Authors vs. Readers - A Comparative Study of Document Metadata and Content in the WWW*

Michael G. Noll
Hasso-Plattner-Institut,
University of Potsdam
14440 Potsdam, Germany
michael.noll@hpi.uni-potsdam.de

Christoph Meinel
Hasso-Plattner-Institut,
University of Potsdam
14440 Potsdam, Germany
meinel@hpi.uni-potsdam.de

ABSTRACT

Collaborative tagging describes the process by which many users add metadata in the form of unstructured keywords to shared content. The recent practical success of web services with such a tagging component like Flickr or del.icio.us has provided a plethora of user-supplied metadata about web content for everyone to leverage.

In this paper, we conduct a quantitative and qualitative analysis of metadata and information provided by the authors and publishers of web documents compared with metadata supplied by end users for the same content. Our study is based on a random sample of 100,000 web documents from the Open Directory, for which we examined the original documents from the World Wide Web in addition to data retrieved from the social bookmarking service del.icio.us, the content rating system ICRA, and the search engine Google. To the best of our knowledge, this is the first study to compare user tags with the metadata and actual content of documents in the WWW on a larger scale and to integrate document popularity information in the observations. The data set of our experiments is freely available for research.

Categories and Subject Descriptors

I.7.4 [**Document and Text Processing**]: Electronic Publishing; I.7.1 [**Document and Text Processing**]: Document and Text Editing—*Document Management*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information Filtering*

General Terms

Experimentation, Human Factors, Measurement

*This is the author's version of the work. It is posted here by permission of ACM for your personal use. Not for redistribution. The definitive version was published in the Proceedings of ACM Symposium on Document Engineering, pages 177-187, Canada, 2007. http://doi.acm.org/10.1145/1284420.1284465

DocEng'07, August 28–31, 2007, Winnipeg, Manitoba, Canada.

Figure 1: A so-called *tag cloud* of the most popular tags on del.icio.us.

Keywords

authoring, del.icio.us, dmoz, dmoz100k06, document engineering, google, icra, metadata, pagerank, social bookmarking, tagging, www

1. INTRODUCTION

The recent emergence and success of so-called tagging with services such as del.icio.us or Flickr have shown the great potential of this simple yet powerful approach to add metadata to documents. Unlike traditional categorization systems, the process of tagging is nothing more than annotating documents with a flat, unstructured list of keywords called *tags*. Users can browse or query documents by tags, and so-called *tag clouds* provide a rudimentary but often sufficient way to find popular and interesting content. Figure 1 shows a so-called tag cloud of popular tags on del.icio.us. It is the collection of the tags most frequently used by del.icio.us users when bookmarking web documents.

Although the number of peer-reviewed research on tagging is still comparatively low, several studies have already analyzed the semantic aspects of tagging and why it is so popular and successful in practice [11, 5, 10, 1]. A common argument is that tagging works because it strikes a balance between the individual and the community: the cost of participation, in particular entering data, is low for the individual, and tagging a document benefits both the individual and the community.

In this paper, we analyze and compare metadata provided by end users via social bookmarking and tagging with traditional types of document metadata supplied by the authors and publishers of web content. We are interested in finding out how much metadata is available for web documents in

the WWW, and whether and how a document's metadata and content supplied by authors differ from metadata supplied by readers, i.e. end users. Our work is based on a random sample of 100,000 web documents from the Open Directory, for which we examined the original documents from the World Wide Web in addition to data retrieved from the social bookmarking service del.icio.us, the content rating system ICRA, and the search engine Google. We describe what can be deduced from the results for further research and development in the areas of document engineering, information retrieval and information filtering.

The rest of this paper is organized as follows. In section 2, we briefly outline the different types and forms of metadata available for describing and annotating web documents. In section 3, we describe how we obtained real-world data for building the experimental data set used for our analysis. We report and discuss the results of our experiments in section 4, and give a summary of our findings in section 5.

2. WEB DOCUMENTS

2.1 Metadata provided by authors and publishers

2.1.1 Traditional HTML metadata

The traditional and most common method of adding metadata to web documents is described in the (X)HTML standards[1], which define elements and attributes for specifying metadata in the document source itself. This implies that this kind of metadata is provided by the authors or publishers of online content. For example, authors should use the TITLE element to identify the contents of a document. While adding a title to a document is common in practice as we will see, other metadata such as META keywords or META description is often neglected by authors out of convenience, ignorance[2], or lack of motivation. The purpose of these elements and attributes has been to help users find relevant content. However, search engines like Google or Yahoo often do not trust and therefore discard HTML metadata elements in web documents because these have been abused by spammers in the past [14]. Since search engines do not guarantee to honor this data at all, an incentive for authors to add this information is often missing.

2.1.2 Content rating systems

While the goal of traditional HTML metadata as outlined above has been to support promotion and retrieval of interesting online documents, the purpose of so-called Internet content rating systems is the opposite: restricting access to online content. Rating systems define special metadata, so-called *content labels*, to describe and rate content depicted in web documents. The creation of these content labels is generally performed on a voluntary basis by the authors and publishers of web documents, who integrate the rating information into the document sources. Most of today's Internet content rating systems are based on PICS[3], the Platform for Internet Content Selection [19]. PICS was originally designed to help parents and teachers control what children access on the Internet, and it is a platform on which other rating services and filtering software have been built. The most prominent content rating system in the Internet today is developed and maintained by the Internet Content Rating Association (ICRA)[4], an independent non-profit organization established in 1999 by a group of international Internet companies and institutions. In the same year, ICRA superseded the older RSAC rating system. The cornerstone of the rating system is the ICRA vocabulary[5], which defines a set of descriptors for classification of online content. The vocabulary covers nudity and sexual content, violence, language, chat facilities, and other topics such as gambling or drugs.

A selection of ICRA descriptors is listed in table 1. An exemplary application which uses content labels is Microsoft's Internet Explorer: the browser ships with a *Content Advisor* feature which can be configured to filter access to web documents based on content labels as.

descriptor	**category**	**meaning**
n 1	nudity	exposed breasts
s 1	sexual material	passionate kissing
v 2	violence	injury to human beings
l 1	language	mild expletives
Format: <meta http-equiv='pics-Label' content='...' />		

Table 1: Exemplary ICRA descriptors for creating digital labels to rate online content.

Rating systems for Internet content sound promising on paper. Obviously, the availability of such manually applied labels could make automated content filtering per se rather trivial and theoretically more reliable than heuristic methods for content classification. However, the viability and success of any kind of content rating system depend heavily on the actual usage of such systems by authors and publishers, and the accuracy and trustworthiness of rating information. In this paper, we continue our previous studies about content rating systems [12] and compare the results with user-supplied metadata as described in the next section.

2.2 Metadata provided by end users

Social bookmarking and tagging services such as del.icio.us, CiteULike and Connotea take a different approach. Here, the recipients and readers of online content supply metadata about web documents in a collaborative fashion. This metadata is not part of the document source but stored at and available from external web services. In the case of del.icio.us, the metadata of a web document is stored as bookmarks of the document's URL with additional tag information. Organizing and sharing bookmarks with the help of tags mitigates some of the problems of traditional, hierarchical bookmarking (for example, where to file a bookmark if it fits to more than one category without filing it twice) and increases findability.

Basically, tagging can be interpreted as a relation

$$R_{tagging} \subseteq D \times U \times T$$

[1] http://www.w3.org/MarkUp/

[2] If, for example, an author relies on software tools to create web documents, the quality of the tools often determines whether meaningful metadata is added to the document or not.

[3] http://www.w3.org/PICS/

[4] http://www.icra.org/

[5] http://www.icra.org/vocabulary/

where D is the set of documents, U the set of users and T the set of tags. The act of bookmarking a document with tags by a user creates one or more tuples as described by the relation above. Documents are identified by their URLs and users by their account name in the bookmarking service.

Golder and Huberman [5] and Ames and Naaman [1] analyzed the structure and dynamical aspects of collaborative tagging systems and user motivations for annotation of resources. The evolution of such systems depends on a variety of factors such as the user interface, tagging rights, user incentives, social connectivity, and the personal characteristics of individual users as described in [17, 10].

The social bookmarking service del.icio.us, which we used as information source for user-supplied metadata in this paper, provides a *free-for-all* tagging system (to use the terms of Marlow et al. [10]) in which users can freely annotate any document with as many tags as they want. The del.icio.us interface affords for *suggested-tagging*, i.e. it supports users in tagging documents by recommending tags and displaying a document's most popular tags.

2.3 Document content

A lot of methodologies and techniques in information retrieval and data mining focus on information extracted from the actual content of documents. Bag-of-words or n-gram approaches are common for classification and clustering tasks, and a plethora of refinements help to increase the performance of these techniques, for example by using stop words to filter out common words, or term-weighting techniques such as TFIDF [16].

The drawback of content-based approaches for the World Wide Web in practice is the difficulty to extract meaningful information from web documents because these may contain lots of different, non-trivial content types such as images, videos, Java applets or Flash. While it is very easy for a human to analyze such content, it is a much harder task for algorithms even with modern processing power. For example, image processing algorithms may be able to identify human faces or nudity in images up to a certain reliability, but such techniques are often restricted to very specific problem domains [20, 15, 8]. Secondly, results of machine learning algorithms depend heavily on quantity and quality of training input, and training input varies with a user's individual preferences and characteristics. An algorithm for binary classification will not yield optimal results if it is not trained with a sufficient number of samples from both classes, even though training tricks such as PEBL [21] may help up to a certain extent.

Other approaches try to circumvent or at least mitigate the problems of content extraction by using different sources of information. For example, ranking and classification techniques may use incoming or outgoing hyperlinks of a web document to infer information about the document and its neighbors [6, 9]. Hybrid solutions combine content-based and link-based approaches, for instance by integrating the anchor text of incoming hyperlinks into the analysis [4, 2].

3. DATA SETS

We have created a data set called *DMOZ100k06* by building an initial random sample of 100,000 URLs from the Open Directory[6], which contained 4,818,944 URLs in over 590,000 categories in December 2006. For each URL (i.e. document) in the sample, we retrieved the actual document plus metadata from the social bookmarking service del.icio.us, from the Internet Content Rating Association, and from Google as shown in figure 2. For this purpose, we implemented custom software tools which relied on the services' official APIs where possible and fell back to alternative techniques for situations where the APIs did not provide the required functionality. The final data set and the data mining tools are available on the author's home page.

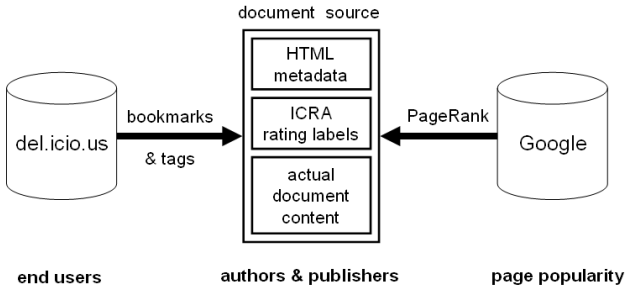

Figure 2: Information sources used for building the experimental data set used in this paper.

3.1 Document source (WWW)

For each URL in the data set, we downloaded the HTML document source from the WWW. We removed such documents from the data set whose corresponding IP addresses could not be resolved via DNS or whose HTML source code could not be successfully retrieved[7] after multiple retries over the period of two weeks. This step reduced the size of the initial data set from 100,000 to 97,578 URLs. We extracted author-supplied metadata and the actual document content by parsing and analyzing the document sources. Content rating information is generally contained within a document's source code just like standard HTML metadata, however we decided to use a different method for analyzing content labels as described in the next section.

3.2 Content rating (ICRA)

In order to ensure the correctness of our experiments, we developed an automated software tool to facilitate ICRA label tests for web documents. This tool queries the official ICRA label tester service in "strict rules" mode for each document in the data set and returns the official test result.

Basically, the label test consists of three sub-tests: label presence, label syntax, and label scope. First, it tests a web document for the presence of a content label; second, it verifies the syntactical correctness of the label; third, it verifies whether the complete web document is labeled including any elements such as hyperlinked images. A label tester result of "red" means that either no label has been found at all or only labels with errors[8] were present; "yellow" indicates a partially but not fully labeled web document, i.e. although the web document carries a label, some elements such as images or banners are not labeled or covered by the existing

[6]http://www.dmoz.org/

[7]More precisely, URLs with an HTTP status code other than 200 as defined in RFC2616 were discarded.

[8]For example, a typographic error in the URL definition of a label.

label. A "green" result is returned for a fully rated website with a syntactically correct label. In addition to these three official results, we have included a fourth result, "error", which indicates the failure of the label test[9]. It is important to note that the ICRA framework counts "yellow" web documents as unrated and does not verify whether rating labels actually match the content they describe.

3.3 Tagging (del.icio.us)

For our study, we use the tagging data available at del.icio.us, one of the most popular social bookmarking services. It has a large community of more than one million registered users [10], who can bookmark and tag web documents and share this information with other users. The data from del.icio.us was collected over a period of three weeks in December 2006.

For each URL, we retrieved the following information:

- the number of del.icio.us users who have bookmarked the URL (= number of bookmarks)
- the list of so-called "common tags" of a URL, i.e. the list of the most popular tags of the document

Del.icio.us limits "common" tags to 25 per URL, which means that the list of all tags attached to a document might actually be larger than 25. This implies that the average numbers related to tags in our experiments are likely to be larger in practice if non-common tags are included in the calculation, particularly if no thresholding is applied to remove "tag noise", i.e. tags associated only once or twice with a URL. The reason for retrieving just the common tags of a document instead of all tags, i.e. even rarely used ones, is due to technical restrictions. We are working on enhancing our data mining tools for del.icio.us so that we can retrieve all tagging information of a document in the future. But even if all tagging information was available, it would be recommended to perform some sort of thresholding or preprocessing anyway. This means that a potential drawback of using common tags is mitigated in practice, particularly when conducting a study on a larger scale.

3.4 Popularity (Google)

For our study, we consider a web document's PageRank as returned by Google as a measure of its popularity in the WWW. PageRank [2] is a link analysis algorithm which assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of estimating its relative importance within the set. The Google PageRank is a score between 0 and 10, where higher numbers denote higher popularity. A PageRank of zero, however, does not necessarily denote a very uninteresting web page; it is a special value that can have several and different meanings, some of which are:

- the PageRank value is not yet calculated because it is a new web document in Google's search index
- the document has been banned by Google (e.g., a spam page)
- the document is considered as duplicate content

We used the Google SOAP Search API[11] to retrieve official Google PageRank information for each document in the data set.

4. RESULTS

The total data set consists of 97,578 documents, starting from an initial random sample of 100,000 documents. 14.1% of the documents in the data set have been bookmarked by users (the number of bookmarks is equivalent to the number of users who have bookmarked a document), and 5.1% of all documents are tagged. Details are shown in table 2.

Total documents	97,578	
Total bookmarks	180,246	
Bookmarked documents	13,771	14.1%
Total (common) tags	25,311	6,090 unique
Tagged documents	4,992	5.1%

Table 2: Overall statistics of the data set.

The probability of a bookmarked document to have at least one tag is 36.2%. A possible explanation for this relatively low number could be that most new users start using del.icio.us by exporting their existing bookmarks from their browser applications and importing the data to del.icio.us on first use of the service. Traditionally, most popular browser applications have not provided means to add tags to bookmarks[12], so in the past there was no additional information available when importing bookmarks. The *import-at-first-use* assumption is backed by the relatively high occurrence of the tag `imported`, which is by default automatically added to bookmarks on import by del.icio.us: `imported` ranks as #18 of all tags in our data set (see table 6). Recently, del.icio.us and other social bookmarking services added the option to automatically add popular tags to bookmarks on import to mitigate the "cold start problem" for bookmarks without tags.

Statistics per document	**Mean**	**std. dev.**
Bookmarks	1.85	47.68
Tags	0.26	1.80
PageRank	3.13	1.66

Table 3: Statistics per document in the data set.

The average Google PageRank of a document in the data set was about 3 (of a maximum of 10). Details are shown in figure 3. The relatively high variances for the number of bookmarks and tags per document in table 3 suggest that looking only at global mean values is not recommended and a more granular analysis is required. We have therefore differentiated documents by PageRank and analyzed each set individually in addition to the total set of all documents.

[9]For example, the label test for a web document can fail because of network connection problems to the corresponding web server at the time of the test.

[10]http://blog.del.icio.us/blog/2006/09/million.html, last retrieved on May 25, 2007

[11]http://code.google.com/apis/soapsearch/. As of December 5, 2006, Google is not longer issuing new API keys for using its SOAP Search API.

[12]Users of Mozilla browsers have the option to add keywords to bookmarks. However, the keywords field is not presented to the user in the standard bookmarking dialog window, and the user has to manually update a bookmark in a later step in order to add any keywords.

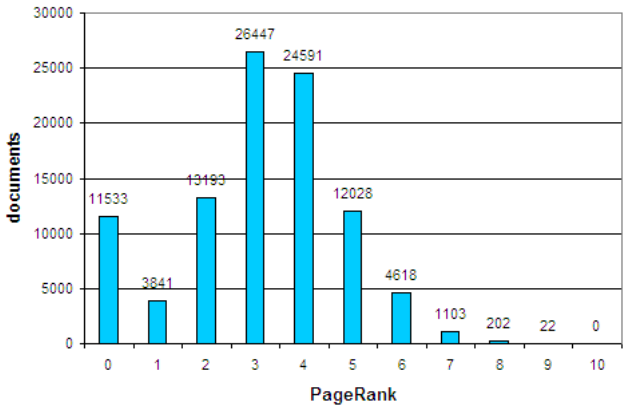

Figure 3: PageRank distribution in the data set.

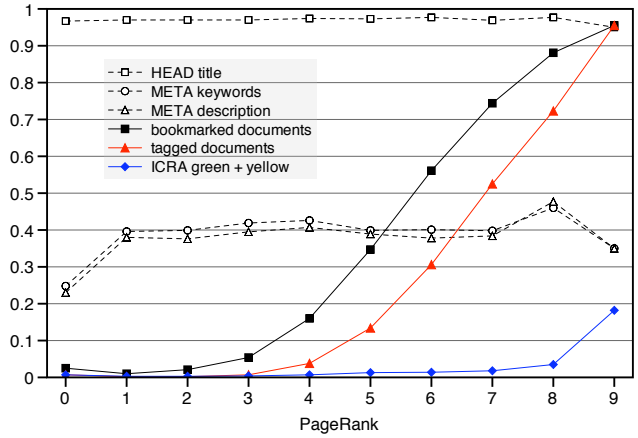

Figure 4: Availability of traditional document metadata and content rating information by authors compared to the graphs for bookmarked and tagged documents based on social bookmarking by end users. For example, 72.3 % of PageRank 8 documents are tagged and 46.0 % of them have META keywords.

4.1 Document metadata from authors

4.1.1 Title, keywords, description

Figure 4 shows the document metadata provided by authors. The first observation is the relatively stable frequency of available metadata throughout all PageRanks. The most frequently used element is the HTML TITLE element with a global average occurrence of 97.14%[13]. The META keywords property slightly outperforms META description with the exception of the small peak at PageRank 8, which is in line with the results of [7]. Both keywords and description occur with a frequency of around 40% in our data set, with two exceptions: web documents with PageRank 0 have a lower frequency of around 25%, and web documents with PageRank 8 have a higher frequency of around 47%.

A possible explanation of the lower frequency of keywords and description in the set of rank 0 web documents is that the reason why these documents have been assigned a PageRank of 0 in the first place is because of improper composition, faulty document structure or markup (the missing META keywords and META description properties being indicators for this). In other words, the reason why we observe a higher-than-average number of "problematic" web documents for PageRank 0 might be that Google intentionally assigns a PageRank of 0 to such documents. The frequency peak at PageRank 8 is harder to explain, and we are unsure how to interpret it. A second test confirmed the peak but the results suggest that more research is necessary for a final conclusion[14].

An interesting observation of figure 4 is the rather constant frequency of author-supplied metadata throughout all PageRanks compared to the frequency of user-supplied metadata which increases with a document's PageRank. Bookmarking or tagging information is significantly more likely to be available for popular web documents than for unpopular ones (more on this later).

4.1.2 ICRA

ICRA content rating information is not widely spread as shown in table 4 and figure 4. Generally, web documents with a higher PageRank are more likely to include digital content labels but the absolute frequency is still very low throughout all PageRanks. Less than 0.1% of web documents in our data set get a "green" test result when looking at individual PageRank slots, with the exception of PageRank 8 web documents with 2.5%. These findings are similar to our previous results from 2005 [12].

red	95.9%
yellow	0.6%
green	0.1%
error	3.5%

Table 4: Results of the content label test for all documents in the data set.

When we combine "yellow" and "green" test results as shown by the blue line in figure 4, the frequencies per PageRank increase slightly for all PageRanks except the higher ranks, which show a larger improvement. PageRank 9 web documents show the highest combined frequency of 18.2%, all of which is contributed by "yellow" test results. This observation suggests that even though highly popular websites seem to be more aware of content rating systems and are more likely to invest time in rating and labeling their content, the content depicted on the web documents is often not fully covered by its labels. One of the reasons we have encountered in practice for this discrepancy is the syndication of content from external partners, e.g. advertisements, over which the original website has no direct control. Document authors should therefore verify prior to publication that the final document including all external references and

[13]This result seems to be in line with the results of [7]. While exact numbers are not given, "the overwhelming majority of pages specify [the title element]".

[14]We built a different sample of 341 PageRank 8 documents which showed a similar trend with a result of of 49.85 % for both keywords and description where each document with keywords also had a description and vice versa. A similar test for 105 PageRank 9 documents resulted in 46.7 % for both keywords and description with the same correlation between availability of keywords and description.

content, i.e. as it is seen by the reader on the live website, is correctly and fully labeled. In previous works [12], we also analyzed the trustworthiness of content labels. We found discrepancies in 18.5% of labels, i.e. label and content did not match, suggesting that it is not advisable to blindy trust in available content labels. The assumption of proponents of content rating systems that every fully rated ("green") website can be correctly classified by its content label is not true in practice, which further lowers the potential usefulness of current content rating systems like ICRA.

In summary, today's content rating systems are a good first step. But their actual usage in practice is negligible, and even if rating information is available, it cannot be trusted blindly - contrary to popular belief. An alternative approach to web filtering is described in one of our previous works [13]. Social web filtering allows the community of end users to collaborate and to provide metadata about web pages by contributing tags and tag votes. This information can then be used to allow or block access to online content according to a user's personal preferences, e.g. to block access to a web page tagged as `porn` by more than 80% of users.

4.2 Document metadata from end users

When designing systems which rely on metadata from social bookmarking and tagging, an important factor to know is the likelihood of a web document to be bookmarked or tagged. Figures 4 and 5 show the metadata supplied by end users. We consider a document as bookmarked if there exists at least one bookmark for the document (same for tags). Since social bookmarking systems like del.icio.us do not allow users to tag a web document without bookmarking it, the number of bookmarked documents is always greater than or equal to the number of tagged documents, i.e. for our experimental set D of documents:

$$\forall d \in D : P_d(bookmarked) \geq P_d(tagged)$$

Our experiments show that the more popular a web document is with regard to its PageRank, the more likely the web document is to be bookmarked or tagged. The detailed results are listed in table 5 (see also the solid black and red lines in figure 4).

PR	bookmarked		tagged	
0	288	2.5%	85	0.7%
1	38	1.0%	6	0.2%
2	275	2.1%	23	0.2%
3	1,435	5.4%	179	0.7%
4	3,945	16.0%	924	3.8%
5	4,178	34.7%	1,614	13.4%
6	2,592	56.1%	1,415	30.6%
7	821	75.5%	579	52.5%
8	178	88.1%	146	72.3%
9	21	95.5%	21	95.5%
10	-	-	-	-

Table 5: Numbers of bookmarked and tagged documents and their relative frequencies in the data set by PageRank. For instance, 2592 documents with a PageRank of 6 were bookmarked, which is 56.1% of all PageRank 6 documents in our data set.

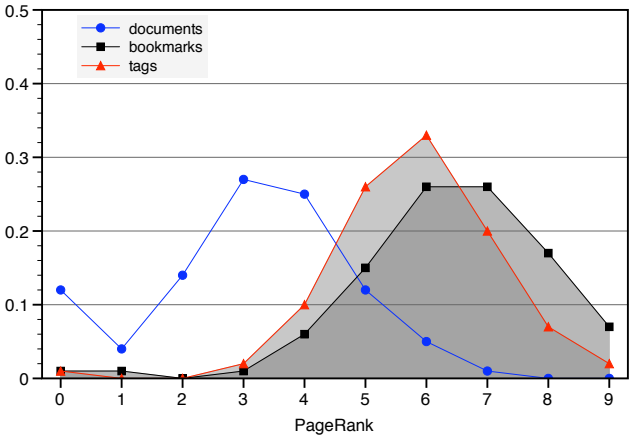

Figure 5: Frequency of metadata supplied by end users via social bookmarking and tagging by PageRank. For instance, 32.8% of all tag annotations in our data set were applied to documents with a PageRank of 6.

In the previous paragraphs, we differentiated between documents without a bookmark and documents with at least one bookmark, i.e. between non-bookmarked and bookmarked documents. A different aspect is the number of bookmarks and tags per PageRank in relation to all bookmarks and tags. Here, we are looking only at bookmarked or tagged documents: we are interested in analyzing where end users create most of the bookmarks and tags, and thus where most of the user-supplied metadata will be available. The results are shown by the black and red lines in figure 5.

The main tagging action focuses on web documents with a PageRank between 5 and 7: 78.7% of the tagging is applied to only 18.2% of web documents in the data set. We see a similar situation for bookmarks: 84.1% of bookmarks are applied to only 18.4% of web documents. Tagging is slightly shifted towards lower PageRanks with an average PageRank of 5.8 compared to bookmarking with an average of 6.4.

A popular argument for the power and success of the "Web 2.0" in the context of information retrieval is that user collaboration and contribution helps to tackle the "long tail": while search engines and IR algorithms may fail to retrieve "rare gems" from the mass of available web documents, humans could augment the search for relevant content by their insider knowledge. For example, even though an interesting web page might not be indexed or ranked high enough by search engines, word-of-mouth propaganda between users through email, social bookmarking or other means could eventually direct visitors to it (a common scenario for the blogosphere).

Our findings however suggest that users tend to focus their bookmarking and tagging activities on popular pages and less on unpopular ones. The majority of web documents in our data set, 52.3%, has a PageRank of 3 or 4 but receives only 7.0% of all bookmarks and only 11.5% of all tags. It is therefore questionable whether the current use of social bookmarking techniques can help to bring order to the web in this respect - at least from a global point of view.

The results could also be an indication that the ranking models and algorithms of search engines like Google are

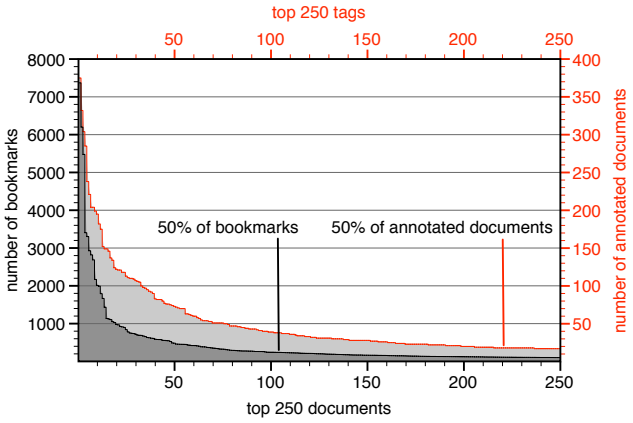

Figure 6: Top 250 documents in the data set sorted by number of bookmarks of the corresponding document, and top 250 tags in the data set sorted by number of documents annotated with the corresponding tag.

quite capable of matching the readers' taste for interesting and relevant content even though models such as PageRank are based on "metadata" supplied by document *authors* (e.g., by regarding hyperlinks to other documents as an indication of their importance and popularity) and not the readers. However, we have to note that with regard to analyzing the correlation of PageRank and social bookmarking, it is also possible that because a document has been bookmarked or tagged by an increasing number of users over time, its initially low PageRank increases as a result. Our experimental data did not include historical information of a document's PageRank, and thus we could not verify this claim in the present paper and have to leave it up to future research.

4.3 Top tags and top bookmarks

We define the *tag count* of a tag as the number of documents that list the tag as one of their common tags. For example, if the tag `document_engineering` is a common tag for 123 documents, its tag count is 123. The bookmark count for a document is simply the number of bookmarks of the document's URL.

The top twenty tags in the data set are shown in table 6. These tags are rather general than specific terms, an observation similar to the one described in [3] (cf. section 4.5).

The top 5.0% of all tags account for 55.1% of the total tag count of 25,311 (see figure 6). 63.3% of all tags had a tag count of only 1, i.e. they were listed only once as a common tag for a document. The top 5.0% of all bookmarked documents account for 74.7% of the total bookmark count of 180,246 (see figure 6). 47.4% of all bookmarked documents had a bookmark count of only 1, i.e. they were bookmarked only once.

The results shown in figure 6 are similar to the findings in [5] for the number of tags in each *user's* tag list: there, a power law graph showed a small percentage of users with very high numbers of unique tags used to annotate their bookmarks followed by users with less and less amounts of unique tags in their vocabulary. Our study suggests that the distributions of bookmarks and tags for *documents* have similar characteristics.

Pos#	Tag name	Pos#	Tag name
1	reference	11	programming
2	software	12	shopping
3	news	13	education
4	music	14	research
5	design	15	history
6	web	16	science
7	tools	17	games
8	art	18	imported
9	blog	19	internet
10	travel	20	fun

Table 6: Top 20 tags in the data set.

4.4 Matching metadata of authors and document content with metadata of end users

We were interested in finding out whether and how much metadata provided by human end users could be derived from a document through direct, automated text extraction. For this, we matched user-supplied tags of a document against the textual content of its TITLE, META keywords, META description, and BODY elements. We also matched tags against the combination of these elements, the results of which are listed as *combined* in table 7 and figure 7. The BODY of a document was cleaned from any HTML markup such as `<h1>` or `<strong>` so that only the actual text content remained. All matching was done case-insensitive. Words in the document as well as tags were splitted at whitespaces and the characters `,.:_#/!?` so that the tag "information_retrieval" would be translated to "information retrieval". The list of special characters used in our study includes the most frequent separators for compound tags found on del.icio.us [18]. This pre-processing step for tags noticeably increased the matching frequency; for instance, the results for "combined" increased from 46% to 58% in table 7, suggesting that special handling of tags can yield significant performance improvements for applications.

BODY	48.9 %
BODY first half	42.9 %
BODY first 2500 characters	41.9 %
HEAD title	25.3 %
META keywords	22.6 %
META description	15.2 %
combined	58.4 %

Table 7: Matching a document's user-supplied tags with its content and metadata supplied by authors (for the complete data set).

Tags appear much more frequently in the body of a document than in its traditional metadata as shown in table 7. Our experiments also suggest that it can be sufficient to parse only the beginning of a web document compared to parsing its full content for getting a relatively high percentage of matches. In a practical setting, savings of 50% in terms of network bandwidth (because only the first half of a

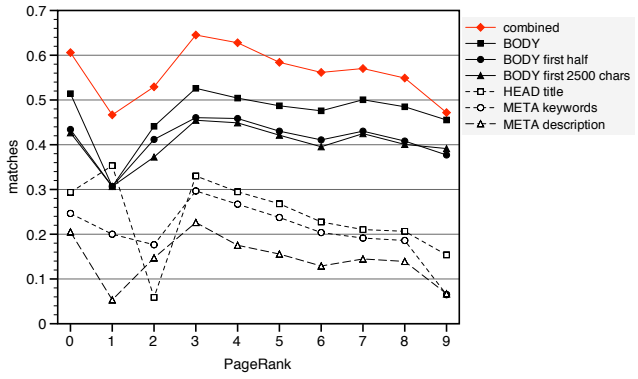

Figure 7: Matching a document's user-supplied tags with its content and metadata supplied by authors (by PageRank). Results for PageRanks 1 and 2 should be treated with care because of the low volume of available tagging information for these documents.

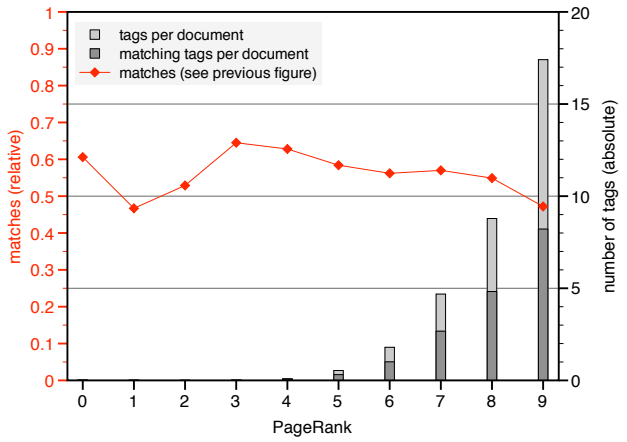

Figure 8: Average number of tags per document compared to the expected number of these tags matching the document's content or author-supplied metadata. The maximum number of (common) tags per document is 25.

document is downloaded) decreases the matching frequency by only 12%: 49% for the full document vs. 43% for the first document half only.

Matching frequency in a document's actual content, i.e. the body, is relatively stable throughout PageRanks of 3 and higher whereas we observe decreasing matches for metadata, i.e. title, keywords and description, with increasing PageRank as shown in table 7. It seems that the more popular a website is, the less does its metadata reflect the perception of the users as expressed by the tags used to annotate the document. The high derivations for PageRanks 1 and 2 are most probably caused by the low absolute number of available tags as listed in table 5.

Another interesting observation is that META keywords match tags better than META description (23% vs. 15%). Ironically, search engines like Google reportedly discard keywords rather than the description in order to prevent abuse and spamming [14].

Figure 8 shows the average number of common tags per document, and the expected number of such tags matching with a document's content or metadata[15]. Generally, the expected amount of matching tags is less than 60% for all PageRanks. This leads to the conclusion that user annotations are indeed providing additional information which is not already contained within the document. We therefore suggest that collaborative tagging could be very useful to augment and improve document classification and retrieval tasks in the WWW.

The figures and numbers above are also an indicator of how well an author's intention of a web document she or he has created - expressed in its title, keywords, description, and body - matches a reader's perception - expressed in tags. Analyzing tags and how they match documents or relate to user click streams derived from web server log files could help authors with optimizing and promoting their content and serve as an additional metric for evaluation tasks. We should bear in mind however that an author's incentive when creating the document and its metadata might be different from the incentive of a reader when bookmarking and tagging it: the former might aim at search engine optimization, while the latter might want to annotate the document for classification, sharing, and later retrieval from his or her personal bookmark collection. For this reason, it could be helpful to infer more information about individual tags. A first step could be to distinguish the different classes of tags such as the three types described in [17]. Knowing that `toread` is a rather *personal tag* could be useful to decide how much weight should be assigned to the tag for a specific task: while `toread` might not help with a general interpretation of the document at hand, it might be valuable for individual users in personalization scenarios or a group of users in a social network. Similarly, one can reasonably assume that a personal tag like `toread` will most likely not match a document's content.

4.5 Most popular non-matching tags

We wanted to find out which are the most popular tags used to annotate web documents but which are *not* in the documents' content or author-supplied metadata. We were also interested to see whether these tags change on different PageRanks. Non-matching tags should give useful insights in which kind of data or information the readers of a document are using to classify and describe it which is not already contained within the document. For this, we modified the traditional TF-IDF formula [16] as follows for ranking non-matching tags for each PageRank.

We define the function $t \dashv d$ to be true if and only if document d is annotated with tag t. We define the function $t \sim d$ to be true if and only if tag t is matching the content or author-supplied metadata of document d.

$$t \triangleleft d := t \dashv d \wedge t \not\sim d$$

$$TF_r(t) := \frac{|\{d \in A_r : t \triangleleft d\}|}{\sum_k |\{d \in A_r : t_k \triangleleft d\}|}$$

[15]Here, "content" means the body with all HTML tags removed from the document source, and "metadata" comprises TITLE, META keywords, and META description.

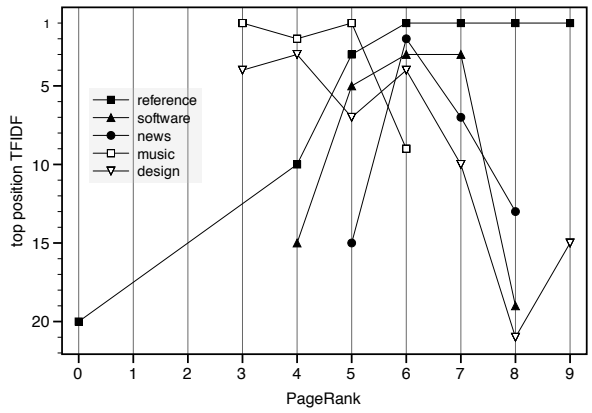

Figure 9: The (modified) TFIDF rank of the overall top 5 tags by PageRank. Ranks below the top 25 are cut off and not shown in the graph.

$$IDF_r(t) := \log \frac{|A|}{|\{d \in A : t \triangleleft d\}|}$$

$$TFIDF_r(t) := TF_r(t) * IDF_r(t)$$

where A is the set of documents $d \in D$ which are annotated with at least one tag ($A \subseteq D$) and A_r is the set of documents $d \in A$ with a PageRank of r ($A_r \subseteq A$). The most important differences between the traditional and the updated formula are that the frequencies are based on *non-matching* terms (here: tags) and that term frequencies are not calculated for individual documents but for sets of documents[16]. The results are shown in table 8 .

We can see that the most popular non-matching tags are rather general terms like `software` instead of being very specific or focused like `compiler`. This result may be surprising because it was unclear so far whether user-supplied information which is not already contained within documents would be either more targeted or more general (or both). Brooks and Montanez [3] found that human-assigned tags - without distinguishing between matching and non-matching tags - produce broad categories for documents whereas automated tagging based on traditional TF-IDF formula, i.e. extracting the top TFIDF-scored words and using them as tags, create smaller and more focused topical clusters of documents. Our results suggest that even the additional information provided by non-matching tags, i.e. user-supplied information to which text extraction techniques have no access, cannot help much with the retrieval of documents on a specific subject. One conclusion would be that when designing an information retrieval system, one should not rely on human tagging to improve *global* recall performance for finding the needle in the stack. On the other hand, the user study of Ames and Naaman for ZoneTag [1], a photo sharing service, found that "social" motivations were the most common motivation for tagging in their scenario. We therefore argue that tags might have more intrinsic value than just broad categorization in the case of smaller networks of connected users who know each other well, because for a specific group of users, even a generally broad term can have a very specific meaning and interpretation.

[16]It would make no sense to make *per document* calculations with non-matching tags in this context. All non-matching tags of a document have per definitionem the same term frequency for the document: 0 (zero).

PR	Ranks from 1 to 5
0	kids, art, books, tutorial, internet
1	searchengine, wlan, disability, adsense, selector
2	ranma, iceland, sussex, turbomachinery, novel
3	music, rpg, gourmet, design, food
4	shopping, music, design, photography, travel
5	music, shopping, reference, fun, software
6	reference, news, software, design, tools
7	reference, imported, software, education, tools
8	reference, tools, imported, safari_export, tech
9	reference, college, tools, searchengine, opensource

Table 8: Top ranked non-matching tags by (modified) TFIDF per PageRank. Results for PageRanks 1 and 2 should be treated with care because of the low volume of available tagging information for these documents.

5. CONCLUSIONS

In this paper, we conducted a quantitative and qualitative analysis of metadata and information provided by the authors and publishers of web documents compared with metadata supplied by end users for the same content. We created a publicly available data set for research based on a random sample of 100,000 web documents from the Open Directory and data retrieved from the social bookmarking service del.icio.us, the content rating system ICRA, and the search engine Google. The most important results and findings of our study are listed below.

1. Availability of traditional document metadata is relatively stable throughout all PageRanks. The opposite is true for or user-supplied metadata: popular web documents are much more frequently bookmarked or tagged than less popular documents.

2. Content rating information based on digital labels provided by document authors for filtering and restricting access to web content is practically not available in the WWW today. We recommend not to rely on content rating systems for such tasks as long as the awareness for and usage of such frameworks has not increased in practice.

3. Within the set of bookmarked/tagged documents, most bookmarks/tags are concentrated on a relatively small subset of documents. Techniques such as PEBL [21] might prove interesting to infer more information about documents with few or no tags by starting from the set of documents with a high number of tags.

4. Tagging is slightly shifted towards lower PageRanks than bookmarking. The mean PageRank is 5.8 for tagged documents and 6.4 for bookmarked documents.

5. Distributions of bookmarks and tags for documents show a power law curve. This observation is similar to the findings of [5] for users.

6. Tags provide additional information which is not directly contained within a document. We therefore argue that integrating tagging information can help to improve or augment document classification and retrieval techniques and is worth further research. Our results suggest that in general, tags may help more with broad categorization of documents than with specific categorization. For example, tagging information could be useful for disambiguation of search keywords and queries in the case of web search personalization.

7. Popular tags are rather general terms. While this result by itself may not be surprising, we could observe that this finding is also true for those tags which provide information that is not already contained within a document.

8. Tags are matching a web document's content (body) significantly better than its metadata (title, keywords or description). Additionally, matching frequency for a document's metadata decreases with higher PageRank. Future research could investigate under which circumstances metadata supplied by readers is a more valuable or trustable resource.

9. For document metadata, the highest matching frequency is between tags and a document's title, which is slightly higher than the matching frequency for META keywords. Both title and keywords match significantly better than META description. Still, the overall matching frequency of metadata is relatively low when compared with the document's content, which might be the result of non-optimal metadata strategies of document authors. Whether this is the cause or the effect (or both) of today's search engines mostly ignoring document metadata is hard to tell.

10. If efficiency of network bandwidth is an issue for practical applications, it can be sufficient to inspect only the beginning of web documents without a big impact on matching frequency between tags and documents.

11. Proper pre-processing of tags, e.g. special treatment of characters used as separators for compound tags, yields significant improvements with regard to matching frequency. Techniques for identifying the type or class of a tag like `@toread` might further help with tag analysis, e.g. for assigning different weights to tags depending on the task at hand. Research in the area of tagging should therefore not only look at identifying synonyms, heteronyms etc. but also analyze how individuals actually use tags "in the wild".

6. REFERENCES

[1] M. Ames and M. Naaman. Why we tag: Motivations for annotation in mobile and online media. In *Proceedings of CHI '07*, 2007.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of WWW '98*, pages 107–117, 1998.

[3] C. H. Brooks and N. Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *Proceedings of WWW '06*, pages 625–632, 2006.

[4] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proceedings of WWW '98*, pages 65–74, 1998.

[5] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

[6] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *Proceedings of VLDB '04*, pages 271–279, Toronto, Canada, 2004.

[7] I. Hickson. Google: Web authoring statistics, http://code.google.com/webstats/. Technical report, Google, Inc., December 2005.

[8] M. J. Jones and J. M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46(1):81–96, 2002.

[9] M.-Y. Kan. Web page categorization without the web page. In *WWW*, pages 262–263. ACM, 2004.

[10] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proceedings of HT '06*, pages 31–40, 2006.

[11] A. Mathes. Folksonomies - cooperative classification and communication through shared metadata. Technical report, UIC, 2004.

[12] M. G. Noll and C. Meinel. Web page classification: An exploratory study of internet content rating systems. In *Proceedings of HACK '05*, Luxembourg, 2005.

[13] M. G. Noll and C. Meinel. Design and anatomy of a social web filtering service. In *Proceedings of CIC '06*, pages 35–44, Hong Kong, 2006.

[14] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *Proceedings of WWW '06*, pages 83–92, Edinburgh, Scotland, 2006.

[15] H. A. Rowley, Y. Jing, and S. Baluja. Large scale image-based adult-content filtering. In *1st int'l Conference on Computer Vision Theory*, 2006.

[16] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[17] S. Sen, S. K. Lam, A. M. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, F. M. Harper, and J. Riedl. tagging, communities, vocabulary, evolution. In *Proceedings of CSCW '06*, pages 181–190, 2006.

[18] E. Tonkin and M. Guy. Folksonomies: Tyding up tags? *D-Lib Magazine*, 12(1), January 2006.

[19] J. Varghese, R. Krishnan, Y. U. Ryu, R. Chandrasekaran, and S. Hong. Filtering objectionable internet content. In *Proceedings of ICIS '99*, pages 274–278, 1999.

[20] Y. Wang, W. Wang, and W. Gao. Research on the discrimination of pornographic and bikini images. In *Proceedings of IEEE ISM '05*, pages 558–564, Washington, DC, USA, 2005. IEEE Computer Society.

[21] H. Yu, J. Han, and K. C.-C. Chang. Pebl: positive example based learning for web page classification using svm. In *Proceedings of SIGKDD '02*, Canada, 2002.