Evaluation of Reader Perception by Using Tags from Social Bookmarking Systems

Stefanie Haustein[1,2], Isabella Peters[1] & Jens Terliesner[1]

s.haustein@fz-juelich.de | {isabella.peters | jens.terliesner}@uni-duesseldorf.de

Background: Other than citation based methods, usage statistics are able to reflect journal usage by the whole readership not just that of readers who appear as citing authors. With the emergence of electronic publishing it has become possible to evaluate journal influence on the publishing and pure readers through click and download rates.

Problem: Despite existing standards like COUNTER, download statistics are flawed. Global usage data is not available and local data often lacks appropriate normalization.

Proposed alternative: Social bookmarking services specialized on STM allow users to store, share and manage bibliographic references online. These bookmarks indicate global usage on article level. Through the tagging function users create additional metadata about the articles' contents, which can be used to analyze reader perception of journal content.

Research question: Are tags consistent with traditional indexing methods or do they provide a new, reader-specific view on journal content?

Method: Tags are compared with title and abstract terms, author keywords, Inspec subject headings and KeyWords Plus™ on article level. Spelling variants are unified automatically to improve term matching. Similarity values are computed to indicate in how far readers apply the same terms as authors, intermediaries or automatic indexing methods.

Data preprocessing and cleaning

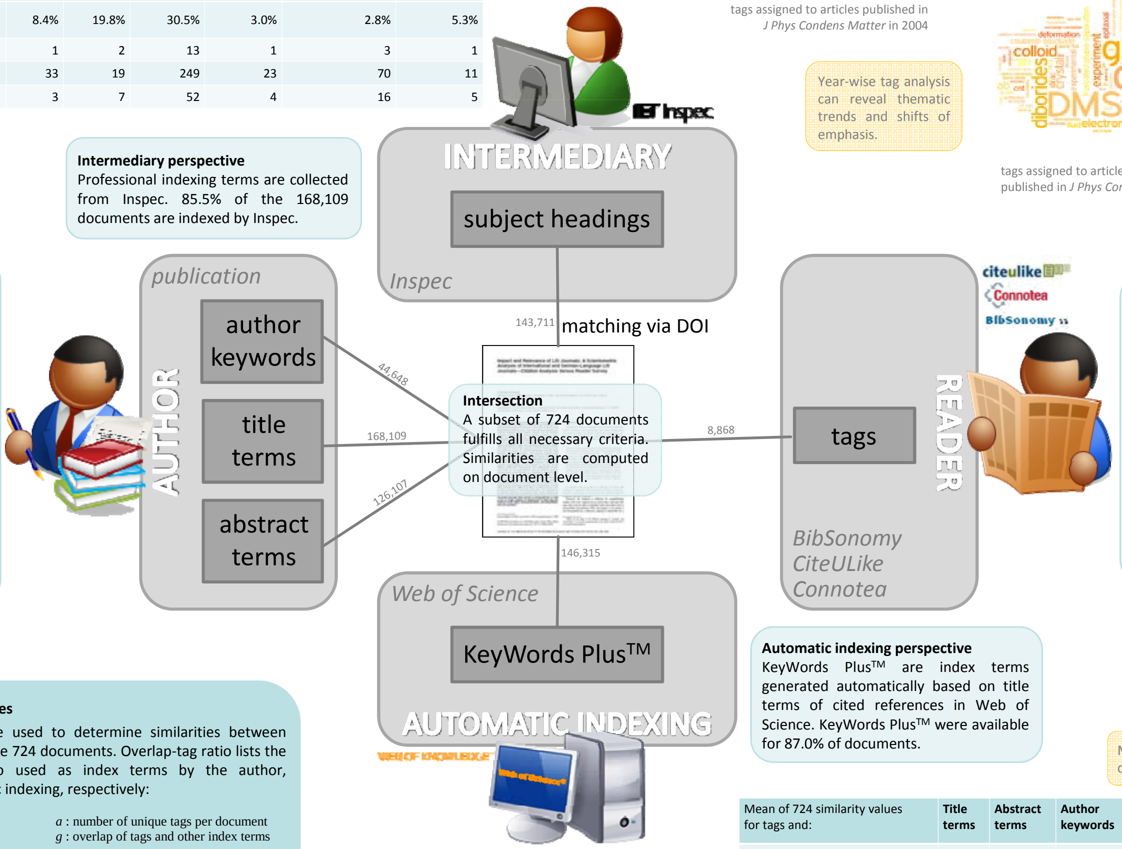
Due to the uncontrolled nature of tags and spelling variants in keywords, title and abstract terms, terms are preprocessed in order to obtain a linguistically homogenous term collection:

- removing special characters *(except "-" and "_")*
- removing stop words
- transferring BE to AE suffixes
- stemming with Porter 2

		Tags	Title terms	Abstract terms	Author keywords	Inspec subject headings	KeyWords Plus™
Reduction of unique terms through preprocessing		8.4%	19.8%	30.5%	3.0%	2.8%	5.3%
Number of unique terms per document after preprocessing	Min	1	2	13	1	3	1
	Max	33	19	249	23	70	11
	Median	3	7	52	4	16	5

Intermediary perspective
Professional indexing terms are collected from Inspec. 85.5% of the 168,109 documents are indexed by Inspec.

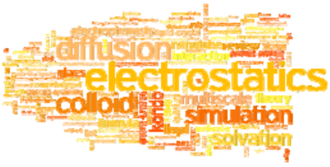
Author perspective
The initial data set contains 168,109 documents published between 2004 and 2008 in 45 physics journals. Bibliographic data was downloaded from Web of Science, DOIs checked and completed via CrossRef.
All 168,109 entries contain a title, 75.0% an abstract and for 26.6% author keywords have been provided by the journal.



Intersection
A subset of 724 documents fulfills all necessary criteria. Similarities are computed on document level.

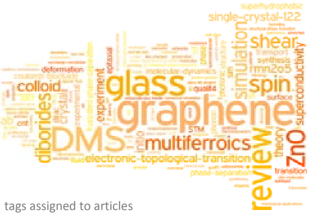
Reader perspective
Social bookmarking data was collected from CiteULike, Connotea and BibSonomy. 10,280 of the 168,109 documents were bookmarked 13,608 times by 2,441 users. 86.3% of all bookmarked publications were tagged. 8,868 publications contained 35,804 tag applications by 1,992 users.

Automatic indexing perspective
KeyWords Plus™ are index terms generated automatically based on title terms of cited references in Web of Science. KeyWords Plus™ were available for 87.0% of documents.



tags assigned to articles published in *J Phys Condens Matter* in 2004

Aggregated on journal level *(i.e. "joursonomy")*, tags reveal the readers' perspective on journal content.



tags assigned to articles published in *J Phys Condens Matter* in 2008

Year-wise tag analysis can reveal thematic trends and shifts of emphasis.

Most tags can be found in the document's abstract.

Mean of 724 similarity values for tags and:	Title terms	Abstract terms	Author keywords	Inspec subject headings	KeyWords Plus™
Mean overlap-tag ratio	36.5%	50.3%	11.8%	13.3%	2.9%
Mean overlap-analyzed term ratio	24.5%	4.8%	10.4%	3.4%	3.0%
Mean cosine similarity	0.279	0.143	0.103	0.062	0.026

Readers use different terms and emphasize other topics than professional indexers.

Measuring term similarities

Three measurements are used to determine similarities between index terms for each of the 724 documents. Overlap-tag ratio lists the percentage of tags also used as index terms by the author, intermediary or automatic indexing, respectively:

$$\text{overlap-tag ratio} = \frac{g}{a}$$

a : number of unique tags per document
g : overlap of tags and other index terms

Overlap-term ratio analyzes the percentage of index terms taken up by tagging users:

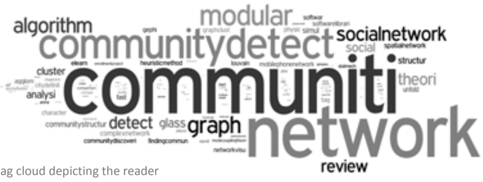
$$\text{overlap-analyzed term ratio} = \frac{g}{b}$$

b : number of unique index terms per document
g : overlap of tags and index terms

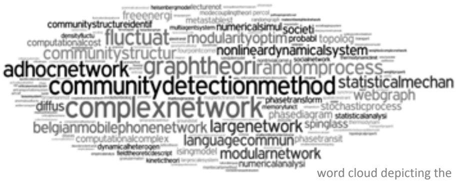
To combine both measurements, similarity between the readers' point of view on the one hand and author, intermediary and automatic indexing perspectives on the other, is calculated by cosine:

$$\text{cosine similarity} = \frac{g}{\sqrt{a \cdot b}}$$

a : number of unique tags per document
b : number of unique index terms per document
g : overlap of tags and index terms



tag cloud depicting the reader perspective on 94 articles published in *J Stat Mech*



word cloud depicting the intermediary perspective on 94 articles published in *J Stat Mech*

References

Haustein, S., Golov, E., Luckanus, K., Reher, S., & Terliesner, J. (2010). Journal evaluation and science 2.0. Using social bookmarks to analyze reader perception. In *Book of Abstracts of the 11th International Conference on Science and Technology Indicators, Leiden, the Netherlands* (pp. 117-119).

Haustein, S. & Siebenlist, T. (2011). Applying social bookmarking data to evaluate journal usage. *Journal of Informetrics*, 5(3), 446-457.

Kipp, M. E. I. (2005). Complementary or discrete contexts in online indexing: A comparison of user, creator, and intermediary keywords. *Canadian Journal of Information and Library Science*, 29(4), 419-436.

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

Noll, M. G., & 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* (pp. 177–186).

Peters, I., Haustein, S., & Terliesner, J. (2011). Crowd sourcing in article evaluation. In *Proceedings of the ACM WebSci'11, Koblenz, Germany* (pp. 1-4).

Results

Low similarity values show that tags assigned to journal articles by users of social bookmarking differ from traditional indexing methods. Mean cosine similarity is highest between tags and title terms.

Journal 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.

Limitations

Social bookmarking in STM is still in its infancy. Thus the database is comparatively small and does not allow for generalization.

The group of users of social bookmarking is not identifiable and may not adequately represents the population of readers of physics journals.

Although matching could be improved through term unification (stemming etc.), similarity is based on lexical not semantic comparison.

[1] Department of Information Science, Heinrich Heine University Düsseldorf
Universitätsstraße 1, 40225 Düsseldorf (Germany)

[2] Central Library, Forschungszentrum Jülich
52425 Jülich (Germany)

Mitglied der Helmholtz-Gemeinschaft