Crowdsourcing in Article Evaluation

Isabella Peters
Heinrich-Heine-University
Universitätsstr.1
40225 Düsseldorf
+49 211 8 81 08 03
isabella.peters@uni-duesseldorf.de

Stefanie Haustein
Forschungszentrum Jülich
Central Library
52425 Jülich
+49 2461 61 61 98
s.haustein@fz-juelich.de

Jens Terliesner
Heinrich-Heine-University
Universitätsstr.1
40225 Düsseldorf
+49 211 8 81 08 03
jens.terliesner@uni-duesseldorf.de

ABSTRACT

Qualitative journal evaluation makes use of cumulated content descriptions of single articles. These can either be represented by author-generated keywords, professionally indexed subject headings, automatically extracted terms or by reader-generated tags as used in social bookmarking systems. It is assumed that particularly the users' view on article content differs significantly from the authors' or indexers' perspectives. To verify this assumption, title and abstract terms, author keywords, Inspec subject headings, KeyWords Plus^{TM} and tags are compared by calculating the overlap between the respective datasets. Our approach includes extensive term preprocessing (i.e. stemming, spelling unifications) to gain a homogeneous term collection. When term overlap is calculated for every single document of the dataset, similarity values are low. Thus, the presented study confirms the assumption, that the different types of keywords each reflect a different perspective of the articles' contents and that tags (cumulated across articles) can be used in journal evaluation to represent a reader-specific view on published content.

Categories and Subject Descriptors

H. 3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *abstracting methods, dictionaries, indexing methods, linguistic processing, thesauruses.*

General Terms

Measurement, Experimentation, Languages, Theory.

Keywords

Crowdsourcing, journal evaluation, tagging systems, tagging, folksonomies, article evaluation, subject headings, content analysis, metadata comparison.

1. INTRODUCTION

The traditional way of evaluating a journal's impact on the scientific community is citation analysis based on the number of citations articles of a particular journal gain. As pointed out in [2] this approach disregards various other aspects contributing to a journal's standing. Therefore, [3] introduce the analysis of usage data from social bookmarking platforms as a crowdsourced alternative to measure journal perception, particularly reader perception of journal contents. Similar projects explore parallel paths and apply social bookmarks to measure author impact[1]. In this study bookmarks of journal articles on the STM-specialized platforms[2] CiteULike, Connotea and BibSonomy are used to evaluate usage of journals and perception of articles.

2. MOTIVATION AND APPROACH

The state of the art of research on comparing folksonomies with other metadata demonstrates the high interest in this topic, while the conducted studies all arrive – more or less – at the same conclusions (amongst others: [5], [7], [8]). There are few overlaps of tags and professionally created metadata; most matches can be found when comparing tags and title terms. [5] suspect that this low overlap of tags and professional metadata is due to the indexing methods' different goals. Tagging users seem to have other demands than professional indexers, who want to index and cover all topics of a document using controlled vocabularies. Users seem to seek out the subject they are interested in and add a tag rather than represent the document completely. These findings confirm our hypothesis that users perceive articles and journals differently from intermediaries and that tags provide a basis for content description and qualitative journal evaluation. Tags are here considered as crowdsourced alternative for professionally created metadata which directly reflect the users' collective intelligence in describing article contents and along with it in unfolding journals' topical foci and research areas.

For the aforementioned analyses (except for [6], [1]) the researchers often do not differentiate between the collection of tags for all documents of a web 2.0 service, the folksonomy, and the docsonomy, which only consists of the tags for a single document. Moreover, a processing of tags and other metadata terms is regularly missing or not explicitly reported in the papers (except for [1]). Calculating overlaps and matches with unprocessed terms and tags as well as simply relying on one-to-one matches on the string levels can lead to erroneous values and along with it to invalid conclusions about the nature of folksonomies and metadata. Because of the uncontrolled nature of tags and the given technical restrictions in some tagging systems (e.g., only allowing one-word-tags for indexing) a great variety of compound terms (e.g., informationretrieval, information_retrieval, information-retrieval) as well as spelling variations (e.g.,

WebSci '11, June 14-17, 2011, Koblenz, Germany.

[1] http://readermeter.org, http://altmetrics.org/manifesto.

[2] www.citeulike.org, www.connotea.org, www.bibsonomy.org.

American English (AE) vs. British English (BE)) occur in folksonomies and docsonomies. On the other side, these term variations cannot be found in the metadata terms, because professionally created index terms are subject to indexing guidelines. Only terms extracted from the articles' titles and abstracts and author keywords may differ from the indexing guidelines because of the journals' particularly prescribed terminology. Therefore, we preprocess tags and metadata to gain a homogenous term basis for comparisons [1], [8].

We match tags and terms for every single document and then calculate average values [1]. This is done to avoid mismatches between tags and metadata of different documents, as we believe term overlap between docsonomy and respective metadata is more valuable than overlap between folksonomy and the entire metadata collection.

3. DATABASE

The data of this analysis is based on a previous study [3], which examined the application of social bookmarking data to journal evaluation. For 10,280 documents published in 45 physics journals between 2004 and 2008 bookmarks were downloaded from CiteULike, Connotea and BibSonomy. Since this study aims to compare the description of content by readers (i.e. users of social bookmarking platforms), authors, intermediaries and automatic indexing, the initial dataset was limited to a subset of documents, for which all of this data was available. The readers' perspective is covered by tags assigned by users of CiteULike, Connotea and BibSonomy. Inspec provides controlled thesaurus and uncontrolled subject headings, which are intellectually assigned by information professionals. The authors' point of view is represented by keywords, which are provided by authors in the publication, as well as by terms extracted directly from document titles and abstracts. Automatic index terms are represented by WoS KeyWords Plus$^{\text{TM}}$, consisting of words extracted from the titles of a publication's references. 724 documents fulfill all necessary criteria. For each of these publications, the required information is downloaded from the particular sources. In order to compare differences and similarities of reader and author, intermediary and automatically indexed terms on document level, each term is connected to its publication via DOI.

4. METHODS

In contrast to previous studies (amongst others: [4], [5], [7], [8]), which mainly compare tags to author keywords or indexing terms for a set of documents or even on database level, we aim to analyze the similarity of tags and titles, abstracts, author keywords, indexer- and automatically generated terms for each document separately. For each publication, the similarity between tags and Inspec terms, KeyWords Plus$^{\text{TM}}$, author keywords, title and abstract terms is computed applying cosine measure. Additionally, the percentage of overlap from tag and particular metadata-perspectives are given. While others calculate one value across the whole collection of articles, we compute the more exact mean of the means of each of the 724 documents.

4.1 Preprocessing and Cleaning

Due to the uncontrolled nature of tags and the different spelling variants of terms in titles, abstracts and keywords, data cleaning and transformation has to be applied to receive a linguistically homogenous tag collection [9]. Initially all special characters (except hyphens and underscores) are deleted and all letters converted to lower case. Stop words are removed from article titles and abstracts by a list of 571 stop words compiled for the SMART project[3] complemented by a list of dataset-specific terms (i.e. *imported*, *fileimport081104*). For comparison of tags with titles and abstracts, tags were split at the separating character (i.e. hyphen or underscore) to allow for a matching of single-word terms of title and abstracts [8]. When comparing tags to author keywords, subject headings and KeyWords Plus$^{\text{TM}}$, hyphens and underscores are deleted within tags and blanks within keywords in order to unify different spellings like *complex_network*, *complex-network*, *complexnetwork* and *complex network*. This initial cleaning process reduces the number of spelling variants in tags by 2.3% from 1,743 to 1,703 unique tags.

Table 1a. Ten most frequently indexed terms representing reader, title and abstract perspectives *(number of unique terms after cleaning)***.**

tags split at separating characters *(1,515)*		**title terms** *(1,858)*		**abstract terms** *(6,289)*	
network	75	model	353	model	2553
quantum	44	network	339	system	2167
theori	40	laser	310	result	1811
review	38	magnet	298	method	1687
dynam	35	quantum	269	right	1682
ion	35	electron	233	reserv	1677
model	33	ion	228	field	1621
electron	30	system	223	effect	1528
magnet	30	complex	215	network	1521
physic	30	effect	206	measur	1459

Since especially tags can appear in multiple forms, we decided to take the preprocessing one step further and try to unify variants as far as automatically possible. Fortunately the great majority of physics publications has been tagged in English, so that the marginal number of non-English terms can be ignored. BE suffixes are transformed into AE by applying a rule-based algorithm. The tag *synchronisation* is thus unified with the AE spelling *synchronization*. A manual check allows for reversing the algorithm in six cases, where the rule-based approach has failed, when changing *-our* to *-or*, *-ogue* to *-og* and *-tre* to *-ter* (*four*, *hour*, *our*, *homologue*, *Lemaitre*). Additionally, all terms are stemmed using the Porter 2 stemming algorithm[4]. This unifies tags like *network*, *networks* and *networking*. These additional cleaning processes reduce the unique number of tags further to 1,596. Thus, the combination of all preprocessing methods reduces spelling variations by 8.4% compared to unprocessed tags. Unifying BE and AE and applying the Porter stemmer alone cause 6.1% improvement. Due to the slightly different methods applied when comparing tags to abstract and title terms, the number of unique tags differs between tables 1a and 1b. Counter-

[3] ftp://ftp.cs.cornell.edu/pub/smart/english.stop.

[4] http://snowball.tartarus.org/download.php.

intuitively the separation of tags leads to a decreased number of unique terms (1,515 instead of 1,596 tags). This is caused by the aggregation of parts of different terms.

The same cleaning methods are applied to the other terms (title, abstract, author keywords, Inspec subject headings and KeyWords Plus^{TM}). Especially abstract and title terms can be improved by these methods: term quantity is reduced by 30.5% and 19.8%, respectively. Expectedly due to their controlled nature, the reduction for author keywords (3%), Inspec headings (2.8%) and automatic index terms (5.3%) is lower. The ten most frequently assigned terms after extensive preprocessing for the whole database are listed in tables 1a and 1b.

Table 1b. Ten most frequently indexed terms representing reader, author, intermediary and automatic indexing perspectives *(number of unique terms after cleaning)*.

tags merged at separating characters *(1,596)*	**author keywords** *(2,287)*	**Inspec subject headings** *(8,242)*	**KeyWords Plus^{TM}** *(2,390)*
network 48	network 138	review 266	dynam 223
review 36	randomgraph 134	synchron 245	complexnetwork 151
theori 29	newapplicationsofstatisticalmechan 95	complexnetwork 244	model 132
simul 26	synchron 91	groundstat 193	system 127
2007 24	complexnetwork 90	fluctuat 184	smallworldnetwork 113
physic 24	networkdynam 87	graphtheori 177	puls 98
laser 20	quantumopt 87	isingmodel 177	scalefreenetwork 96
model 20	groundstat 80	montecarlomethod 171	radiat 95
communiti 18	iiivsemiconductor 66	anneal 146	simul 89
electron 16	Isingmodel 66	randomprocess 144	generat 84

4.2 Term Comparison

It is analyzed, if terms match between the different indexing methods on document basis. Following this more exact approach, the number of cleaned unique tags, author keywords, KeyWords Plus^{TM}, Inspec, title and abstract terms has to be determined for each of the 724 journal articles, before the overlap between the different entities can be computed. The overlap counts the number of exact character strings, which appear in both, the reader and author, intermediary, automatic indexing, title or abstract data, respectively. The highest share of overlap is detected between tags and abstracts: 77.6% of the 724 articles, share at least one term in tags and abstracts. For 66% at least one tag appears in the title. This is followed by the overlap of tags and intermediary terms (33.4%) and author keywords (29.3%). Only 10.5% of all documents have at least one tag and automatically generated KeyWord Plus^{TM} in common. Again, the unification of AE and BE and stemming successfully increases the share of documents with at least one mutual term. The number of at least one overlap of tags and author keywords, Inspec headings, KeyWords Plus^{TM}, title and abstracts improved by 26.2%, 21%, 20.6%, 9.4% and 8.5%, respectively.

5. RESULTS

In contrast to [1] who use the harmonic mean we calculate the arithmetic mean of the similarity values of the 724 documents. First, the percentage of overlap is computed in contrast to the total number of unique tags per document on the one hand and to the number of the particular meta terms on the other, in order to detect the share of common tags from each of the perspectives. The overlap-tag ratio lists the percentage of overlapping tags in contrast to all unique tags assigned to the particular document and is defined as "overlap tag ratio" g/a where a stands for the number of unique tags per document and g represents the overlap between tags and terms (author keywords, Inspec headings, KeyWords Plus^{TM}, title or abstract terms, respectively) per document. Most tags are represented in the abstracts, which is to be expected, since the number of abstract terms is much greater than that of the other metadata.

The "overlap-analyzed term ratio" g/b calculates the same overlap from the other perspective. Here b stands for the number of unique terms per document and g represents the overlap between both sets per document. On average, 24.5% of title terms are taken over by users, when tagging articles. Strikingly, only 3.4% of indexer terms are adopted by users. While this might have dramatic consequences on information retrieval in Inspec, it reveals a wide difference between the reader and indexer perspectives on published contents.

To combine both measurements, the similarity between the readers' point of view on the one hand and author, intermediary and automatic indexing perspective on the other hand is calculated by cosine $g/\sqrt{a*b}$ where a stands for the number of unique tags per document, b for the number of unique terms and g represents the overlap between tags and terms per document. If a publication is tagged by its readers with exactly the same terms the author used to describe it, the similarity of author and reader indexing is 1. It is 0, if the two sets of terms are completely different.

Similarity is computed alike for tags and Inspec terms, KeyWords Plus^{TM}, title and abstract terms. While there is no document, where all tags and indexer or abstract terms match exactly, there are documents with 100% matches of tags and titles, KeyWords Plus^{TM} and author keywords, respectively. The highest number of documents with a cosine of 1 can be found for author keywords (6 documents), followed by title (3) and KeyWords Plus^{TM} (2). On average, there is hardly any overlap between reader and professional (0.062) and automatic indexing (0.026) methods. The mean cosine value is highest for title terms (0.279), abstracts (0.143) and author keywords (0.103). Overall, cosine similarities are however very low because a great share of documents do not have common tags and indexing terms. This implies that social tagging represents a user-generated indexing method and provides a reader-specific perspective on article content, which differs extremely from conventional indexing methods. When applied to journal evaluation, tags from users of social bookmarking tools can depict a reader specific view on published content.

The following example illustrates this result. The *Journal of Statistical Mechanics (J Stat Mech)* was chosen, because it had the highest tagging frequency within the dataset: 110 unique users tagged 94 articles published in the journal between 2004 and 2008. The difference in perception displayed with tags and Inspec subject headings (both preprocessed and cleaned) is shown in figure 1.

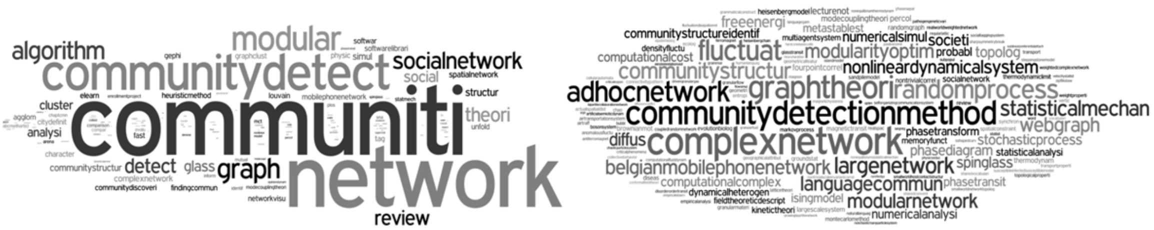

Figure 1. Term clouds depicting reader perspective (left) and intermediary perspective via Inspec subject headings (right) on 94 articles published in *J Stat Mech.*

6. CONCLUSIONS

Our study shows that author-generated, indexer-generated and user-generated index terms each reflect a different view on article content. Most term matches can be found in the comparison of tags and abstract as well as title terms, but on average nearly half of tags used (49.7%) do not occur in abstracts and 63.5% are completely different from title terms. The comparison of tags and author keywords, Inspec subject headings and KeyWords Plus^{TM} results in fewer matches. These results confirm our basic assumption that journal and article evaluation can profit from the application of user-generated tags for content analysis, as they add a third layer of perception besides the author and indexer perspectives. Due to the dynamic nature of social bookmarking and tagging, these descriptions evolve in real time. They offer direct channels to the readers' opinions and depict trends in the language of a specific discipline. We demonstrate that extensive preprocessing and cleaning (i.e. removal of special characters, unification of AE and BE spellings and stemming) of all term sets lead to a more homogenous collection of terms, which improved the calculation of overlapping terms: that is, 8.5% increase of overlap between tags and abstract terms and 26.2% improvement for the comparison to author keywords. These findings indicate that term comparisons without applying extensive cleaning are misleading and show distorted results. Cleaning methods are still limited. Since terms are not compared semantically, some problems remain (synonyms, homonyms, different languages). The advantage is however, that all facets of the users' description can be depicted. It is strongly recommended to match tags and metadata on document level, as this gains more accurate results than calculating similarities on folksonomy level, when analyzing indexing consistency of users, authors and intermediaries.

Our future work comprises the evaluation of readers' perception using weighted tag and term information. In broad docsonomies [10] not only all unique tags assigned by users can be considered, but also frequency information about how often a particular tag is used, can be applied. If we assume, that frequently used tags are more important for a document's content, we can calculate weighted overlap values. This would reveal whether authors and users attach importance to the same or different topics of a document's or a journal's content.

7. ACKNOWLEDGMENTS

Isabella Peters and Jens Terliesner are supported by the Research Funding Organization DFG (STO 764/4-1). They are associated with the research group "Science and the Internet" funded by the Strategic Research Fund of the University of Düsseldorf.

8. REFERENCES

[1] Good, B., Tennis, J., and Wilkinson, M. 2009. Social tagging in the life sciences: Characterizing a new metadata resource for bioinformatics*. BMC Bioinformatics*, 10(313). DOI= 10.1186/1471-2105-10-313.

[2] Haustein, S. 2011. Wissenschaftliche Zeitschriften im Web 2.0. Die Analyse von Bookmarks zur Evaluation wissenschaftlicher Journale. In *Proceedings of the 12th International Symposium for Information Science* (Hildesheim, Germany, March 09-11, 2011). 148-159.

[3] Haustein, S. and Siebenlist, T. 2011. Applying social bookmarking data to evaluate journal usage [Preprint]. *Journal of Informetrics*. DOI= 10.1016/j.joi.2011.04.002

[4] Jeong, W. 2009. Is tagging effective? Overlapping ratios with other metadata fields. In *Proceedings of the International Conference on Dublin Core and Metadata Applications* (Seoul, Korea, October 12-16, 2009). 31-39.

[5] Lin, X., Beaudoin, J., Bul, Y., and Desai, K. 2006. Exploring characteristics of social classification. In *Proceedings of the 17th Annual ASIS&T SIG/CR Classification Research Workshop* (Austin, USA, November 03-08, 2006).

[6] Lu, C., Park J., and Hu, X. 2010. User tags versus expert-assigned subject terms: A comparison of LibraryThing tags and Library of Congress Subject Headings. *Journal of Information Science*, 36(6), 763-779.

[7] Lux, M., Granitzer, M., and Kern, R. 2007. Aspects of broad folksonomies. In *Proceedings of the 18th International Conference on Database and Expert Systems Applications* (Regensburg, Germany, September 03-07, 2007). 283-287.

[8] Noll, M. G. and Meinel, C. 2007. Authors vs. readers. A comparative study of document metadata and content in the WWW. In *Proceedings of the 2007 ACM Symposium on Document Engineering* (Winnipeg, Canada, August 28-31, 2007). 177-186.

[9] Peters, I. 2009. *Folksonomies. Indexing and Retrieval in Web 2.0*. De Gruyter Saur, München.

[10] Terliesner, J., and Peters, I. 2011. Der T-Index als Stabilitätsindikator für dokument-spezifische Tag-Verteilungen. In *Proceedings of the 12th International Symposium for Information Science* (Hildesheim, Germany, March 09-11, 2011). 123-133.