Blog Annotation: From Corpus Analysis to Automatic Tag Suggestion

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

LIPN, Paris 13 University – Sorbonne Paris Cité & CNRS, France,
99, av. J.-B. Clément, F-93430, France
{garridomarquez,jgflores,fl,
adeline.nazarenko}@lipn.univ-paris13.fr

Abstract. Nowadays, blogs cover a large audience and they become part of mainstream media. Tags and categories are structural elements of a blog post intended to increase a blog's visibility and enhance navigation and searching. We suppose that those annotations are made on subjective grounds rather than in a systematic way. This paper presents a 11 million words corpus of blogs posts in French dedicated to the analysis of blog post tagging and categorization practices. We present experiences on automatic tag and category suggestion based on this corpus. Preliminary results show that around 27% of the overall tags can be predicted from lexical frequency analysis of blog posts. Furthermore, a first comparison experience with an existing tag suggestion tool shows that an important proportion of the tags used for blog description are not present in the blog post. Preliminary observations of annotations in time might suggest that taking into account diachronic information from previous post might improve the tag suggestion process.

Keywords: Annotation, blogs, tag suggestion, tagging, corpus of blogs, corpus analysis

1 Introduction

The rise of blogs followed that of the web at the end of the 90's. The blogosphere boomed in the early 2000s and it has been part of the mainstream media for more than a decade. Originally a blog was mostly a personal journal published on a website containing multiple entries. It is estimated that 152 millions blogs were active by the end of 2010 [2]. In the single Wordpress domain, 22.8 billion pages are viewed every month while 56 millions of new posts are published. Nowadays, the blogosphere and its hundreds of millions of blogs has become an essential mean for sharing information.

* 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.

The elemental unit of blogs is the post, a piece of content normally written by one single author. To ease the classification of blog posts, it is very common that bloggers annotate them with categories and/or tags. In the first case, the blog classify its contents into a predefined set of categories, each category corresponding to a group of posts that are somehow related. Category systems may be organized into taxonomies where posts can be classed in more than one category. A different way to mark a post is by adding keywords, called tags, that somehow summarize the topic. These tags come from an open vocabulary, and both their variety and their amount might grow as the blog post number increases. The evolution of the blog's subject might enrich the variety of these tags as well.

Annotating blog posts with tags makes easier the searching of content in the blog. It enhances navigation as well, by allowing to group posts of a particular subject or related content. They might increase blog's visibility in the web by letting web search engines index them with tags. On the other hand, adjoining tags or categories is mostly based on a distributed, subjective or any other arbitrary criteria. We suppose that the study of tagging and categorization practices could bring a better understanding of the semantic relations underlying between tags, categories and posts.

The paper is structured as following: In section 2 we present related work on blog post analysis and tag and category prediction. Section 3 presents an 11 million words corpus of blog posts in French on law, cooking and technology. Section 4 introduces our first experiences on tag and category prediction. Section 5 discusses the results while section 6 gives a hint of future perspectives for this work.

2 Previous works

Even if ”one of tagging's biggest appeals is its simplicity and ease of use” [1], we tend to believe that the resulting annotations are not systematic at all because they almost always depend on the user, as the example 1 from section 4.2 suggests.

We distinguish three types of approaches proposed to automatically identify possible tags for a blog post:

1. Predicting tags from a fixed set by using machine learning. In [5] Katakis and al. present a system for tag recommendation in social bookmarks. The system recommendations are meant to be particular for each author: it recommends the most popular tags present in the post and previously associated to the user.
2. Computing a topic description over the set of tags [10,6,4,7]. For example, in [10] Tsai uses topic modeling for mining the tags in blogs according to topics. Each tag is represented by a probability distribution over topics, and each topic represented by a probability distribution over terms for that topic. The technique is based on LDA for topic modeling and dimensionality reduction.

The most suitable terms for tagging can be identified by computing the topics.

3. Searching for tags on similar posts. The system AutoTag described in [8] estimates similarity between blog posts with information retrieval measures and selects the most similar posts to the one at hand. Then it extract a list of tags ranked by their frequency in the selection of posts. At the end, a filtering and reranking step boosts the score of tags previously used by the user, and then the best tags are proposed.

There are various developed tag suggestion tools, available as APIs or web services, to help bloggers to annotate their posts. They rely on post content and propose the most relevant extracted keywords as tags. *Zemanta*[1], *Yahoo! Content Analysis*[2] and *Open Calais*[3] can match the tags with entities in external descriptive resources. Others like *AlchemyAPI*[4] and *Thoth*[5] can make use of blog-level statistics or sophisticated natural language processing techniques. They are designed as independent tools, but provide plugins for major blog platforms. Some like *Climate tagger*[6] work for specific domain content such as documents abour climate.

Despite the existing annotation methods and tools, blog annotation often remains manual and unsystematic, which might hinder the usability of the blogs. With 10 years of hindsight, we can analyze blog annotation in the long term so as to understand how tag suggestion works and why the use of tag annotation tools isn't as widespread practice as one could expect from the huge amount of bloggers. This historical perspective, impossible ten years ago, allows a long term analysis, which we consider the most important value of the corpus.

3 Corpus analysis

Our corpus includes around 11 millions of words. It is composed of 20 blogs in French dealing with different topics. In order to analyze the annotation practices, we focus on blogs containing tag and category annotations. We also focus of textual parts only, leaving aside the images and video contents. The main topics included in the blogs are cooking, law, technology, and video games. From Table 1 we can observe high standard deviations on every description feature: number of authors, categories, tags and size. Some blogs have a lot of contributing authors, up to 143, while many of them have a single author (9) or less than 5 authors (13). The number of the posts as well as their size also vary significantly (from 184 to 6,585 posts per blog and from 55 to almost 1,300 words per post). It is worthy to notice that four of these blogs contain more than 1 million of words each.

[1] http://www.zemanta.com/

[2] https://developer.yahoo.com/contentanalysis/

[3] http://www.opencalais.com/

[4] http://www.alchemyapi.com/

[5] https://fr.wordpress.org/plugins/thoth-suggested-tags/

[6] http://www.climatetagger.net/

Table 1. Corpus description

Blog	**posts**	**authors**	**cats**	**tags**	**Kb**	**words**
jeuxvideo6	184	6	18	556	968	66,991
technologie2	243	1	38	40	1108	55,073
droit3	283	1	13	77	3704	366,816
technologie5	305	1	16	295	2112	177,034
technologie3	343	13	41	397	2120	193,160
technologie5	374	2	25	358	2816	317,551
cuisine3	474	1	50	243	2048	152,377
droit1	485	2	4	84	4736	466,702
cuisine1	514	1	60	460	2180	133,063
technologie4	573	1	12	321	2508	110,111
jeuxvideo5	1135	2	37	2467	5716	387,632
cuisine2	1166	1	26	695	10064	1,051,706
jeuxvideo1	1423	3	43	1772	9672	868,019
technologie1	1423	17	56	1231	6740	416,498
jeuxvideo4	1501	17	40	3146	9048	698,151
cuisine4	1721	1	25	265	9092	891,033
droit4	1752	1	15	0	14104	1,333,494
droit2	1769	143	48	741	10440	771,041
jeuxvideo2	2483	6	33	2978	17060	1,349,318
jeuxvideo3	6587	67	91	4650	31148	1,598,143
average	1236.9	14.35	34.55	1038.8	7369.2	570,195.65
std dev	1426.12	33.77	20.57	1292.08	7221.87	475,284.97
max	6587	143	91	4650	31148	1,598,143
min	184	1	4	0	968	55,073
total	24738	287	691	20776	147384	**11,403,913**

3.1 Annotation activity

Table 1 shows stats about the annotation activity. The average number of categories is 34 with a high standard deviation. Categories for a single blog range from 4 to 91. Tags for a single blog range from 40 to 4650. The standard deviation for tags is high as well. Table 2 reviews the annotation activity at a *per blog post* level. The mean number of categories per post ranges from 1 to 4.27.

Overall, each blog has its own annotation profile. Furthermore, the tagging activity might be more arbitrary than the category attribution, so one could wonder if a more consistent semantic annotation system is possible for the blog annotation activity by using a tag suggestion tools, such as those cited in Section 2. On the other han, figures suggest that categories are semantically more structured than tags.

We suppose that the tags arise from a wide variety of sources: the post content, the pool of existing tags, external resources (the web, another blog, search engines): *i.e.* new tags do not seem derived from the post content.

Table 2. Categories and tags per post

	Categories				Tags			
Blog	**mean**	**min**	**max**	σ	**mean**	**min**	**max**	σ
cuisine1	1	1	1	0	2.12	0	17	3.42
cuisine2	1	1	1	0	5.45	1	20	3.51
jeuxvideo1	3.41	1	11	1.87	4.95	0	19	1.69
technologie1	1.07	1	6	0.29	3.16	0	16	3.55
technologie5	2.31	1	8	1.22	4.20	0	24	4.34
droit1	1	1	1	0	2.41	0	6	1.31
jeuxvideo2	4.27	1	12	1.74	8.84	1	21	3.19
technologie2	1.88	1	5	0.96	0.79	0	5	1.09
jeuxvideo3	0.99	0	1	0.07	4.09	0	28	2.24
jeuxvideo4	2.22	1	3	0.71	6.07	0	45	3.92
droit2	1.72	1	31	2.81	3.19	0	19	2.82
cuisine3	1	1	1	0	5.20	1	14	1.88
technologie3	1.31	1	4	0.60	2.54	0	6	1.34
droit3	1.41	1	5	0.68	2.94	0	9	2.28
cuisine4	1	1	1	0	4.04	0	11	1.68
droit4	3.14	1	7	1.08	0	0	0	0
technologie4	1	1	1	0	3.13	0	13	2.1
jeuxvideo5	2.94	1	10	1.22	3.79	0	13	1.7
technologie5	4.18	1	12	2.03	6.72	0	18	3.17
jeuxvideo6	1.01	1	2	0.1	5.34	0	20	2.84

3.2 Evolution over time

Our corpus covers a period of 10 years of blogging, from 2005 to 2010. It allows to study the temporal evolution of the blog annotation practices. Let's consider the example of the *droit2*, which was launched in 2007 and is still active in 2016.

Figure 1 shows the distribution of the posts over 10 years (120 months) of the *droit2*. A peak of the blog post activity can be observed between Months 37 and 75. This blog has the highest number of authors (see Table 1), but an average number of categories (1.72) and tags (3.19) per post. Between 2008 and 2011, posting activity was very intense and by a high number of authors, which didn't imply higher than average tagging nor categorizing variety. The 143 authors of the *droit2* annotated with less tags and less categories per post than the single author of the significantly smaller *technologie5* blog.

Figure 2 shows three different tag profiles for blog *droit2*. Like most of the categories, they present irregular distributions. One could argue that it is related to the news and that *Contrefaçon*, for instance, appears and disappears in the annotations because of an underlying issue that becomes hot and then fades in the news. However, we observe that the histogram of *Contrefaçon* actually follows the activity in the blog (measured as the number of posts per month, on Figure 1). It therefore reflects a rather stable distribution. On the opposite, the histogram of *Avocats* shows a surprinsingly uneven distribution for a term like

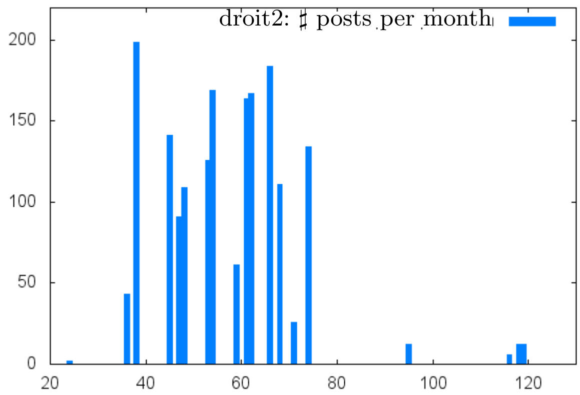


Fig. 1. Distribution of the *droit2* posts over a period of 100 months. X-axis is the date in months (from the 20^{th} to the 120^{th} month of the global blog corpus). Y-axis is the number of posts.

lawyers, which would be expected to be very common in a legal blog. And also, *Internet* has the same profile as *Avocats* at the same time, hinting to a possible correlation.

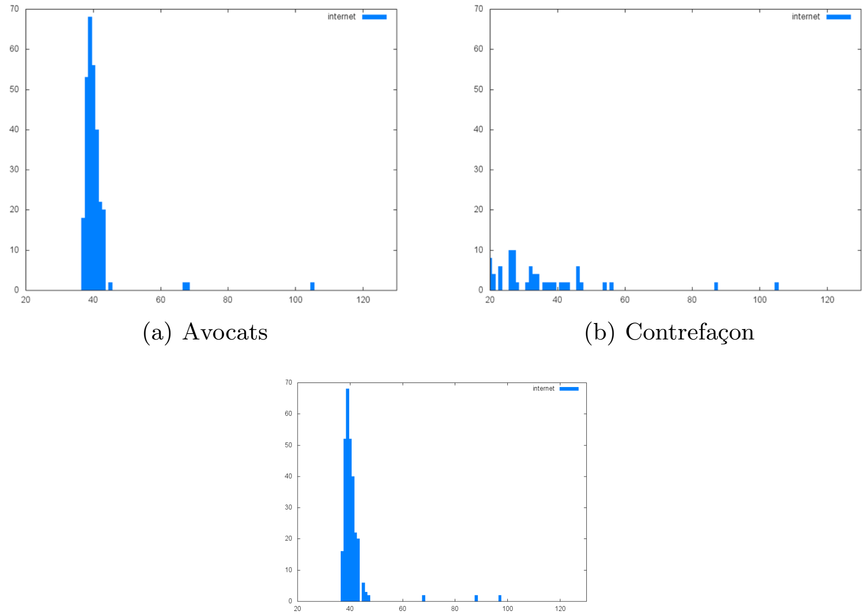
These observations suggests that the tagging and categorizing activity are not indexed to the post frequency over time. From these figures we suppose that the blog annotation activity is rather arbitrary and not very systematic, which limits the utility of tags and categories while searching information withing blogs. It is therefore important for bloggers to be assisted with tag and category suggestion tools.

4 Annotation strategies for blogs

4.1 Tagging from post contents: term frequency strategy

Tag prediction based on word frequency is a traditional approach for tag suggestion [1].We analyse a prediction strategy based on simple term frequency; *tf-idf*; and the combination of both, giving a higher weight to the tags present in the first two strategies. Ten tags were automatically generated and compared with author's hand made tags. Because of the variation on the tags per post, we consider the recall measure (R@10) as the most appropriate for this kind of evaluation.

Table 3 shows the results for frequency based strategies. Data in bold show the best and the worst precision and recall measures. All of them got their highest recall on *technologie3* blog, while the worst recall was for *droit1* blog. We interpret a high recall for a blog as a systematic use of the tags that are frequent in the content of blog posts. On the other hand, a low recall is a sign of outside the post lexical choices as tags. Both blogs show a systematic tagging



(a) Avocats

(b) Contrefaçon

(c) internet

Fig. 2. Profile of three tags over time on the *droit2*

approach: *technologie3* blog (whose main subject is technology, therefore physical objects) systematically uses tags from the blog posts, while in the *droit1* blog, authors systematically tag with words that do not occur (or with a very low frequency) in the blog posts. Our best strategy was able to find an average of 27% of author's tags.

4.2 Human annotations versus automatic tagging in blogs

As mentioned in Section 2.1, different platforms and services provide multipurpose meta-data extracted from unstructured text content. Specific tools have been developped for automatically suggesting tags to blog authors. In the case of IBM's Alchemy there is a wordpress plugin called AlchemyTagger, which suggests a set of tags based on a JSON or XML response of AlchemyAPI.

We annotated blog posts with tags suggested by AlchemyAPI, selected the top 10 suggestions according to the relevance score provided by this API and compared to author's tags. We will explain some of our observations with the following examples.

The first exampleis a brief article about the legal authorities involved in the creation of a made-in-europe mark for products. It was tagged only with one tag by its author (*"Avocats; John Doe"*). This tag is composed of two terms *Avocats (Advocates/Lawyers)* and the name of a person *John Doe*. The first term is very

Table 3. Term frequency based tagging strategy

	TF			TFIDF			Mix		
Blog	**P@10**	**R@10**	**F@10**	**P@10**	**R@10**	**F@10**	**P@10**	**R@10**	**F@10**
cuisine1	0.14	0.25	0.18	**0.18**	0.32	0.23	0.13	0.35	0.19
cuisine2	0.14	0.28	0.19	0.14	0.28	0.19	0.11	0.32	0.17
jeuxvideo1	0.10	0.21	0.14	0.10	0.21	0.14	0.08	0.26	0.13
technologie1	**0.17**	0.29	0.21	0.16	0.27	0.20	**0.13**	0.31	0.19
technologie5	0.11	0.24	0.15	0.12	0.28	0.17	0.09	0.30	0.14
droit1	**0.02**	**0.09**	**0.04**	**0.02**	**0.10**	**0.04**	**0.02**	**0.11**	**0.03**
jeuxvideo2	0.13	0.16	0.15	0.13	0.15	0.14	0.11	0.20	0.14
technologie2	0.07	0.43	0.11	0.06	0.37	0.10	0.06	0.45	0.10
jeuxvideo3	0.07	0.18	0.10	0.07	0.17	0.09	0.06	0.2	0.09
jeuxvideo4	0.10	0.20	0.13	0.10	0.19	0.13	0.08	0.23	0.12
droit2	0.07	0.18	0.10	0.07	0.15	0.09	0.06	0.20	0.09
cuisine3	0.15	0.29	0.20	0.15	0.29	0.20	0.12	0.33	0.17
technologie3	0.14	**0.54**	**0.22**	0.15	**0.59**	**0.24**	0.11	**0.62**	**0.19**
droit3	0.04	0.10	0.05	0.04	0.10	0.05	0.03	0.12	0.05
cuisine4	0.11	0.28	0.16	0.13	0.33	0.18	0.09	0.34	0.14
technologie4	0.06	0.15	0.08	0.04	0.11	0.06	0.04	0.16	0.07
jeuxvideo5	0.09	0.24	0.13	0.11	0.30	0.16	0.08	0.31	0.13
technologie5	0.11	0.17	0.13	0.10	0.15	0.12	0.09	0.19	0.12
jeuxvideo6	0.09	0.16	0.11	0.10	0.19	0.13	0.08	0.20	0.11
average	0.10	0.23	0.13	0.1	0.24	0.14	0.08	**0.27***	0.12
max	0.17	0.54	0.22	0.18	0.59	0.24	0.13	**0.62**	0.19
min	0.02	0.09	0.04	0.02	0.10	0.04	0.02	**0.11**	0.03
std dev	0.04	0.11	0.05	0.05	**0.12**	0.06	0.03	0.12	0.05

generic. It is related to the general topic of the blog, which is law, and not to the specific topic of the post. The second term, *John Doe* does not correspond to the author, neither to a person mentioned in the text: it belongs to one of the main owners of the blog. We would like to remark that none of these terms in the tag come from the content of the post.

The 10 tags top-ranked by AlchemyAPI for this post are: *"made in", "made in europe", "marquage d'origine", "Direction Générale", "made in france", "régime européen uniforme", "Europe", "position officielle", "Union française", "Après l'UFIH"*[7]. This proposed tag set proposed was extracted from the actual content of the post.

With 4 tags, the second post example, about the payment in advance of legal proceedings in france by buying revenue stamps online, is richer in author's tags. These tags are *"Action en justice", "John Doe", "Doe avocats", "Procédure"*[8]. Only one tag out of the four comes from the content of the post

[7] Translation:*"Made in", "Made in Europe", "origin marking", "General Management", "Made in France", "uniform European system," "Europe," "official position", "French Union", "after UFIH".*

[8] "Legal Action", "John Doe", "Doe lawyers", "Proceeding".

itself: "Procédure". "John Doe", "Doe avocats" might actually be intended to improve the visibility of the author and his firm by appearing as meta-data that could be indexed by internet searching engines. The set of tags proposed by Alchemy included *"paiement d'un timbre", "Le paiement s'effectue", "Une fois l'achat", "adresse mail", "également possible", "carte bancaire"*[9]. In this case, it is not possible to say that they summarize the topic of the post, but they are elements related to specific parts of the content.

In these examples, the authors of the blog *droit2* don't use tags taken from the post content. Table 3 confirms this observation: the blog *droit2* has a very low recall of tags coming from the post text. The results for both the frequency strategy and the much more complex analysis tool were particularly low, not that they suggest bad quality tags, but because they propose tags extracted from the text of the post, a policy that contrasts with the author tagging. We also observed that for a certain period of two years almost every post in this blog was tagged with the name "John Doe"; this seems to be a policy to increase the visibility of a certain lawyer's firm. Overall, if we consider the direct match with users' tags, the precision and recall of Alchemy suggestions are close to 0. An in-depth evaluation would be needed to on the one hand to evaluate the quality of author tags and on the other hand the existing tag suggestion tools.

4.3 Representing categories: Words or keyword tags

Unlike tags, categories are an established set, well defined before a post is written. Categories can be created at the moment, but in general they come from a closed vocabulary that classing the posts by topics. Categories frequently define taxonomies which can be seen as a very primitive semantic model for the blog. When a blog has a significant amount of examples to represent well enough its current categories, it would be possible to train a supervised classifier to predict the category of a new post.

We trained four popular supervised classifiers relying on vector machines (SVM) with a linear kernel, Multinomial naive bayes (NB), K nearest neighbor with K=5 (5NN), and Random Forest using 25 weak decision tree classifiers (RF). We measured accuracy with a 10-fold cross validation for every blog in the corpus.

Posts were represented as a bag of words with two different sets of features. The first space of features of the training set contains all the post's words without stop words. The second space of training set features contains the all the post's tags. Table 4 presents the results per blog of this experiment for all the classifiers and both feature sets.

The tags used as features seem to have a similar or even better categorization power than the blog post words. This makes sense because at the end tags are supposed to be a good feature set defined by the author to summarize the content of a post, while categories are meant to focus on the topic. Therefore we might

[9] "Payment of a stamp", "Payment done", "Once the purchase", "email address", "also possible", "credit card"

think that categories hold certain semantic relation with tags, so that categories can be represented by tags. It is worth mentioning the blog *droit4* was excluded from this experience because its posts.

Table 4. Supervised learning for post categorization based on words and tags

	Words				Tags			
blog	**SVM**	**NB**	**5NN**	**RF**	**SVM**	**NB**	**5NN**	**RF**
cuisine1	0.71	0.29	0.60	0.61	0.50	0.55	0.17	0.50
cuisine2	0.73	0.27	0.62	0.67	0.79	0.72	0.62	0.80
jeuxvideo1	0.64	0.57	0.64	0.64	0.75	0.75	0.73	0.78
technologie1	0.61	0.35	0.59	0.49	0.73	0.69	0.64	0.72
technologie5	0.96	0.96	0.94	0.96	0.96	0.94	0.96	0.98
droit1	0.96	0.63	0.76	0.95	0.79	0.75	0.76	0.85
jeuxvideo2	0.94	0.54	0.64	0.92	0.93	0.83	0.59	0.92
technologie2	0.52	0.28	0.55	0.44	0.67	0.63	0.63	0.62
jeuxvideo3	0.52	0.21	0.31	0.51	0.58	0.45	0.42	0.63
jeuxvideo4	0.65	0.46	0.59	0.60	0.78	0.72	0.67	0.74
droit2	0.88	0.51	0.55	0.90	0.81	0.59	0.54	0.91
cuisine3	0.52	0.29	0.54	0.54	0.60	0.44	0.42	0.64
technologie3	0.35	0.23	0.49	0.29	0.46	0.36	0.34	0.39
droit3	0.45	0.45	0.59	0.45	0.57	0.48	0.38	0.67
cuisine4	0.83	0.46	0.72	0.59	0.76	0.73	0.64	0.76
droit4	0.81	0.77	0.79	0.77				
technologie4	0.56	0.38	0.23	0.54	0.76	0.72	0.70	0.76
jeuxvideo5	0.58	0.37	0.49	0.53	0.32	0.31	0.17	0.30
technologie5	0.59	0.59	0.63	0.60	0.64	0.63	0.44	0.64
jeuxvideo6	0.42	0.16	0.63	0.26	0.55	0.55	0.39	0.55
average	0.66	0.44	0.59	0.61	0.68	0.62	0.54	**0.69**

We would like to remark that current systems for annotating blogs do not propose categories from the predefined taxonomy of categories the blog holds. We hypothesize this is mostly due to their not taking historical information of the blog into account to do so. only the content of the post to be tagged.

5 Discussion

We consider that a corpus made of ten years of blogging has unique features for the study of annotation practices. First, posts have a chronological order and a significant size. Second, blogs have a collective authorship, which in practical terms means a group of persons trying to tag and classify (that is, to apply an implicit semantic type system) to text fragments. And third, because 10 years of blogging can produce a reasonable amount of data to study the evolution of this relation between a text fragment and a keyword (the tag) or an implicit taxonomy (the category). Last but not least, to our knowledge there's no other corpus of French blog post with the size nor the topic variety like the one we are presenting in this paper. This two-level structure (tags, somehow arbitrary, and categories, considered as more stable from a semantic point of view) provides a reference to evaluate automatic annotation. In addition to that, it allows the experimental evaluation of hand annotation quality. The comparison of Alchemy's with author's tags shows an empty intersection but, from a subjective perspective, the annotation of Alchemy looks better than that of authors. The TFIDF based tag suggestion system suggests that the posts word frequency plays a role on this implicit and somehow arbitrary semantic type system that authors might have in mind when tagging and categorizing blog posts. In at least one blog from our blog set (*technologie3*), the authors have a consistent policy of tagging with words from the title or the blog posts. In another blog (*droit1*), the authors have adopted the opposite policy: they tag with terms that do not occur in the blog posts. Further experiences on tag quality might give a better insight on the quality of authors' tag annotations and categories. Furthermore, the chronological order of posts and tags might lead to interesting lexical analysis. For instance, how is it that *lawyers* and *internet* have exactly similar distributions over time? Have they been supported by a common topic? Once in use, even out of order tags are not changed and they generally retain at least some low level activity, obscuring the landscape with non informative noise. At the same time, categories seem to evolve slowly, but they need to be divided when their growth is too fast. Furthermore, some theme can be appreciated differently some time after the annotation has been made. For instance, considerations evoking that Great Britain might leave Europe written in 2005 are still interesting now, but may not have been adequately annotated at this time for 2016 readers. Backward annotation might be an important feature to propose to bloggers.

6 Conclusion and Perspectives

In this paper we presented an 11 million word corpus of French blogs which provides a valuable testbed to analyze blog annotation practices. Blogs subject's include cooking, technology, law and video games. The blog corpus gathers ten years of posts with authors, tags and categories in a chronological order. Future experimental perspectives for this resource will focus on the semantic analysis of bloggers annotation practices and its evolution over time. The chronological study of the annotation practice suggests that considering the life span of tags and categories enriches their relations and opens the perspective of a re-annotation process. We also presented a first experiment on tag suggestion for blog posts. It is based on term frequency, the method chosen for the first experiments in English language ten years ago. Our motivation is that a corpus gathering ten years of authors tags and categories on a chronologicaly ordered text covering a wide variety of subjects is a unique experimental platform for the study of annotation practices. The comparison of the actual tags with those proposed by the analyzed strategies in section 4.1 show us that, in order to offer a better annotation system, we should not only rely on the content of the posts.

Based on section 3, we think that the historical archive of posts is also a good source of knowledge to be considered and analyzed to improve the prediction of tags. External knowledge resources could also be exploited to enrich this systems. In future works, we plan to implement a systematic and semantically consistent method for tagging and categorizing blog posts. This would imply an evaluation on author's tags and category quality to answer the underlying question of our work: Are tags arbitrary or systematic? Can authors produce solid categories and tags for their posts in an objective way? Our results suggest that in 27% of the cases a very simple TF-IDF tag suggestion system can extract author's tags from frequent terms contained in the post. We plan to improve this prediction measures and to exploit external semantic sources in order to propose appropriate and consistent tags for blog posts.

References

1. Brooks, C.H., Montanez, N.: Improved Annotation of the Blogopshere via Autotagging and Hierarchical Clustering. Proceedings of the 15th international conference on World Wide Web (WWW 06) pp. 625–632 (2006)
2. Chapman, C.: A brief history of blogging. `http://www.webdesignerdepot.com/2011/03/a-brief-history-of-blogging/` (2011), [Marketing, Web Design, WordPress Mar 14, 2011]
3. Chen, Y., Tsai, F.S., Chan, K.L.: Machine learning techniques for business blog search and mining. Expert Systems with Applications 35(3), 581–590 (2008)
4. Christidis, K., Mentzas, G., Apostolou, D.: Using latent topics to enhance search and recommendation in Enterprise Social Software. Expert Systems with Applications 39(10), 9297–9307 (2012)
5. Katakis, I., Tsoumakas, G., Vlahavas, I.: Multilabel Text Classification for Automated Tag Suggestion. Proceedings of the ECMLPKDD 2008 Discovery Challenge (2008) 9(3), 1–9 (2008)
6. Li, F., He, T., Tu, X., Hu, X.: Incorporating word correlation into tag-topic model for semantic knowledge acquisition. 21st ACM International Conference on Information and Knowledge Management, CIKM 2012 pp. 1622–1626 (2012)
7. Li, Z., Xu, C.: Tag-based top-N recommendation using a pairwise topic model. Proceedings of the 2013 IEEE 14th International Conference on Information Reuse and Integration, IEEE IRI 2013 pp. 30–37 (2013)
8. Mishne, G.: AutoTag: A Collaborative Approach to Automated Tag Assignment for Weblog Posts. Proceedings of the 15th international conference on World Wide Web (WWW 06) (2006)
9. Sood, S.C., Owsley, S.H., Hammond, K.J., Birnbaum, L.: TagAssist: Automatic Tag Suggestion for Blog Posts. Proceedings of the 1st International Conference on Weblogs and Social Media (ICWSM 2007) (2007)
10. Tsai, F.S.: A tag-topic model for blog mining. Expert Systems with Applications 38(5), 5330–5335 (2011), `http://dx.doi.org/10.1016/j.eswa.2010.10.025`