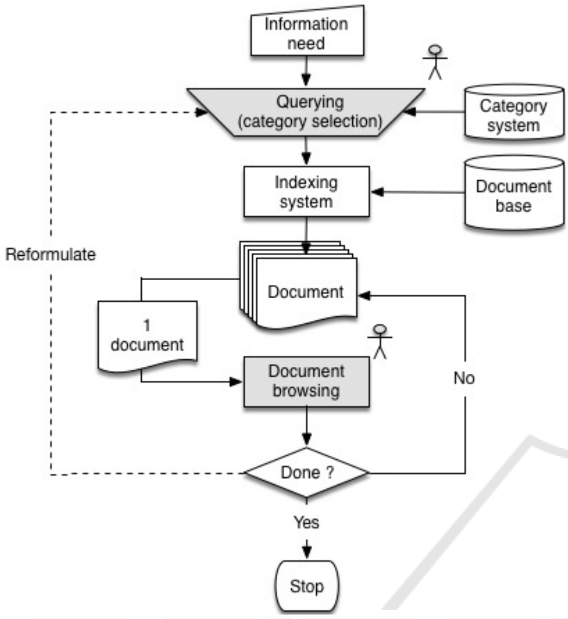

Figure 1: Model of access process (the user is a reader).

Several cases must be considered, depending on the number of categories that can be associated to a document[4]. In *mono-categorial systems*, each document is associated to exactly one category, whereas a document can be associated to several categories in *multi-categorial systems*.

We consider that the users formulate their query all at once, even if the query consists of a list of categories. A query corresponds to a *basic category* or to a *compound category* (a set of basic categories) and the system returns the set of documents which have been indexed with that/all these category/ies.

3.2 Blog Posts Indexing

We experiment our approach for controlling the drift of indexing system on the FLOG corpus (Garrido-Marquez et al., 2016), where each blog consists of a flow of documents or posts. The posts are traditionally annotated for easy access and consultation by its potential readers. The indexers are usually the blog author(s) or editor(s). Quite often, the blogs support two types of annotations: the tags corresponding to freely chosen keywords, and a smaller set of categories used as a controlled vocabulary. Even if blogging systems propose different types of annotations for indexers, the blogging activity is done freely, without any guideline for post annotation. The indexing practices vary a lot from one blogger to another, even among the same blog.

Table 1 gives an overview of the FLOG corpus, composed of 20 blogs in French covering 4 different domains related to cooking, technology, law and video games over a 10-year period (2005-2015). 8 blogs rely on a mono-categorial system while the others often associate several categories to a single post. The table shows how heterogeneous blogging practices are w.r.t. blog size (from 184 to 6587 posts resp. for `jeuxvideo6` and `jeuxvideo3`), growth speed (some multi-author blogs may publish several posts per day), size of the vocabulary of categories (indexers cannot manage a set of 60 or 91 categories as they do for a set of 4 categories), mean size of a category (from few to several hundreds of documents).

It also appears that the quality of post categorization varies a lot from one blog to another. Some blogs, such as `jeuxvideo3`, are hardly manageable and bloggers seem to have difficulties mastering their own categorization system. Actually, the indexers have introduced 91 different categories, ranging from 1 to 932 documents, the largest covering 1/6th of the blog.

3.3 The Drift of Indexing Systems

The blogs of the FLOG corpus cover a rather long period (10 years in most cases), which allows a diachronic analysis. Of course, indexers introduce new categories when the underlying topic of the blog evolves (Fig. 2), but often not enough to counterbalance the increasing number of posts or to give an efficient access to users (*e.g.* the curves of `droit4` or `jeuxvideo3`). Some categories become huge with time (Table 1 shows that 8 out the 20 blogs have categories with more than 500 posts) while others remain only scarcely used or simply stop being used.

4 QUALITY MEASURES

Assessing the quality of indexing systems and controlling their drift require adequate measures.

4.1 Balance

The first measure of quality for an indexing system refers to the amount of information that it conveys.

[3] We do not consider here the case where the user reformulates the initial query.

[4] We assume we have a flat vocabulary of categories.

Table 1: Blogs from the FLOG corpus (all of them in French). The size of a category corresponds to the number of posts that are associated with that category.

blogname	#posts	#cats	min cat. size	mean cat. size	max cat. size
cuisine1	452	60	1	7.5333	112
cuisine2	927	26	1	35.6538	155
cuisine3	395	50	1	7.9000	42
cuisine4	1561	25	1	62.4400	401
droit1	243	4	8	60.7500	154
droit2	931	48	1	32.7292	277
droit3	283	13	1	30.6154	135
droit4	1752	15	1	366.6000	1481
jeuxvideo1	1422	43	1	112.7674	866
jeuxvideo2	1234	33	5	156.4242	1120
jeuxvideo3	5486	91	1	60.0110	932
jeuxvideo4	1501	40	1	83.1250	913
jeuxvideo5	1134	37	2	90.1892	548
jeuxvideo6	184	18	2	10.3333	17
technologie1	1423	56	1	27.1786	522
technologie2	243	38	1	11.9737	62
technologie3	343	41	1	10.9512	73
technologie4	573	12	2	47.7500	177
technologie5	132	16	1	18.4375	117
technologie6	374	25	2	62.4800	262

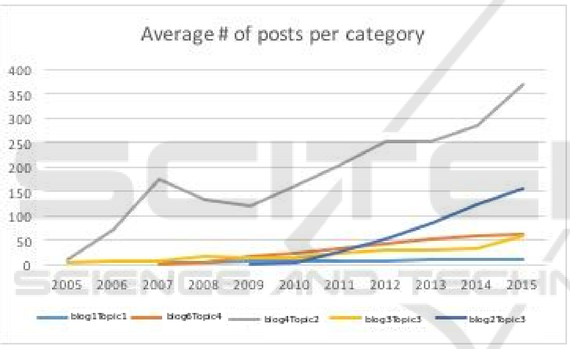

Figure 2: Evolution of the category-based indexing systems of 5 blogs from the FLOG corpus.

We model the searching process as a document draw: a document is (imperfectly) identified by its index or by its category. Since Shanon's work, the amount of information of a finite space of events is measured by its entropy, *i.e.* $H = -\sum_{e \text{ atomic}} P(e)Log(P(e))$. The entropy of an indexing system is therefore:

$$H(b) = -\sum_{i=1}^{|V_C|} \frac{freq(c_i)}{N} Log(\frac{freq(c_i)}{N})$$

where $\left| \begin{array}{l} b \text{ is a document base with } N \text{ documents} \\ |V_C| \text{ is the number of categories} \\ freq(c_i) \text{ is the size of the } i^{th} \text{ category} \end{array} \right.$

To measure the entropy of a multi-category system, we transpose it into a new formal mono-categorial one having the same power of documentary discrimination but a broader vocabulary of categories and we compute entropy on the formal categories[5].

This measure can be used to compare two systems having the same number of categories. To disregard the number of categories, we use a derivative measure that has been defined to compare the diversity of animal population over different territories (Pielou, 1966): the *balance* is a normalized entropy compared to the maximum entropy that can be reached with the same number of categories. It is defined as follows:

$$balance(b) = -\frac{1}{\log(|V_C|)} \cdot H(b)$$

The balance ranges from almost 0, when most documents are in the same category, to 1, when all categories have the same frequency. Intuitively, we consider that a balance of 0.8 or lower is poor.

Fig. 3 presents two graphs showing the evolution of `cuisine1` with this respect[6]. The top one shows the balance over time. The black thick curve shows that the balance is rapidly deteriorating (from 0.75 to

[5] Let's take a simple example of a multi-categorial system to illustrate this point. In a given document base, the documents d_1 and d_2 are associated to the category c_A, d_3 is associated to c_B whereas d_4 and d_5 are associated with both c_A and c_B. c_A and c_B are two basic categories, respectively containing d_1, d_2, d_4, d_5 and d_3, d_4, d_5. The compound category $c_A \cdot c_B$ contains d_4 and d_5. That multi-categorial system can be transposed into the following formal mono-categorial system:
$c_1 = \{d_4, d_5\} \Leftrightarrow c_A.c_B$
$c_2 = \{d_1, d_2\} \Leftrightarrow c_A.\neg c_B$
$c_3 = \{d_3\} \Leftrightarrow \neg c_A.c_B$

[6] The full quality results of the corpus are available on http://lipn.univ-paris13.fr/g̃arridomarquez/corpusanalysis/

0.69). The other curves show what would have happened if the new categories had not been added: it appears that the new categories (from 23 in 2007 to 60 in 2015) actually aggravate the imbalance. The bottom graph gives the distribution of posts over the 60 categories of the blog in 2015: it shows an overwhelming category that clusters 25% of the posts and a long trail of very small categories with 1 or 2 posts. Although one could expect a mono-author and mono-categorial blog such as `cuisine1` to be well structured, its balance appears to be poor. This shows how difficult it is for an human indexer to maintain a balanced category system over time.

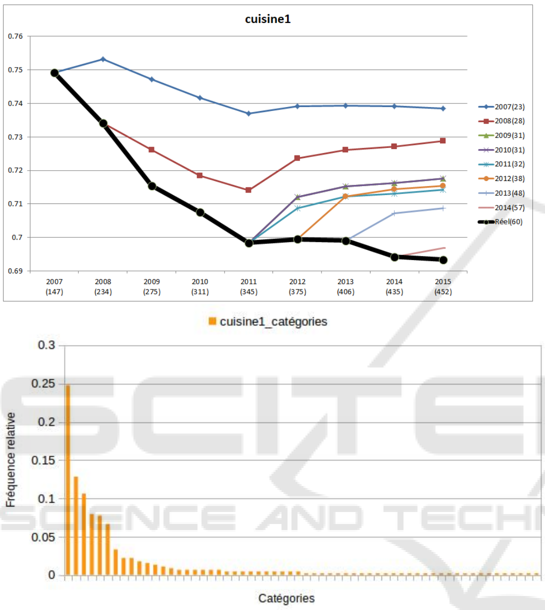

Figure 3: Drift of the balance - an example.

4.2 Access Cost

We also consider the effort required for retrieving a document. Based on the model presented in Fig. 1, we consider that the *access cost* depends on the efforts required 1) for selecting a basic or compound category that is used for querying the document base (*querying cost*), and 2) for browsing the documents returned by the system (*browsing cost*).

If the categories form a flat vocabulary, the querying cost is related to the size of that vocabulary, $|V_C|$, as one has to choose 1 category among $|V_C|$ ones. In the compound case, one has to choose successively each of the basic categories that compose the compound one. In our simple access model, every returned documents is browsed, which implies that the browsing cost depends directly on the number of documents returned by the system for the selected category, be it basic or compound.

The cost is modeled as the expected cost of access over the whole set of queries that return a non-empty set of documents. The following formula covers the cases of the mono- and multi-categorial systems, on which we focus in this paper:

$$Cost(b) = E_{q \in Q}\left(\alpha \sum_{i=0}^{l(q)-1} (|V_C| - i) + \beta \mathit{freq}(q)\right)$$

$$\text{where} \left| \begin{array}{l} b \text{ is a document base} \\ V_C \text{ is the vocabulary of basic categories} \\ Q \text{ is the set of non-empty queries} \\ l(q) \text{ is the number of categories composing } q \\ \mathit{freq}(q) \text{ is the number of doc. associated to q} \\ \alpha,\ \beta \text{ are formal parameters (see below)} \end{array} \right.$$

By default, we assume that all categories are equiprobable and that the querying and browsing costs have the same weight in the global access cost ($\alpha = \beta = 1$)[7]. To appreciate the cost of an indexing system, we compare it to the optimal cost that can be obtained for a mono-categorical indexing system containing N documents (hereafter, *reference cost*). This minimal cost is obtained for the indexing system consisting of $\sqrt{N}$ categories, each associated with $\sqrt{N}$ documents ($cost(b) = 2\sqrt{N}$).

Figure 4 shows the evolution of the cost for the blog `cuisine1`. The black thick curve shows that the cost increases with time (from 29.4 in 2007 to 67.5 9 in 2015). Such an increase is expected as the blog is constantly enriched with new posts and categories but the black curve exceeds the other colored ones, which show what would have been the cost if no new category had been added. In this case, the introduction of new categories degrades the access cost, which shows how difficult it is, for human indexers, to control the introduction of new categories.

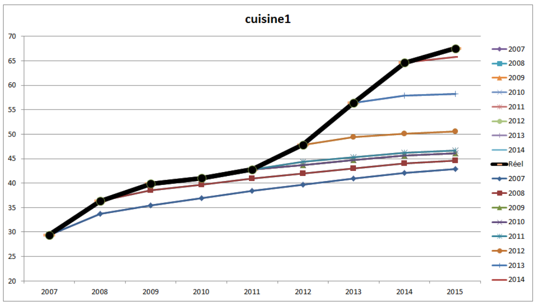

Figure 4: Drift of cost - example.

[7] In concrete applications, these relative weights can be tuned to take into account the types of documents or the functionalities offered by a users' interfaces.

4.3 Redundancy

The balance and the access costs are global measures that may hide local problems which matter for efficiency. To analyze the relevance of the combinations of categories in multi-categorical systems, we compute different redundancy scores between pairs of categories. This paper focuses on the overlap score.

A simple Jaccard Index (Manning and Schütze, 1999) measures the *overlap* between two categories:

$$O(c_1,c_2) = \frac{|c_1| \cap |c_2|}{|c_1| \cup |c_2|}$$ where c_1 and c_2 are two categories considered as sets of documents

This score ranges from 0 for categories that have no document in common to 1 for categories that share exactly the same set of documents.

Fig. 5 presents the overlapping analysis of Blog `technologie6` in the form of a heat map. Each cell corresponds to a pair of categories and the color indicates the degree to which they overlap. In multi-categorical systems, a certain degree of overlapping is naturally welcome, but we consider that the overlap is too large if the size of the intersection between two categories c_1 and c_2 is bigger that the complementary parts in c_1 or c_2 (score of 0.3 or light blue color).

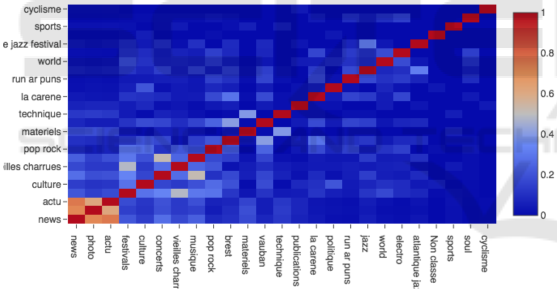

Figure 5: Overlapping map of Blog `technologie6`. The categories are ordered from the bigger (at the bottom left corner) to the smaller (at the upper right corner).

The yellow-orange cells on the `technologie6` map points out categories that are both very large and very redundant with each other. Those `news`, `photo` (picture) and `actu` (a French term for *news*) categories cover a large part of the blog. It seems that the indexer uses French and English terms indifferently or even jointly. There are additional light cells on the map that denote rather high degrees of redundancy.

5 RESTRUCTURING AN INDEXING SYSTEM

As indexers have difficulties in maintaining efficient indexing systems for document flows, they need a tool to support their indexing work. This is the goal of the following algorithm.

Restructuring an index is considered as a interactive process, where the system helps the indexer to get a global vision of the index under construction. It makes suggestions as how to restructure the index to increase its efficiency in terms of information access. However, the indexer bears sole responsibility of choosing and applying the suggested restructurations.

5.1 Algorithm

Figure 6 gives an overall idea of how the indexer can interact with the tool which analyses the document base, based on the above mentioned quality indicators, makes a diagnosis and proposes some improvements to the indexer(s), who may accept or refuse them, depending on

- the semantic feasibility of the proposed transformations, that only the indexer can appreciate;
- the restructuring cost that is approximated by the number of documents that need to be re-classified;
- the expected gain of the transformation in terms of indexing efficiency, computed under the hypothesis of an optimal restructuring[8].

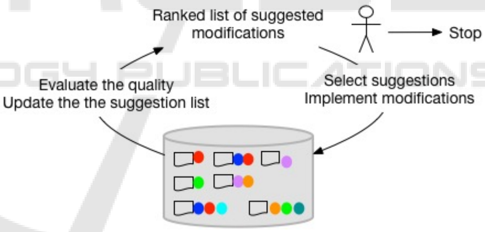
Ranked list of suggested modifications
Stop
Evaluate the quality
Update the the suggestion list
Select suggestions
Implement modifications

Figure 6: Interactive restructuring of an indexing system (here, the user is the indexer).

Algorithm 1

Entry : a system of categories and an indexed document base

Output : an updated system of categories and an updated indexed document base

User : person in charge of indexing the document base

Method : When the indexing diagnosis is triggered (by the user or on a regular base), do until the suggestion list is empty or the user stops.

1. Compute the quality measures (balance, cost, redundancy) on the indexed document base
2. Propose a ranked list of suggestions to the user (see Algorithm 2)

[8] *E.g.* splitting a big category in the optimal number of categories of equal size.

3. Wait for the user to implement one or several suggestions

Algorithm 2

Entry :

- a system of categories
- a document base indexed with those categories
- a set of quality measures (balance, cost, redundancy) associated to the indexed document base

Output : a ranked list of modifications

User : person in charge of indexing the document base

Method :

1. If the balance is low, add
   - splitting suggestions for the biggest categories
   - merging suggestions for some small categories
2. If the browsing cost is high (*i.e.* the granularity of the categories should be refined), add
   - splitting suggestions for the biggest categories
   - a complex restructuring suggestion (*e.g.* one may decide to add new categories that are orthogonal to the existing ones)
3. If the category access cost is high, add a complex suggestion (one option is to manually turn the system into a multi-categorial or hierarchical one)
4. Perform the redundancy analysis:
   - if there are generic categories that include many others, add suppressing suggestions for them
   - if there are couples of overlapping categories, add merging suggestions for them
   - if there are categories included in others, add
     – suppressing suggestions for the smaller ones
     – a complex restructuring suggestion (*e.g.* partition the big categories into smaller ones so as to form a hierarchy)
5. Rank the suggestion list according to
   - the internal complexity of the modification type (*e.g.* splitting one category is much simpler than restructuring the whole document base)
   - the expected impact of the modification in terms of balance, cost and redundancy
   - the restructuring effort needed, computed as the number of documents that must be reclassified
   - the internal priority of the categories to be modified (this depends on their age and activity rate[9]).

Two types of suggestions are made. The simple operations are supported by the system which propose the categories to merge or split, even if the indexers still have to decide how to reclassify their documents. The complex operations are left to the indexers: the system simply points out the need for restructuring.

The algorithm has several thresholds, which have been fixed on an intuitive basis so far. As future work, we plan to start with arbitrarily high thresholds (to maximize the number of suggestions and limit false negative) and tune them incrementally as the indexer uses the system and accepts/refuses the suggestions.

5.2 Index Analysis and Restructuring

This section shows the impact of the restructuring algorithm on some of the blogs of our corpus.

Correcting the Balance. `jeuxvideo3` (Table 2) is a typical dynamic mono-categorial blog. The number of categories is multiplied by 7 and the cost by 7.5 but it remains close to the optimum (reference cost).

Improving the balance takes priority as it declines significantly, from 0.86 to 0.76. The mean frequency of the 91 categories is 60 but five categories exceed 4 times this size. The system first suggests to split some of these big categories, starting with the one which contains 15% of the posts. Formally, to limit all categories to 4 times the average size, the indexer has to re-categorize 1240 posts and create 5 new categories. The balance should improve to 0.832. The total cost is expected to increase slightly (up to 153).

A second suggestion consists in merging the 31 categories with very few posts into one "miscellaneous" category. This requires reclassifying 118 posts but reduces the number of categories to 66, increases the balance to 0.890 and reduces the cost to 149.12.

Note that the indexer may reach the same result in a different way. However, if local improvements of the balance become too intricate, the only possible escape is to restructure the indexing system as a multi-categorial one. In that case, the restructuring cost is difficult to estimate.

Reducing the Access Cost. `technologie2` is a small multicategorial blog, reaching 243 posts after eight years of activity (Table 2).

The balance is good, constantly near or beyond 0.9 but the access cost is always more than twice the reference cost, with the querying cost accounting for almost 90% of the total cost. Not only is the number of basic categories high, but there are twice as many compound categories and the multi-annotation is not uniform (in 2015, 45% of the posts are associated to a single category, whereas 25% have 3 to 5).

The algorithm suggests first to reduce the number of categories and multi-categories. One simple proposal would be to delete the `domotique` category that is uninformative (it is the biggest category with $1/4^{th}$

[9] Large categories deserve attention if they are still very active and keep on growing but small categories that do not seem active any more are good candidates for merging.

Table 2: Evolution of the quality of Blog jeuxvideo3 (top) and Blog technologie2 (bottom).

Year	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
# posts	92	110	193	338	473	805	1494	2293	3686	5486
# categories	13	14	17	22	31	36	52	68	79	91
Access cost	20.1	21.8	28.4	37.4	46.3	58.4	80.7	101.7	125.7	151
Reference cost	19.2	21	27.8	36.76	43.50	56.74	77.30	95.77	121.42	148.13
Balance	0.86	0.84	0.84	0.84	0.81	0.81	0.82	0.79	0.77	0.76

Year	2007	2008	2009	2010	2011	2012	2013	2014	2015
# posts	45	79	94	100	109	158	205	232	243
# categories	17	22	23	23	24	27	33	38	38
# compound categories	23	42	48	51	56	69	86	98	101
Access cost	29.29	39.76	42.68	42.94	44.45	49.33	65.56	76.58	77.50
Optimal cost	13.42	17.78	19.39	20	20.88	25.14	28.64	30.46	31.18
Balance	0.931	0.935	0.938	0.937	0.933	0.941	0.937	0.900	0.896

of the blog posts, it includes several smaller categories and corresponds to the title/topic of the blog). It would only require re-classifying the posts which are not already associated to another category and it would improve both the cost and the balance.

6 DISCUSSION

This paper presents a method for helping indexers to control the quality of the document classification or indexing systems they provide to facilitate access to information for readers. On a corpus of blogs covering a decade, we observe that the indexing practices vary a lot and that the quality of the indexing systems in terms of information and cost of access for the readers often deteriorates with time.

Our algorithm helps indexers to get a global vision and control the quality of the indexing systems they build incrementally. It relies on quality measures that are used to audit an existing indexing system and makes suggestions as how to restructure it. Our simulation results show that those restructuring suggestions can actually improve a lot the quality of the indexing systems, sometimes with only a moderate number of posts to re-classify.

This algorithm has been designed as a generic tool that can be used in different contexts. The next step will be to integrate it in a blog management platform. It will have to be tuned for each specific blog based on the indexer's requirements or expectations.

REFERENCES

Battail, G. (1997). *Théorie de l'information*. Masson.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

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

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to Algorithms*. The MIT Press.

Funk, M. E. and 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., and Nazarenko, A. (2016). A French weblog corpus for new insights on blog post tagging. In Moreno Ortiz, A. and Pérez-Hernández, editors, *8th Int. Conf. on Corpus Linguistics*, volume 1, pages 144–158, Malaga, Spain.

Jan, J.-C., Chen, C.-M., and Huang, P.-H. (2016). Enhancement of digital reading performance by using a novel web-based collaborative reading annotation system with two quality annotation filtering mechanisms. *Int. Journal of Human-Computer Studies*, 86:81 – 93.

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

Manning, C. and Schütze, H. (1999). *Foundations of statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Mathet, Y., Widlöcher, A., Fort, K., François, C., Galibert, O., Grouin, C., Kahn, J., Rosset, S., and Zweigenbaum, P. (2012). Manual Corpus Annotation: Giving Meaning to the Evaluation Metrics. In *Int. Conf. on Computational Linguistics*, pages 809–818, Mumbaï, India.

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

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

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