Uncovering Hidden Qualities – Benefits of Quality Measures for Automatically Generated Metadata

Sascha Tönnies[1] and Wolf-Tilo Balke[1,2]

[1] L3S Research Center, Appelstraße 9a, 30167 Hannover, Germany
[2] IFIS TU Braunschweig, Mühlenpfordstraße 23, 38106 Braunschweig, Germany
toennies@L3S.de, balke@ifis.cs.tu-bs.de

Abstract. Today, digital libraries more and more have to rely on semantic techniques during the workflows of metadata generation, search and navigational access. But, due to the statistical and/or collaborative nature of such techniques, the underlying quality of automatically generated metadata is questionable. Since data quality is essential in digital libraries, we present a user study on one hand evaluating metrics for quality assessment, on the other hand evaluating their benefit for the individual user during interaction. To observe the interaction of domain experts in the sample field of chemistry, we transferred the abstract metrics' outcome for a sample semantic technique into three different kinds of visualizations and asked the experts to evaluate these visualizations first without, later augmented with the quality information. We show that the generated quality information is indeed not only essential for data quality assurance in the curation step of digital libraries, but will also be helpful for designing intuitive interaction interfaces for end-users.

Keywords: Digital Libraries, Information Quality, Semantic Technologies

1 Introduction

Digital Libraries have to handle a vast amount of data ranging from individual papers or reports in journals, conference proceedings etc. up to complete digitized books. Making such data searchable relies mainly on the amount and quality of the provided metadata. On a purely bibliographic level this metadata is relatively easy to derive and maintain, in contrast on the content level the problem of deriving correct metadata obviously grows with the density of information. Whereas the information contained in short conference papers can be manually extracted and annotated quite easily, capturing the content contained in a book definitely needs automatic means of extraction. Today semantic techniques relying on statistically approaches like term co-occurrences or frequencies are already commonplace. But the quality of metadata derived by such techniques is largely uninvestigated. Thus a main topic of future digital library research has to be a quality assessment of such techniques. Obviously the quality – like in information retrievals' precision / recall analysis – can only be evaluated comparing the techniques output with manually provided judgments.

In our previous research about digital libraries [1] and large digital book collections [2] we proposed three general metrics, i.e. Degree of Category Coverage (DCC), semantic word bandwidth (SWD) and relevance of covered terms (RCT), for measuring the quality of semantic techniques used for taxonomy / folksonomy creation. These quality measures were derived by observing the workflow of a domain expert using the example of (but not limited to) the field of chemistry. First evaluations already pointed out the frameworks' usefulness but a thorough investigation is still needed.

In this paper, we evaluated the metrics' usefulness taking the Semantic GrowBag technique as an example of automatic metadata generation techniques. Our contribution is threefold:

- First, we performed a user study with domain experts to assess the general applicability und usefulness of our quality measures in the field of chemistry.
- Second, we applied the quality measures to the Semantic GrowBag technique to demonstrate the added value of purely statistically techniques in metadata generation.
- Third, we showed that already simple diagram types are sufficient to transport the information provided by quality measures.

2 Related Work

Recently digital libraries have made the step towards using semantic techniques for *automated metadata generation*. Typical examples include for instance exploiting term co-occurrences or language models to find relevant keywords, categorization, or even inter-document relationships like for instance in JeromeDL [3]. But already in early projects the need to evaluate the quality of the metadata became clear, although it has usually been restricted to general user satisfaction studies.

An excellent overview of related work in the field of quality assessment of manually and automatically generated metadata and semantic annotation techniques is done in [1]. But of course the annotated information also has to be displayed to help users in efficient document retrieval and selection. Thus, up to a certain degree also *information visualization techniques* are related to our paper. The classical representations of taxonomies used for navigational access are acyclic directed graphs, usually trees. Also, the Semantic GrowBag technique [5] used in our evaluation generates graphs of the automatically generated taxonomies. On the other hand, light-weight ontologies or folksonomies are often simply represented by tag clouds. Due to the popularity of tag clouds a lot of work has already been done to investigate their possibilities for searching and browsing in large document collections. For instance, [6] tries to answer the question whether tag clouds provide sufficient value for information seeking. Tag clouds are characterized as particularly useful for browsing or non-specific information discovery. Moreover, tag clouds provide a compact visual summary of the content and scanning the tag cloud requires less cognitive load than formulating specific query terms. In contrast, building tag clouds requires a lot of effort, if the esthetic sensation of a user should be matched. Therefore, several approaches, e.g. [7], and [8] investigate the influence of text size and position as well as the algorithms to use for tag cloud generation. Still, all this work relies on the frequency of terms and not on the underlying quality.

3 Evaluating the Value of Quality Measures for Semantic Techniques

We conducted a user study with domain experts, in our case practitioners in the field of organic chemistry, for evaluating the three metrics DCC, SWD and RCT defined in [1]. The aim of the study was first to get a feeling whether the defined metrics are useful and second how important the information provided is compared to classical forms of visualization.

For the experiments we used a corpus of 4554 documents extracted from the Journal of Synthetic Organic Chemistry (SYNTHESIS) published by Thieme Publishers, Stuttgart, Germany. For all papers we extracted the author keywords (9554) and eliminated all those with little discriminating power (terms like '*experiment*' or '*synthesis*') occurring in many papers. For the remaining set of about 1600 terms folksonomies were generated using the Semantic GrowBag technique [5], which relies on higher order co-occurrences of keywords in relation to the respective documents. For the actual experiments we randomly chose ten query terms for each expert to evaluate the quality of the respective folksonomy. For each query term we generated three different kinds of visualization: the original GrowBag graph (Fig. 1, query term in black box), the respective Tag Cloud and a concentric circle diagram (CCD) (Fig. 2, query term in the center).

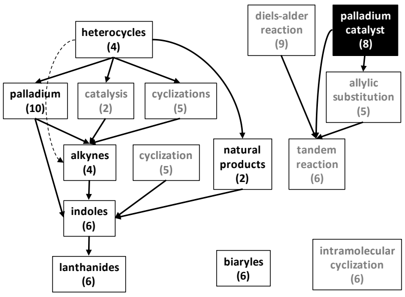

Fig. 1. The generated GrowBag graph for the keyword *palladium catalyst*.

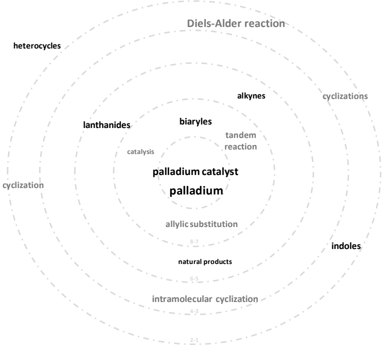

Fig. 2. The generated concentric circle diagram for the keyword.

Basically the information provided by our three quality metrics, i.e. related category, overall specificity topical and distance to query term, can be represented by the visual features text color, text size and spatial layout. Please note, this information can be easily visualized in the CCD, whereas the other two visualizations lack the possibility to visualize the distance in an intuitive way. For the tag cloud we tried several clustering algorithms and visualized the terms in clusters, thus sacrificing the compactness of display for the possibility to show spatial relationships. However, since intra-cluster similarity usually clashed with the individual terms' relationship to the query users tended to be confused by that notation. Hence, in our experiments we tested the advantages of compact visualization versus the benefits of the information provided by the distances to the query term.

3.1 Designing the user study using the Semantic GrowBag technique

To control the environment we ensured that all participating domain experts were recruited from the field of organic chemistry and in particular familiar with the focus area of the SYNTHESIS journal. Hence, we could expect only slight variations with respect to the individual knowledge spaces. Although the experts did not know about the specific semantic GrowBag technique used for deriving the graphs, all participants had prior experience with the use of ontological information retrieval and were proficient in using computing devices.

As stated above for all users we randomly selected ten query terms and confronted them with the three different visualizations in individual questionnaires. Filling in the questionnaire took only about half an hour of time. To illustrate the design of our user study let us focus on an example evaluation workflow for the query term '*palladium catalyst*'. The respective visual representations are shown in figures 1 – 3. Each questionnaire was divided into three major blocks:

- The first block of questions in the questionnaire focused on the first impression with respect to the diagrams. Users were asked to rank the different diagram forms for each query term regarding the intuitive understandability, i.e. the degree of ease to grasp the concepts contained.
- After evaluating the first impression the second block should prove if the users intuitively interpreted the diagrams in the correct way. Therefore, each metric and the correlation between the metrics' outcome and the diagrams were explained. With this knowledge the users were again asked to rank the different diagram forms.
- The third block actually measured the correctness of the three quality metrics. Therefore the domain experts were asked to rank the visualized metrics' outcome for each query term.

In more detail, the metrics explained in the second part of the evaluation concluded in the following visualization. Focusing again on the example query term '*palladium catalyst*', all terms can be categorized into the two categories, i.e. reactions (light) and chemical substances (dark). The size of each keyword represents the overall specificity, e.g. '*Diels-Alder reaction*' is quite specific term as it represents a concept of specific reaction with given reactants, products, solvents and reaction conditions. This way a domain expert reading the term 'Diels-Alder reaction' - maybe in connection with a substance - has a good impression of a reaction scenario and possible products. In contrast, the '*tandem reaction*' is an unspecific term describing a broader concept of a reaction type with much more space for interpretation. The term '*tandem reaction*' just defines a cascade of reactions from an educt to a product without the isolation of any intermediate product: the actual reactions of the cascade are not defined in detail by this concept. As already stated above, the closeness of a term in relation to the query term can only be visualized within the CCD, i.e. the distance of each keyword to the circles' center. Thus, the query term '*palladium catalyst*' is located in the circle center. Closely related terms like e.g. '*palladium*' and '*tandem reaction*' are located nearby, whereas only loosely related terms are located far away e.g. '*heterocycles*'. The closeness of '*palladium catalyst*' and '*palladium*' is obvious as a palladium catalyst contains the metal palladium, which in turn defines the functionality of the catalyst. Tandem reactions are most often catalyzed reactions with a high stereo selectivity induced by various classes of palladium catalysts. An example for a loosely related term is '*heterocycles*', which represents a general concept of a substance class with a rather weak relation to the term '*palladium catalyst*'.

For the last part of the experiment the domain experts were given a scale divided into five degrees (0 - 4) of satisfaction (see table 1).

Table 1. Evaluation scale for part 3 of the experiment

Value	DCC: percent of occurring concepts	SWD: percent of matching proportional font sizes	RCT: percent of matching distances
4 - completely satisfied	> 90%	> 90%	> 90%
3 - mostly satisfied	~ 75 %	~ 75 %	~ 75 %
2 – satisfied	~ 50%	~ 50%	~ 50%
1 - partially satisfied	~ 25%	~ 25%	~ 25%
0 – unsatisfied	≤ 10%	≤ 10%	≤ 10%

3.2 Experimental results

In the first part of the experiment we evaluated the first impression and the intuitive understandability of the respective visualizations. We expected a high rank of the tag cloud as it is a compact and well known kind of visualization. Surprisingly, as can been seen in figures 2 the concentric circle diagrams (CCD) were already ranked considerably higher immediately claiming about 95% of the position one ranks with an average rank of 1.07. In contrast the tag cloud visualization just got an average rank of 2.1 and the remaining 5% of position one ranks. The (somewhat harder to understand) ontology graph was never ranked at position one and only got an average rank of 2.82.

It is interesting to note that in topically focused document collections quality information blended into navigational information or categories is indeed attractive for users. This also shows that even the simple CCD is already an intuitive way of visualization for our quality metrics. Also a later interview with selected domain experts confirmed this: they explained the lower rank of the tag cloud, because the co-occurring terms were sometimes misleading. But the adoption of the distance in the CCD clarified intuitively that some terms may belong to the query term in a rather loosely coupled way.

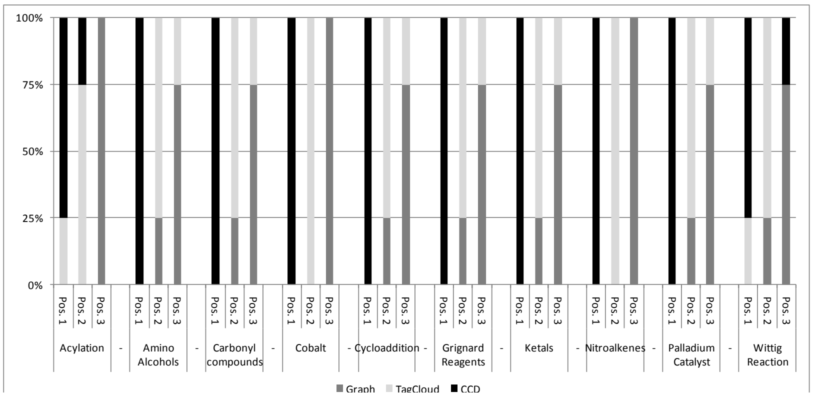

Fig. 3. First impression results

In the next step of the experiment, the visualizations' semantics in terms of encoded quality measures was explained in detail and the domain experts were asked to re-rank the three kinds of visualization. As expected, the CCD was still most often ranked at position one (see figure 3). However, a marginal loss of 2.5% in position one ranks occurred, still resulting in 92.5% of top positions and an average rank of 1.1. At this stage, although gaining 7.5% of the overall position one ranks, tag clouds experienced a slight drop in the average rating (2.15). This can particularly be attributed to their limitations becoming clear during the explanation of the semantics: users better understood their power of compact representation, but also their difficulties in discriminating terms. Again, the graph representation was never ranked first but the re-ranking still resulted in a slightly better average rating of 2.75. Further interviews with the domain experts have shown that they liked the tag cloud more than the CCD in situations, when confronted with very sparse CCD diagrams. This shows that there is a tradeoff between compactness of the visualization and the transported information. One possibility to handle this tradeoff would thus be a digital library interface where first a tag cloud of the digital collection is shown and once a user selects a term for deeper investigations, the respective CCD is shown.

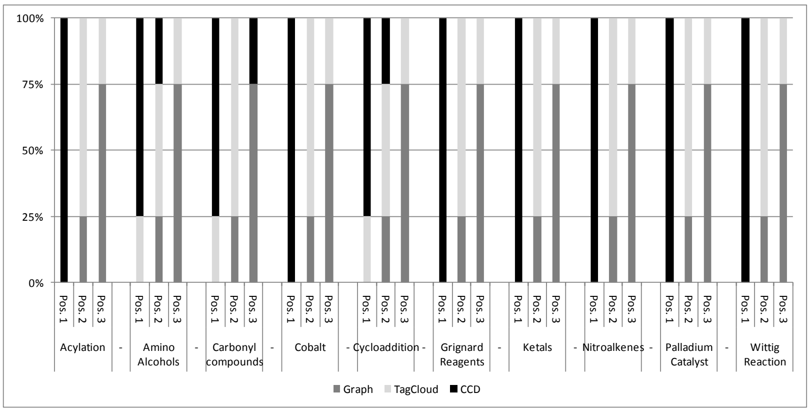

Fig. 4. Second impression results

Due to the fact, that the CCD is the only diagram which represents our metrics' entire outcome and that the rank has only slightly been decreased after explaining the quality metrics contained, our three metrics indeed seemed reasonable for the domain experts. A deeper investigation as last part of the evaluation further substantiates this prediction. We asked all experts to consider the three metrics individually and evaluate the terms provided by the Semantic GrowBag technique for the ten test queries. As can be seen in figure 4 on a scale from 0 to 4 none of the metrics' outcome has been ranked less than 2, i.e. the 50% mark of satisfaction. In average the degree of domain coverage (DCC) was ranked with 3.20, the semantic bandwidth (SWD) with 2.82 and the relevance of covered terms (RCT) with 3.18. On average the domain experts were mostly satisfied with the quality of the Semantic GrowBag technique's generated metadata and also the proposed quality metrics' usefulness in quality assessment was confirmed.

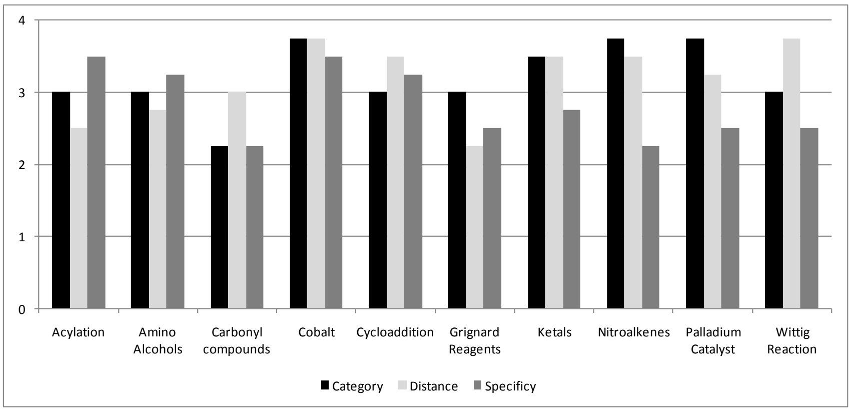

Fig. 5. Rating of the correctness of the quality aspects from unsatisfied (0) to completely satisfied (4)

4 Conclusions and Future Work

Since customers from academia and industry depend on timely and correct information provisioning, the focus of today's digital libraries has to be on information quality. Therefore, digital libraries to a large degree still rely on manual curation (in contrast to e.g., IR-based indexing in Web search engines). However, given the exponential growth of digitally available documents and the high costs of manual labor, also digital libraries soon will have to employ social or semantic techniques for automated metadata generation. But in order to still uphold the high quality standards, the quality of all metadata, and thus, ultimately the quality of the techniques used for metadata generation has to be assessed.

To address this problem, we proposed a set of three general purpose quality metrics for metadata generation by supervising the interaction of domain experts with metadata in the example field of chemical literature. In this paper we investigated the usefulness of the proposed metrics and their information gain. For this purpose, we conducted a user study with domain experts again relying on the field of chemistry as an example application. We showed that it is indeed useful to measure the quality of a semantic technique: the domain experts were easily able to asses the outcome of the technique and gained insights into what quality to expect during their information gathering. The Semantic GrowBag, a statistic technique relying on term-co-occurrences for deriving metadata, was graded with an average of about 'mostly satisfied', i.e. about 75% complete, related, and specific. Moreover, by also providing the quality values visually for each term within the navigation elements, domain experts were less confused (especially when interacting with a low grade folksonomy).

For the investigation during our survey we used a very simple kind of diagram derived from fisheye view interfaces visualizing quality values of each term by its distance to the query term: the concentric circle diagram. Still, we were able to show that users are more satisfied by the experience of using this kind of diagram than the semantically rather shallow, yet popular tag clouds. In future work we also want to address the problem of different interfaces for visualizing quality information. The problem here seems to be twofold: one aspect is the uses of semantic techniques while indexing documents, which suggests a kind of 'quality dashboard' for the curator. The second aspect has to deal with the navigational elements used during user interactions for information seeking.

Acknowledgments. This work was supported by the German Research Foundation (DFG) within the program 'Scientific Library Services and Information Systems'.

References

[1] S. Tönnies and W. Balke, "Using Semantic Technologies in Digital Libraries – A Roadmap to Quality Evaluation," *13th European Conference, ECDL 2009, Corfu, Greece, September 27 - October 2, 2009*, Berlin, Heidelberg: Springer Berlin / Heidelberg, 2009, pp. 168-179.

[2] S. Tönnies and W. Balke, "User-centered Content Provisioning over Large Collections of eBooks," *Proceedings of the 2009 2nd ACM Workshop on Research Advances in Large Digital Book Repositories, BooksOnline 2009, Corfu, Greece, October 2, 2009*, 2009.

[3] S.R. Kruk, T. Woroniecki, and A. Gzella, *JeromeDL – a Semantic Digital Library*.

[4] S. Tönnies and W. Balke, "Using Semantic Technologies in Digital Libraries – A Roadmap to Quality Evaluation," *13th European Conference, ECDL 2009, Corfu, Greece, September 27 - October 2, 2009*, Berlin, Heidelberg: Springer Berlin / Heidelberg, 2009, pp. 168-179.

[5] J. Diederich and W. Balke, "The Semantic GrowBag Algorithm: Automatically Deriving Categorization Systems," *11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, 2007.

[6] J. Sinclair and M. Cardew-Hall, "The folksonomy tag cloud: when is it useful?," *Journal of Information Science*, vol. 34, 2007, pp. 15-29.

[7] A.W. Rivadeneira, D.M. Gruen, M.J. Muller, and D.R. Millen, "Getting our head in the clouds: toward evaluation studies of tagclouds," *Conference on Human Factors in Computing Systems*, 2007.

[8] S. Lohmann, J. Ziegler, and L. Tetzlaff, "Comparison of Tag Cloud Layouts: Task-Related Performance and Visual Exploration," *12th IFIP TC 13 International Conference*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 392-404.