How to measure the consistency of the tagging of scientific papers?

Boris Veytsman
bveytsman@chanzuckerberg.com
Chan Zuckerberg Initiative

ABSTRACT

A collection of scientific papers is usually accompanied by tags (keywords, topics, concepts etc.), associated with each paper. Sometimes these tags are human-generated, sometimes they are machine-generated. The evaluation of the tagging quality is an important problem. We propose a simple metrics of tagging consistency for scientific papers: whether these tags are predictive of citations. Since the authors tend to cite papers about the topics close to those of their publications, a consistent tagging should be able to predict citations. We present an algorithm to calculate consistency, and show experiments with human- and machine-generated tags. We show that the addition of machine-generated tags to the manual ones can enhance tagging consistency. We further introduce cross-consistency metrics, the ability to predict citation links between papers tagged by different taggers, e.g. humans and computers. Cross-consistency metrics can be used to evaluate tagging quality of a tagger when the amount of labeled data by the known good tagger is limited.

CCS CONCEPTS

• **Information systems → Document topic models**; *Digital libraries and archives*; *Similarity measures*; *Relevance assessment*.

KEYWORDS

topic modeling, tagging, tagging evaluation

ACM Reference Format:
Boris Veytsman. 2019. How to measure the consistency of the tagging of scientific papers?. In *Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL '19).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Boris Feldman, a formidable Principal University Librarian I've met in 1980s, loved to say that a scientific library without an index is a huge pile of used paper. Today he would probably say that such library is a pile of old magnetic media.

To make such index we add tags to the publications: concepts, topics, keywords, etc. This can be done manually or, as it is happens more and more often today, using machine learning methods [9]. While comparing tagging methods it is important to have evaluation metrics. One should be able to determine whether the given tagging is "good" or "bad". It is customary to test the machine produced tags by comparing them with the tags made by humans on a "golden set" of manually tagged papers. There are, however, certain problems with this approach. First, human tagging is expensive—even more so for scientific papers, where human taggers should have a specialized training just to understand what the papers are about. Second, even the best human taggers' results are inconsistent [8]. The situation is exacerbated when the tagging dictionary is sufficiently large. For example, the popular US National Library of Medicine database of Medical Subject Headings (MeSH, https://www.nlm.nih.gov/mesh/) has about 30 000 entries. A superset of MeSH, Unified Medical Language System (UMLS, https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html) contains a staggering amount of 3 822 832 distinct concepts. It is doubtful a human can do a good job choosing the right tags from such a huge collection. A domain expert uses for tagging a subsystem of the dictionary, covering their area of expertise. This presents obvious difficulties for tagging papers about multidisciplinary research. These papers may require a combination of the efforts of several highly qualified taggers.

Another problem is the evaluation of tag augmentation. Suppose we have papers tagged by humans, and we want to add machine-generated tags, for example, to improve searches in the collection. Do the new tags actually add to the quality of the library or subtract from it? How can we answer this question if the new tags tags are by definition different from those produced by humans?

This suggests a need for a tagging quality metrics which is not based on the direct comparison with the human generated tags. This paper proposes such a metrics.

The idea for this metrics is inspired by the works on graph embeddings [4, 6]. In these works one labels graph nodes and compares different sets of labels to select the best one. The usual comparison criterion is whether the labels can predict graph edges: nodes connected by an edge should have similar labels, while nodes not connected by an edge should have dissimilar labels. To use this approach for evaluation of tagging we need to represent the tags as labels, and the papers as nodes on a graph. A natural choice for scientific publications is the citation graph: and edge from paper A to paper B means that paper A cites paper B. This leads to the following assumptions:

(1) Scientific papers cited by the given paper A are more similar to A than the non cited papers.

JCDL '19, June 02–06, 2019, Urbana-Champaign, IN

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

(2) A good tagging system must reflect this, giving cited papers the tags similar to those of the citing paper.

In other words, a good set of tags must be able to predict citations. We will call this property a *consistency*: a good tagger consistently gives similar tags to papers cited by the given one.

It is worth stressing that consistency is just one component of the taggig quality. If a tagger consistently uses keyword *library* instead of keyword *bread* [1], this measure would give it high marks, despite tags being obviously wrong. A way to overcome this deficiency is to calculate *cross-consistency* with a known "good" tagger. For example, we may tag some papers manually, and some papers using machine generated tags. Then we predict citation links between these papers. The success of the prediction measures the similarity between these taggers. One of the applications of cross-consistency is the expansion of the number of labeled papers for evaluation of machine-based taggers. We can create a set of manually tagged papers, then generate tags for the papers cited by those in the set using a machine-based tagger. Since a typical paper cites many publications, this approach significantly expands the quantity of data available for training and testing.

To create a metrics based on these ideas one should note that citation links strongly depend on the time the candidate for citation is published. Even a very relevant paper may not be cited if it is too old or too new. In the first case the citing authors may prefer a newer paper on the same topic. In the second case they may overlook the most recent publications. Therefore we formalize our assumptions in the following way:

> A consistent tagging system should be able to predict citation links from a given paper to a set of simultaneously published papers.

The rest of the paper is organized as follows. In Section 2 we discuss the algorithm to calculate the consistency of the given tagging system. Experiments with this measure are discussed in Section 3. In Section 4 we present the conclusions.

2 ALGORITHM

The algorithm to calculate consistency metrics is Algorithm 1. We randomly select n seed papers. For each seed paper we take up to k random papers from its reference list, and label them with the label $y = 1$. These are our *cited papers*, or positive samples. For each cited paper we randomly choose k papers from the corpus with the same publication date as the cited paper (more precise, papers with the publication dates within g time interval, where g is the chosen date granularity). These are our *negative samples* (Figure 1). We label the chosen negative samples with the label $y = 0$. As the result we have a vector of labels y_i of size $k + km$, with k ones and km zeros. We tag the seed paper, positive and negative samples, and calculate the number of overlapping tags between the seed paper, and each of the samples. This gives us $k + km$ overlap numbers t_i for each seed paper. We solve the classification problem $\mathbf{y} \sim \mathbf{t}$ and calculate its ROC curve [3]. The area under curve (AUC) for the solution reflects the consistency of the tagging.

Algorithm 1 can be used for calculation of both consistency and cross-consistency of taggers. In the latter case we just choose different sources of tags for seed papers and samples.

Algorithm 1 Calculation of consistency measure for the given tagging system

Input: Parameters: n randomly selected seed papers, number k of randomly selected cited papers per seed paper, number m of negative samples per cited paper, granularity g of publication dates

1: **for all** seed papers $\mathfrak{s}$ **do**
2: Select min(k, bibliography length) random papers from the reference list of $\mathfrak{s}$ (positive samples). Label them as $y_i = 1$
3: **for all** cited papers $\mathfrak{c}$ **do**
4: Select m random papers with the publication date within interval g of the publication date of $\mathfrak{c}$ (negative samples). Label them as $y_i = 0$.
5: **end for**
6: Tag the seed paper, positive and negative samples.
7: **for all** positive and negative samples $\mathfrak{p}$ **do**
8: Calculate the number t_i of overlapping tags between the seed paper $\mathfrak{s}$ and the sample $\mathfrak{p}$
9: **end for**
10: Calculate AUC for the classification problem $\mathbf{y} \sim \mathbf{t}$
11: **end for**
12: **return** The set of AUCs. ▷ *The average AUC is the consistency measure, while the variation provides the error estimate*

Table 1: Tags coverage by different sources. The last column shows the fraction of papers with at least one tag.

Tagging source	Tags per paper		Coverage
	Mean	Median	
MANUAL	5.96	6	0.73
NEJI	12.22	4	0.55
DNORM	0.85	0	0.35
GNAT	0.18	0	0.08
MANUAL + NEJI	17.30	10	0.77
MANUAL + DNORM	6.74	6	0.74
MANUAL + GNAT	6.14	6	0.73
MANUAL + NEJI + DNORM + GNAT	18.26	11	0.77
NEJI + DNORM + GNAT	13.25	5	0.60

3 EXPERIMENTS

For experiments we used random papers extracted from the PubMed database (https://www.ncbi.nlm.nih.gov/pubmed/). These papers have MeSH tags attached manually by the human taggers. We added to them additional tags by processing papers' titles and abstracts using several packages. Gene names were identified using GNAT [5], diseases were identified using DNORM [7], and additional UMLS concepts were mapped using NEJI [2]. The coverage of papers by different tagging sources is shown in Table 1.

The number of seed papers n in Algorithm 1 was chosen to be $n = 100$. Based on the preliminary experiments we chose the following hyperparameters which produced a good convergence of the metrics: the number of cited papers per see paper $k = 10$, the

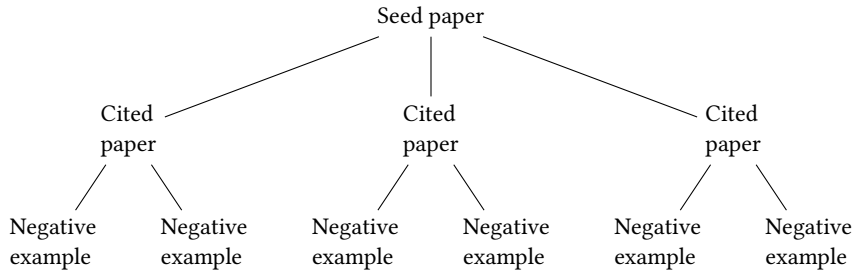

Figure 1: A seed paper, its cited papers and negative examples (here $k = 3$, $m = 2$).

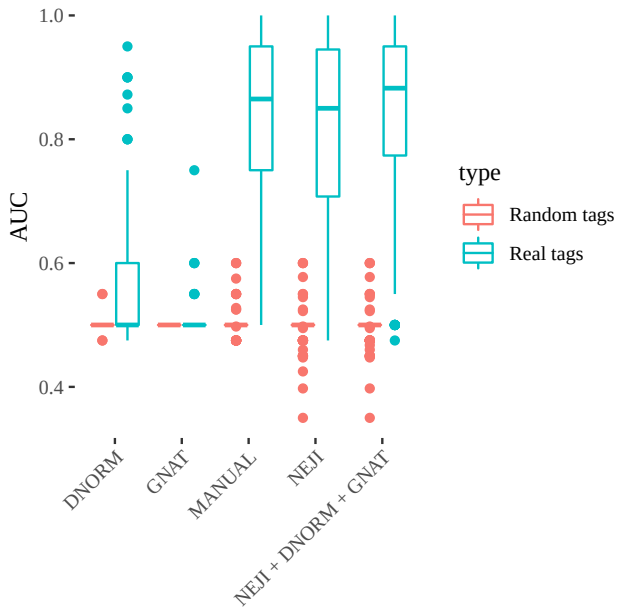

Figure 2: Consistency measure for tagging sources

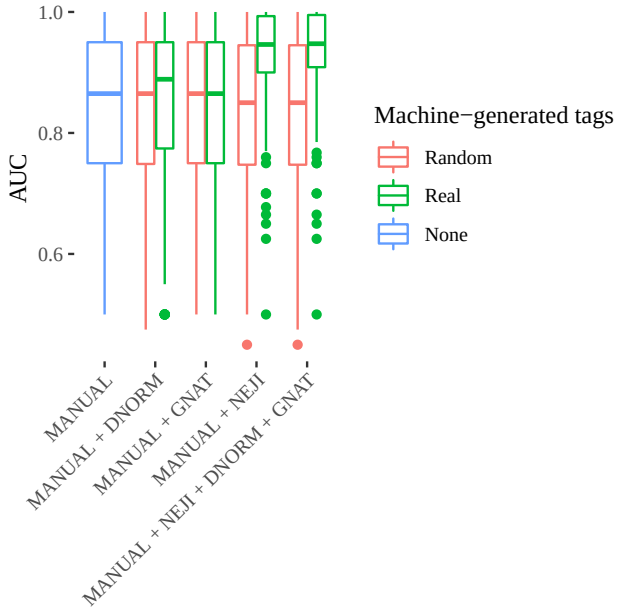

Figure 3: Consistency measure for adding machine-generated tags to manually created ones

number of negative samples per cited paper $m = 2$. The date granularity chosen was one month: two papers were considered having the same publication date if their year and month of publication coincided.

The baseline to compare the results against was constructed by randomizing the tag sets. For each paper (seeds, references, and negative samples) we constructed the set of tags from all the sources. Then we randomly shuffled the papers, so each paper got a set of tags from some other paper in the sample. We expect the average AUC for this random tag set to be 0.5, reflecting the lack of discrimination between positive and negative samples.

On Figure 2 we show the consistency measure for the tagging sources: DNORM, GNAT, MANUAL, NEJI as well as the combined automatic taggers (NEJI + DNORM + GNAT).

Adding machine-generated tags to the manual ones is explored on Figure 3. Here we take manually created tags and add machine-generated ones from different sources, again using random tags as a baseline.

On Figure 4 we show cross-consistency between the manual tags and NEJI-generated ones (since GNAT and DNORM are used to predict only specific concepts like genes and diseases, they are omitted from the experiment). We used one source for tagging seed papers, and another source for tagging samples (cited papers and negative samples).

4 DISCUSSION AND CONCLUSIONS

First, there is clear difference between the consistency of the randomly generated tags and the real ones (Figure 2). As expected, the consistency of the random tags is concentrated at AUC = 0.5, with some outliers both above and below this value. In contrast, the consistency of the real tags is almost always above AUC = 0.5. An exception is tagging sources of low coverage like GNAT (see Table 1), where consistency is close to 0.5. Obviously when the coverage is low, most positive and negative samples have zero overlap with their seed papers, which lowers AUC. Unexpectedly, the

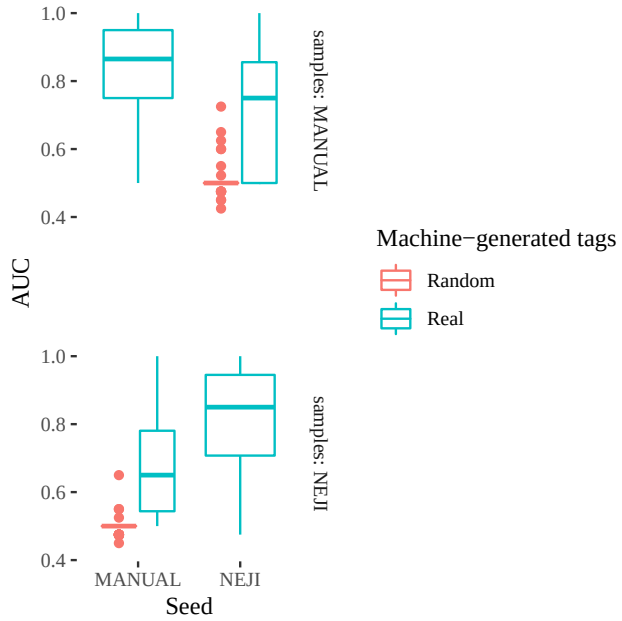

Figure 4: Cross consistency between manual tags and NEJI generated ones. X axis shows the source for the seed papers, Y axes shows the source for samples

consistency of high coverage machine generated sources like NEJI is on par with the human tags.

Tags augmentation is explored on Figure 3. As expected, adding random tags to the manually generated ones does not noticeably change the consistency of the result. However, adding "real" machine generated tags improves our metrics. This is an evidence that the measure itself is reasonable.

The cross-consistency between manual tags and machine-generated ones is shown on Figure 4. Here we used different sources for seed papers and for samples. While cross-consistency is lower than the internal consistency of each tagger, is still is significantly higher for the real tags than for the random ones.

In conclusion, a simple metrics tagging consistency: whether it is predictive of citations,—seems to be informative about the tagging process and can be used, along with other measures, to assess and evaluate it. The cross-consistency between different taggers can be used to estimate their similarity, especially when some taggers (e.g. manual tagging) are too expensive to run on a large set of papers.

ACKNOWLEDGMENTS

The author is grateful to Sunil Mohan, Dewey Murdick, Shankar Vembu, Ivana Williams and Liu Yang for their criticism and help.

REFERENCES

[1] Jorge Luis Borges. 1944. *Ficciones.* Editorial Sur, Buenos Aires.
[2] David Campos, Sérgio Matos, and José Luís Oliveira. 2013. A modular framework for biomedical concept recognition. *BMC Bioinformatics* 14, 1 (24 Sept. 2013), 281. https://doi.org/10.1186/1471-2105-14-281
[3] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27, 8 (2006), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010 ROC Analysis in Pattern Recognition.
[4] Aditya Grover and Jure Leskovec. 2016. Node2Vec: Scalable Feature Learning for Networks. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16).* ACM, New York, NY, USA, 855–864. https://doi.org/10.1145/2939672.2939754
[5] Jörg Hakenberg, Martin Gerner, Maximilian Haeussler, Illés Solt, Conrad Plake, Michael Schroeder, Graciela Gonzalez, Goran Nenadic, and Casey M Bergman. 2011. The GNAT library for local and remote gene mention normalization. *Bioinformatics (Oxford, England)* 27, 19 (October 2011), 2769–2771. https://doi.org/10.1093/bioinformatics/btr455
[6] William L. Hamilton, Rex Ying, Jure Lescovec, and Rok Sosic. 2018. Representation Learning on Networks. WWW-18 Tutorial. http://snap.stanford.edu/proj/embeddings-www/
[7] Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. DNorm: Disease name normalization with pairwise learning to rank. *Bioinformatics* 29, 22 (2013), 2909–2917. https://doi.org/10.1093/bioinformatics/btt474
[8] Christopher D. Manning. 2011. Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?. In *Computational Linguistics and Intelligent Text Processing*, Alexander F. Gelbukh (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 171–189.
[9] Shikhar Murty, Patrick Verga, Luke Vilnis, Irena Radovanovic, and Andrew McCallum. 2018. Hierarchical Losses and New Resources for Fine-grained Entity Typing and Linking. In *The 56th Annual Meeting of the Association for Computational Linguistics (ACL).* Melbourne, Australia.