Tags are not metadata, but "just more content" – to some people*

Bettina Berendt
Institute of Information Systems
Humboldt University Berlin
Berlin, Germany
http://www.wiwi.hu-berlin.de/~berendt

Christoph Hanser
Institute of Information Systems
Humboldt University Berlin
Berlin, Germany
http://christoph.usegroup.de/research

Abstract

The authoring of tags – unlike the authoring of traditional metadata – is highly popular among users. This harbours unprecedented opportunities for organizing content. However, tags are still poorly understood. What do they "mean", in what senses are they similar to or different from metadata? Different tags support different communities, but how exactly do they reflect the plurality of opinions,what is the relation to individual differences in authoring and reading? In this paper, we offer a definition and empirical evidence for the claim that "tags are not metadata, but just more content". The analysis rests on a multi-annotator classification of a blog corpus using the WordNet domain labels system (WND), the development of a system of text-classification methods using WordNet and WND, and a quantitative and qualitative comparative analysis of these classifications. We argue that the notion of a "gold standard" may be meaningless in social media, and we outline possible consequences for labelling and search-engine development.

Keywords

Blog tags, Text categorization, semantics, metadata, human-computer interaction.

1. Introduction

Tags play a key role in social media in general and Weblogs in particular. Tags are created both by authors of the tagged content (e.g., as information added to a blog text) and by people who re-use that content (e.g., in sites like `www.flickr.com`, `del.icio.us` or `www.bibsonomy.org`). Tags are used by people who read social-media content, and by machines that process it (e.g., `www.technorati.com`'s search engine). The vast popularity of tags however still leaves many questions open: How can they work as metadata when they do not come from a controlled vocabulary of limited size and with well-defined meaning, but from a huge and dynamically changing pool of words and phrases? Why is tag creation so popular, when metadata generation in general is one of the least popular activities for people not paid for it?

It has been suggested that tags are in fact not metadata, but "just more content". But what does this claim mean?

(Q1) Tags are different from traditional classification keywords in the sense that they are different *text/world model classes*. This idea has been investigated in the literature on the differences between folksonomies and traditional taxonomies or ontologies, e.g, [18]. However, are tags (the model) also different from metadata *in their relation to the underlying content*?

(Q2) Can these differences contribute to an explanation of the popularity of tags for both readers and authors? May they give us an indication of how tags and tagged material are read, and maybe how they are written? Can we find out about individual differences between social-media users?

(Q3) Can investigating these questions deepen our understanding of "what content really is about", including the questions of how – and whether – gold standards for determining the quality of content classifiers can be established?

(Q4) What does this imply for search engines working on tagged material? How can they be improved to satisfy individual users on the one hand, and groups on the other?

(Q5) Last but not least, which computational and empirical methods are suited to analyzing these questions?

This paper addresses these questions by an empirical analysis that uses (a part of) a standard blog corpus: the corpus offered by Blogpulse for the *Weblogging Ecosystem* workshop 2006.[1] The analysis consisted of gathering classifications from different human annotators and from different IR / text mining methods and semantic resources, and of quantitative and qualitative analyses of their outputs. In addition, a comparison with results on non-social-media data (the Reuters RCV1 news dataset) was performed.

The contributions of this paper are answers to the five questions: First, we show evidence that tags are indeed "just more content", and we argue that the essential characteristic is a difference in the tag–text relationship (complementary rather than summarizing/disambiguating, as in traditional metadata). Second, we show that this assessment may depend on the reader – it seems that some readers take the tags to be an indicator of content, while others more or less ignore them. Thus, tags seem to be "just more content" to some people only. Third, we argue that this highlights the importance of deviating from the monolithic notion of "content"

*We thank Roberto Navigli and Anke Lüdeling for many helpful discussions, and four anonymous reviewers for excellent suggestions and questions.

[1] Thus, our results refer to tagging systems in which authors tag their self-produced, textual content. Future analyses of different tagging systems, e.g. comprising third-party tags and pictorial content, will serve to generalize or differentiate these findings.

that is inherent in many text-analysis methods, and that alternatives to gold standards are needed to capture these semantic intricacies (which may characterize social media to a larger extent than other material). Fourth, we outline implications for search-engine design. Fifth, we critically review our research method, and outline future research.

2. Related work

Marlow, Naaman, boyd, and Davis [14] distinguished between different system designs and architectures, classifying for example by object type (textual, non-textual), source of material (user-contributed, system, global), and tagging rights (self-tagging, permission-based, free-for-all). They also identified different user incentives (e.g., play vs. retrieval). They suggested that system features and user motivation are important determinants for the tag types that emerge, referring to the tag types or functions proposed by Golder and Huberman [6]: identifying what (or who) the tagged item is about, identifying what it is, identifying who owns it, refining categories, identifying qualities or characteristics, self reference, and task organizing. Another important distinction between tagging systems is whether tags are provided *by* self or others ($\sim$ tagging rights) and *for* self or others [7].

Both [6] and [14] empirically measured tag usage, focusing on tags-as-tokens and counting the number of different tags per tagger, the frequency of usage of a tag, etc.; both point out that their findings may, to some extent, be specific to the tagging systems they investigate (`del.icio.us`: Web resources, self-contributed/global content, free-for-all; `flickr`: pictures, self-contributed, mainly self-tagging).

Author tags have been investigated with respect to their ability to describe a blog post's semantics [3]. The authors found that tags are useful for grouping articles into broad categories, but less effective in indicating the particular content of an article. They then used automatic extraction of highly relevant words (the 3 words from an article with the top TF.IDF scores) to produce a more focused categorization of articles than that based on tags.

Tags are generally freely chosen by the user (although tag-recommender systems encourage the re-use of existing tags). Their technical realization and the grouping of blog posts into the tag directory is supported by blogging software. The blog tagging site `www.technorati.com` currently (Feb. 2007) tracks 13.6 million tags, with the number increasing by about 0.5–1 million per month. These tags are not part of a controlled vocabulary, and they are not aggregated into a hierarchy. (Compare this with the current state of the `www.dmoz.org` directory: more than 70,000 editors have categorized more than 4 million sites, agreeing on below 600,000 categories organized into a hierarchy with only 16 top-level categories.)

Such *folksonomies* (collaborative categorization using freely chosen keywords) differ from established classification methods in library science that rely on controlled vocabularies, taxonomic organization, and/or given analytical facets. Folksonomies appear to be superior for fast-changing domains with unclear boundaries [18]. However, the advantages hinge on large numbers, of people and of documents. They also appear to require high levels of coherence and interaction between their users. It is not clear how to make effective use of an unstructured plethora of tags, and spam blogs are beginning to threaten tag validity.

Brooks and Montanez [3] observe "that bloggers are not settling on common, decentralized meanings for tags; rather, they are often independently choosing distinct tags to refer to the same concepts" – and they remark that "whether or not the meanings of these distinct tags will eventually converge is an open question."

The large number of tags suggests that they resemble natural language more than the controlled, limited-size vocabularies traditionally used for classification; in other words, that they are "just more content". It is unclear, however, what exactly this means. A simple (fictitious) blog can serve as an example of one possible meaning:

> *Tags:* Robbie Williams
> Concert in Chicago.
> This gig was awesome.

We complement and extend prior research in the following ways:

(1) We concentrate on blog tags as a third important class of tagging systems (textual content, self-contributed, self-tagging).

(2) We provide a quantitative and qualitative semantic analysis of blog and tag content, and automated methods for the quantitative analysis part.

(3) We propose a new setting that supports the analysis of tags as a means of communication: We relate author tags (written by self for others) to reader annotations (written by others and, by our instructions, for others). To make the empirical study feasible and the different sources of semantics comparable, we had to fix certain parameters: the material is user-contributed by the authors and system-supplied for the readers. We asked readers to choose from a controlled vocabulary, and we mapped all sources of semantics to the same categories. In future work, all these factors should be varied systematically to deepen the understanding of tags.

We use a corpus whose overall characteristics (e.g., tag usage) have been explored in other studies, e.g., [21, 17, 16].

3. Automated blog domain classification

The initial goal of our studies was to assess the quality of different methods for text content classification, derived from standard text mining / IR, to blogs. In particular, we were interested in methods that provide for a "cold start", i.e. that need no labelled data for classifier learning. Semantics-based methods are an obvious choice for this. Such methods could be used to organize an unknown corpus of blogs into a small and manageable number of meaningful categories, which can aid search (and, by extension, also blog authors who want to improve the findability of their content by labelling). Thus, we aimed at augmenting folksonomy-style tagging by more standard ways of assigning metadata.

We used the corpus offered by Blogpulse for the *Weblogging Ecosystem* workshop 2006[2] to refer to a standardized set of texts. We drew a random sample of 100 blog posts from that corpus that were (a) written on 4th July 2006 (the first day of the large corpus), (b) written in English, and (c) tagged by their authors. Choice (a) was made to avoid the likely thematic bias introduced by the London bombings on 7th July, which reverberated strongly in the blogosphere at large and also in the Blogpulse corpus (cf. [21, 17] for investigations of the same corpus that did center on the London bombings). Choices (b) and (c) were made to ensure that standard text-analysis tools would be able to process the data, and that

[2] `www.blogpulse.com/www2006-workshop/#data` .
We thank Natalie Glance for supplying these data.

tags could be analyzed without missing data (only 24.1% of the posts in the large corpus had any tags at all, and only 68% of these tags were useful, i.e. different from entries like "General"). The relatively small number of posts was chosen to ensure high-quality collaboration by the volunteer annotators.

Our basic model of content classification was the assignment of one or more semantic labels to one blog post. We used two resources: WordNet [5], a computational lexicon of English, encoding concepts in the form of synonym sets (*synsets*), and the domain labels from IRST [13], providing a mapping between WordNet synsets and 165 taxonomically-structured domains (e.g. DENTISTRY is a kind of MEDICINE, MEDICINE is an APPLIED SCIENCE, etc.). Notice that one term may belong to different domains, depending on the senses it denotes: For example, sense #5 of "operation" in WordNet is mapped to the domain label MEDICINE, while sense #7, "operation" in the sense of data processing, is associated with COMPUTER SCIENCE, as well as to further domains. The labelling system is based on the Dewey classification, and its hierarchy has 4 levels. The information is complementary to WordNet (which has its own hypernym relations) in that a domain may include synsets of different syntactic categories and from different WordNet sub-hierarchies. Domains may group different senses of the same word together, with the side effect of reducing word polysemy in WordNet.

Each classification method assigned zero, one, or more WND labels to each blog post from the corpus to characterize the blog post's content. The first five "methods" were five human labellers, the sixth was their aggregate judgment, designed to constitute a reference classification against which to evaluate the remaining, automated methods.

3.1 Consensus classification: A "gold standard"?

Five graduate students of Information Systems with a good command of English volunteered to label the corpus. They received a plain-text file containing the corpus entries, a spreadsheet file with the domain tree, and another spreadsheet file for entering their answers. They were given written instructions how to recognize the blog post's domain (using an example) and how to handle the annotation environment. Every annotator was allowed to assign an arbitrary number of domains per post (0–3 were recommended), and they were allowed to choose domains from any position in the hierarchy [12].

This resulted in 500 label sets, of which 23 were empty, 304 contained 1 domain, 160 (12; 1) contained 2 (3; 4) domains. Annotation behaviour proved to be quite heterogeneous: 13 of the 23 empty sets were due to annotator 3, while annotators 1 and 5 labelled all posts. The average number of domains per post varied between 2.08 (annotator 3) and 2.65 (annotator 2). Most of the chosen domains were on the second level of the hierarchy; the occurrence ratios (number of times a domain on the ith level was chosen / number of distinct domains on that level) were 0.83, 8.0, 1.96, and 0.14 for the four levels.

The five annotation sets were then merged to arrive at a *consensus classification* for each post p. It consists of all labels from any of the annotation sets $A_i, i = 1, ..., 5$ that possess at least a "minimal consensus". In this case, we opted for a θ of 0.2, i.e. for each post (including those which were not rated by all 5 annotators), we required at least 2 votes for this domain:

$$consensus(p) = \{d : \frac{|\{A_i(p) : d \in A_i(p)\}|}{|\{A_i(p)\}|} > \theta\} \quad (1)$$

To investigate the "reliability of judgement" [4], an inter-annotator agreement (IAA) was calculated. A further motivation for this computation was to obtain an upper baseline for the expected quality of the automated methods: Given that computers can usually not perform better than people in annotation tasks [25], an automated method can hardly be expected to find the correct domain if even human annotators had problems agreeing.

The standard measure for calculating IAA, kappa [4], even in its adaptation to multiply-labelled data [19], turned out not to be usable, mainly because the probability of chance agreement as well as the probability of total agreement were too low. Instead, for each blog post we used the average similarity, over the 5 annotators, of the individual annotators' label sets to the consensus classification.

We evaluated similarity with different measures. In the current paper, we report only the results using popular and proven Jaccard coefficient, cf. [8]. It defines the similarity between two label sets A_1, A_2 as:

$$sim(A_1, A_2) = \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|}. \quad (2)$$

The average Jaccard coefficient that we use to measure the similarity between two methods is the average, over all blog posts in the sample, of those values.

The resulting IAA was 0.39. The similarities of annotators to the consensus classification differed (between 0.31, annotator 3, and 0.47, 2), as did their pairwise similarities (between 0.26, 2 vs. 5, and 0.4, 1 vs. 4).

3.2 Performance of automated domain classification

The methods were defined as the combination of (a) the extraction of a bag of words (BOW) from the blog post and (b) a word sense disambiguation strategy. BOWs were sets of tokenized and POStagged words (using the Stanford POS tagger, [22]). The disambiguation strategy was based on the selection of one or more WordNet synsets per word as interpretations of this word, and the WND labels assigned to the chosen synset(s).

In this paper, we report on the results of 10 combinations arising from the feature sets: (a1) author tag(s) (tag), (a2) nouns from the title (titleN), (a3) nouns from the blog post body (bodyN), (a4) the top 5 TF.IDF keyphrases (noun n-grams) from the body (TF.IDF), (a5) the top 5 IDF keyphrases (noun n-grams) from the body (IDF); and (b1) the top sense of the word (T), (b2) all senses of the word (A). (a1–a4) are straightforward and standard choices to capture text content. (a5) derives from the observation that because of the rich context of blogs, captured for example in hyperlinked sources, important terms may not actually be frequent in the post itself, such that their being unusual (high IDF) creates a better indicator of importance [10]. IDF was calculated on the corpus of all 429,183 blog posts from the 4th July that were contained in the original Blogpulse corpus. (b1) and (b2) are standard simple ways of disambiguation (e.g., [9]) and have been successfully used together with WND labels to disambiguate blogs in [1].

Each method returns, for each post, a set of domains. First, for each term all WordNet synsets are retrieved. For each

synset, all WND domains to which it maps are retrieved. A disambiguation strategy based on selecting the most frequent domains was developed, which reflects the basic WND disambiguation strategy (see [13], and [1] for a similar strategy for blog disambiguation). This strategy involves the computation of the relevance r of a domain d for a synset s, for a term t, and thence for post p:

$$r(d,p) = \frac{1}{\text{no. of terms in } p} \sum_{t \in p} r(d,t) \quad (3)$$

with

$$r(d,t) = \frac{1}{\text{no. of synsets for } t} \sum_{t \overset{WordNet}{\longrightarrow} s} r(d,s)$$

$$r(d,s) = \frac{1}{\text{no. of domains for } s} \text{ if } s \overset{WND}{\longrightarrow} d, 0 \text{ else}$$

The domains were then thresholded on $r(d,p)$ in order to determine whether they would become part of the method's result domain set or not (similar to equation (1), with the difference that there each annotator's vote has a weight of 1). Note that the number of synsets for a term is always 1 in the "top sense" methods and ≥ 1 in the "all senses" methods. Thus, the terms (and in the "all senses" methods also the synsets) take the role of voters, just as the annotators did in the computation of the consensus classification.

Performance relative to the consensus classification was again measured by the average Jaccard similarity.

The resulting similarities were:[3] tags (T): 0.09, tags (A): 0.11, titleN (T): 0.07, titleN (A): 0.05, bodyN (T): 0.14, bodyN (A): 0.12, TF.IDF (T): 0.09, TF.IDF (A): 0.09, IDF (T): 0.09, IDF (A): 0.08.

This performance looks very poor at first sight. It should be recalled, however, that even the individual human annotators reached at most a similarity of 0.47 (the average was 0.39). Still, the results do show that machine performance was disappointing. Reasons include: (a) Blog posts contain a lot of jargon, misspellings / unusual spellings, grammatical mistakes like omitted sentence modifiers, etc. This resulted in many words that could not be handled properly by one or several of the text processing steps (tokenization, stemming, POS tagging) or that are not entries in WordNet. (b) Some labels are distinct, but stand in a close semantic relationship (e.g., THEOLOGY, which is a subdomain of RELIGION). (c) In a number of cases, annotators gave labels that are different in the WND hierarchy, but that may appear similar to a layperson, especially when they do not consider the hierarchical context of the label (e.g., LINGUISTICS and PHILOLOGY). (d) The hierarchical organization of the labels sometimes obscures existing semantic relationships (e.g., a text about NASA and its stellar observations was classified by one annotator as ASTRONOMY, which is a PURE SCIENCE, by another as ASTRONAUTICS, which is part of ENGINEERING, which is an APPLIED SCIENCE). (e) The consensus often contains a large number of domains coming from different annotators; automated-method result sets tended to be smaller.

We first addressed (e) because we were interested in exactly these disparities. We offer a partial solution approach to (b), and further evidence of (a) and (c), in Section 5.2. We return to (d) in the Outlook.

[3] for a graphical representation, see the "consensus" curve in Fig. 1 below.

4. Different methods and different annotators

In a follow-up study, we investigated whether we would be able to improve on the results and what we could learn from a deeper analysis of the data.

4.1 Tags complement content

The results of Section 3.2 raised the question whether all methods "err in the same way", or whether they make different mistakes.

To investigate this question, we first compared all methods pairwise. The results, shown in Table 1, indicate that methods did err in different ways, but that (a) for each BOW-choice, the choice of disambiguation strategy is comparatively unimportant (these pairs have the highest similarities), (b) the greatest dissimilarity occurs between tag-based methods and nearly all body-based methods.

Finding (b) represents **(A1.1)**: Author tags contain semantically different information than the body of a blog post.

We therefore combined methods that were both dissimilar and (viewed in isolation) relatively good at assessing the human annotators' merged labelling (see Section 3.2). The prediction of the combination of two methods was defined as the union of their respective label sets. The results were:[4] tags (T) & bodyN (T): 0.16, tags (T) & bodyN (A): 0.15, tags (A) & bodyN (T): 0.16, tags (A) & bodyN (A): 0.15, titleN (T) & bodyN (T): 0.15, titleN (T) & bodyN (A): 0.14, titleN (A) & bodyN (T): 0.13, titleN (A) & bodyN (A): 0.11, tags (T) & TF.IDF (T): 0.13, tags (T) & IDF (T): 0.12, titleN (T) & TF.IDF (T): 0.11, titleN (T) & IDF (T): 0.11, tags (A) & titleN (T): 0.12.

The resulting increase in classification quality compared to the single methods, together with the top rank of the tag & bodyN method, suggest[5] **(A1.2)**: Author tags and the body of a blog post contain complementary information. In this sense, they conform to the following

Definition: In a corpus of posts consisting of body elements (text, title, ...) and author tags, the **tags are not metadata but content** if (1) the tags have a low similarity with the body (such that body features cannot be used to predict the tags, or vice versa), and (2) the combination of body and tags predicts the human consensus classification of content better than either body or tags alone.

However, this raises the question *how* tags can be more content: by partitioning information into interdependent components, as suggested by the fictitious example in Section 2, or in some other way. This will be investigated next.

4.2 Tags provide additional information

In some cases, body content was dominated by words that suggested one domain, and the author tag added a meaning component. This is related to the fictitious example in Section 2, but differs from it in that the tag does not really disambiguate (select from possible meanings), but rather adds meaning of the kind "this post is related to ...".

[4] also part of the "consensus" curve in Fig. 1

[5] Due to the relatively small sample and the exploratory nature of the study described in Section 3, we did not perform any tests of statistical significance. Larger samples, confirmatory designs, and significance tests will be the subject of future work (see Outlook).

	tags(T)	tags(A)	titleN(T)	titleN(A)	bodyN(T)	bodyN(A)	TF.IDF(T)	TF.IDF(A)	IDF(T)
tags(A)	.65								
titleN(T)	.23	.14							
titleN(A)	.11	.14	.49						
bodyN(T)	.09	.09	.07	.05					
bodyN(A)	.08	.10	.07	.07	.30				
TF.IDF(T)	.19	.14	.19	.12	.24	.15			
TF.IDF(A)	.09	.14	.09	.11	.14	.20	.36		
IDF(T)	.10	.09	.18	.14	.19	.13	.45	.22	
IDF(A)	.07	.11	.11	.12	.13	.14	.24	.41	.50

Table 1: *Average Jaccard similarities between methods*

Tags: General, Music.
Alpione.com.
Yes! Ive been looking for a way to easily transfer songs on my iPOD to my computer. I want to back all of them up to DVD, but Apple makes it very difficult to pull them off. Thats for copyright purposes, Im sure, but there are legitimate uses for it as well. iPOD Agent, a $15 shareware program, allows you to do that and much more. Synchs contacts/notes/etc with Outlook, gets horoscopes/movie times/weather/RSS feeds. Good stuff. iPOD Soft Go get it! Adam[6]

All methods ignored the "General". Three of the annotators agreed with the blogger that this was about MUSIC, and also added that it was about COMPUTER SCIENCE. The two remaining classified the post as PLAY or FREE TIME. The body methods did not pick out "song", the only word that may indicate music, and classified as COMPUTER SCIENCE and LITERATURE (due to the copyright mentioning). The tag-based methods could obviously find MUSIC. The benefit of combining body information with tag information in such cases reflects the observations of [3] mentioned in Section 2.

In this case, domain knowledge or named-entity recognition of the iPod could have helped the body methods. However, the question would have been whether MUSIC would have been selected even then, because it may be argued that an iPod is a technical device that is at most *related to* music. We will return to this point in the Outlook.

4.3 Tags reflect differences between annotators

In the example shown above, MUSIC was a good prediction for 3 of the 5 annotators. Neither of the automatically found domains corresponded to that of the remaining 2. However, on closer inspection we found that this type of consensus did not occur in many other blogs.

We mentioned above that one difficulty of the (single) methods may be the aggregation inherent in combining the very heterogeneous human annotators' classifications. We therefore analyzed the similarity of each method investigated thus far to each individual annotator's labels.

The results are shown in Fig. 1. They confirm our expectation in that (a) combined methods are better at capturing "consensus classifications", but they also show that (b) different, single methods may be superior to combined methods for individual people. In particular, they indicate that annotator 1 would profit most from a tag-based classification, while annotator 3 would profit most from a body-based classification (specifically, one that focusses on the unusual words that are characterized by a high IDF).

This observation merits a closer, qualitative investigation. We identified all posts for which the similarity between annotator 1 and annotator 3 was 0, and for which the tag+bodyN method (in any of the A/T variants) returned a good prediction value for the consensus classifications.

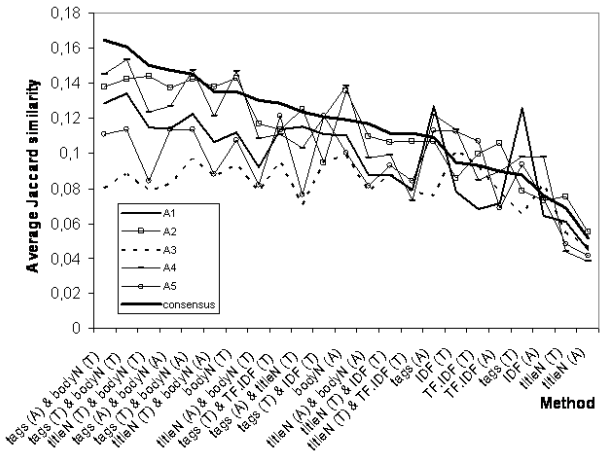

Fig. 1: *All methods relative to consensus and individual annotators.*

Of course, we cannot find out what exactly motivated each annotator's vote. However, we can look for cues in the material. This search led to posts like the following:

Tags: Radio & TV, Islam.
GetReligion.
Todays New York Times includes this report about Sleeper Cell, a 10-part Showtime series about a faithful Muslim named Darwyn (yes, we get it) who infiltrates a terrorist group. The Times mentions the producers goal of high realism, but also must grant that, while some Muslim FBI agents exist , theres no way to know if any such agent has infiltrated a terrorist cell. Still, its easy to sympathize with series star Oded Fehr (pictured), an Israeli actor playing a terrorist, and Cyrus Voris, one of the producers, as they discuss the shows idealism: "You learn there are peace-loving souls in every religion," said Mr. Fehr, who once served in the Israeli military. "We have to respect and strengthen the peace-believers, and hopefully find a way turn the terrorists." In that sense, the production, for all its violence - including the Sopranoesque rubout of a cell member by his fellow crew - is perhaps most ambitious for the idealism that courses through it. "I dont know if a guy like Darwyn is out there somewhere in the U.S.," said Mr. Voris, a creator of the show. "But I sure hope so. Talk about wish fulfillment." Besides which, Showtime has a bit to atone for while it promotes the hostilities of Penn & Teller toward all things religious (including Christo-

[6] permalink: `http://www.alpione.com/?p=114`

pher Hitchens whipping post, Mother Teresa).[7]

Annotator 1 classified this post as RELIGION and TV, while annotator 3 classified it as POLITICS. The tags methods clearly mirror the choices of annotator 1 (religion and (radio and) TV is the only available information for this method). Note that the start of first sentence also only offers cues for these interpretations. In contrast, the IDF method picked out the words "terrorist" and "sleeper", which have clearly political meaning (senses #1 and #2, respectively), as well as the proper names (which do not map to domains) "Teresa" and "Cyrus" and the compound "series star" (which is not in WordNet and was therefore ignored in the domain mapping).

The different results on this blog may also be interpreted as showing that annotator 1 tried to form an opinion on the post's content immediately after starting reading, while annotator 3 read to the end. This conjecture is borne out by their classifications of another post:

> *Tags:* Art.
> Cool Hunting.
> I first wrote about Ludwika Ogorzelec s Space Crystallization Cycle after seeing her show here in NYC last February. Her prolific installation of site specific cellophane lattice has graced a broad range of settings since the series began a couple years ago. The latest... farmland. Farming With Mary is a Queensland Australia project that brought environmental artists from all over the globe to the farming community. Ludwika installed three pieces, each comprised of about 5km of cellophane, on a farm in Tuchikoi in the Mary Valley Region. She also installed one piece in Noosa Woods. Pictures after the jump.[8]

Annotator 1 classified this post as ART, annotator 3 as PHOTOGRAPHY. The choice of ART by both the annotator and the tag-based methods is obvious. There is however only one indicator of photography: the "pictures" in the last sentence – in fact, as the inspection of the post on the Internet shows, the post contains photographs of the cellophane-lattice installations mentioned. (In this case, however, neither TF.IDF nor IDF singled out this word as important. Only the bodyN method had access to it. Annotator 3's interpretation requires knowledge of blogspace authoring conventions.)

This leads us to emphasize **(A3.1)**: Texts have different meanings for different people. This is true outside and inside social media, only it tends to be forgotten when canonical labellings are created and treated as "gold standards".

With regard to the question of how tags add content, we might say that they do not (or not always) add information that is missing in the text, but select or highlight aspects of meaning that for some readers may be less relevant or not even present in the body. This is related to the disambiguating function of metadata, but it is not the same.

In addition, we derive **(A2)**: A comparison of different text-analysis methods and annotators allows us to better understand the tag-text relations, and they may give some indications as to how different people read different texts. This observation also allows us to better understand the popularity of third-party tagging as in `del.icio.us`, and the popularity of tag choice based on community (assuming that "my community" tends to "label as I do").

Another answer is **(A4)**: Exploiting such observations could make search engines and community-finders smarter: the structural understanding of reading styles behind tag preferences could be superior to simple customization or collaborative-filtering options. However, general search engines must cater both for individual users and user groups they have information about, and for users they have no information about (e.g., because these people prefer to stay anonymous). For the latter user type, system output should be geared to satisfying a wide range of users, i.e. ideally reflect consensus classification.

[7] permalink: `http://www.getreligion.org/?p=894`
[8] permalink: `http://feeds.feedburner.com/ch?m=57`

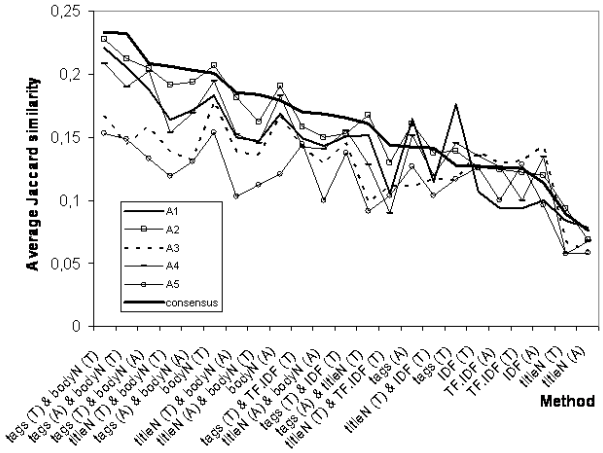

Fig. 2: *All methods (coarsened to level 2) relative to consensus and individual annotators.*

In addition, these findings qualify **(A1)** given above: This observation is in many cases true only for some people.

The improved classification quality and the insights about individual differences still leave us with many open questions regarding the overall predictive quality of the method. In the next section, we will investigate this.

5. Investigations of the method

The analysis of results suggested two basic problems: First, the classification system and the instructions are not ideally suited to the annotators. Second, blog texts have well-known divergences from more official publications (cf. interpretation (a) in Section 3.2). To get a more appropriate picture of the potential inherent in the method itself, we therefore developed and tested some variations.

5.1 What is the basic level of classification?

We asked our annotators to use the label(s) that they thought most fitting for each particular post. This of course invites problems like the classification of one post as RELIGION by one person and THEOLOGY by the other (see interpretation (b) in Section 3.2). It might have been easier for labellers to rate each post from a more restricted catalog.

A good choice for such a catalog would consist of basic-level domain categories [20]. An inspection of the WND hierarchy suggests that for laypeople, level 2 comes closest to being a suitable basic level. For example, in the level-1 domain DOCTRINES, level 2 contains PSYCHOLOGY, ART and RELIGION, and level 3 contains MYTHOLOGY, OCCULTISM, and THEOLOGY (as subdomains of RELIGION). In the level-1 domain APPLIED SCIENCES, level 2 contains COMPUTER SCIENCE, ENGINEERING, ARCHITECTURE, and level 3 contains TOWN PLANNING and BUILDING INDUSTRY (as subdomains of ARCHITECTURE). Level 4 is even

more detailed, and level 1 is too coarse.

Figure 2 is the equivalent of Fig. 1 for level 2. Each method's coarsened label set was created as the union of (a) its original labels that were from the 2nd or 1st level and (b) the 2nd-level superdomains of all its other original elements. This transforms the similarity between the label sets {religion}, {theology} from 0 to 1. The figure shows that prediction quality improves substantially, but that the relative quality of methods and the individual–group phenomena are hardly affected.

Still, the absolute level of predictive quality remains low. The domain labels, in their reliance on the Dewey classification system, may not be the most intuitive classification system for laypersons. A true basic-level category system for blogs remains to be established – and it is unclear whether it will ever be established, because this idea flies in the very face of folksonomy-style tagging.

In addition, although it is generally recommended to use "naive" annotators [4, 11], a manual inspection showed that some people clearly misunderstood some posts or failed to differentiate between domains (see interpretation (c) in Section 3.2). Another problem may have been that the annotators, while fluent in English, were not native speakers.

5.2 Classifying blogs and news

WordNet and the WND hierarchy are – like any other semantic resource – limited by the state of their editing. In particular, such hand-curated resources have problems accommodating emerging domains (for example, "Internet" should be integrated as a subdomain of TELECOMMUNICATION or COMPUTER SCIENCE). Also, mappings may fail to capture the complexity of topics, especially emerging ones. For example, "blogging" might be mapped to COMPUTER SCIENCE, LITERATURE, or TELECOMMUNICATION (WordNet maps it to literature).

The analysis of blogs may be even more sensitive to these problems than that of other text genres. The reasons include syntactical and writing-style features (see interpretation (a) in Section 3.2), the usage of hypertextual and other context for expressing content, and a meandering topic-drift that appears to characterize a number of blogs (annotators called them "unfocused"). As a result, the quality of the tagged POS and the agreement of annotators were often low.

A text category that may eschew many of these problems is news: written by experienced writers (who know how to convey content to readers with limited attention span, and who tend to use common words), focusing on one topic, generally proof-read and syntactically correct. Due to a combination of these features, news routines (such as clear content categories), and the specifics of the best-known news corpus, they also have canonical, agreed-upon domain classifications.

The Reuters news dataset RCV1 [12] originates from 1996. (Thus, it also avoids all Netspeak-related problems.) All news feeds have been manually annotated with 103 topic codes. We tested our basic method on this corpus in order to obtain a "baseline for well-behaved texts" and thus an assessment of how much harder blogs are to categorize than news.

We manually mapped Reuter's topic codes to the WND hierarchy, which is a nontrivial task since domains are different both in their extension and in their structure. Like the WND hierarchy, the topic-code system reflects the purposes of its intended usage context and may be argued to be unbalanced; for example, there are 60 topic codes concerning business, but only one about "science and technology" [12]. Of course, this mapping introduces additional variance into the comparison, but without evidence to the contrary we do not expect a systematic bias, and thus the resulting investigation can still give insights into the usability of the general method.

629 news feeds from August 20, 1996 were randomly selected. Their titles, headlines, texts, topic codes, and their mapped WordNet domains were extracted and stored. The resulting XML corpus was transformed by the procedures described above. The comparison of blogs and news was carried out in several settings; in the following, we just report the one with the best values (all nouns, all senses; in addition, terms were weighted with their TF.IDF weights).

The results showed the expected main effect of text genre, and they replicated the effect of hierarchical coarsening. In addition, an interaction effect could be observed: while micro-averaged F1 rose from 0.23 (blogs) to 0.35 (news) for labels from arbitrary levels, it rose from 0.34 to 0.61 for level-2 coarsening. (These experiments were only evaluated with respect to F1; however other experiments indicate that comparisons based on Jaccard similarity would have given similar results, only with overall smaller numbers.)

The domain ECONOMY (a news-corpus focus, see above) was recognized correctly in 93.7% of cases in the news corpus, but only in 23.5% of cases in the blog corpus. Domains such as HEALTH, RELIGION and ART showed much poorer recognition. This raises questions for future work: which domains are easier to recognize, in which text genre, and how does this interact with the classification hierarchy?

The results indicate **(A3.2)**: blogs are harder to categorize than news; one of the reasons may be (apart from the "text quality" and referential differences) that blog content itself is more fluid and reader-dependent than news texts.

6. Conclusions and outlook

We have presented an analysis of human annotators' labelling of a blog corpus, and the comparison of a semantics-enhanced method for content classification with these human judgements. Our main results are a definition and empirical evidence of the claim that author tags are not metadata, but content. We showed that this means two things: To reflect the opinions of a larger group of readers (and possibly also authors) of blogs, tags and body elements (e.g., nouns) should be combined. However, there appear to be subgroups of users for whom tags *or* body elements are the most suitable indicators of content; they may even largely ignore the respective other elements in their assessment of what a blog is about. Our practical conclusion was that this should be reflected in the design of search engines and labelling recommenders.

We close with a four-part answer to question 5.

(A5.1) Extracting features from different BOWs gives clues about blog post content. In particular, analyzing the author tags and body nouns (potentially weighted) was useful. However, the processing needed to obtain good results (e.g., POS tagging) is often severely hampered by blogs' low syntactical quality or dynamically-changing Netspeak. In addition, the referential context of a blog post (given, for example, by hyperlinks) needs to be taken into account. The advantages of integrating information behind referential links has been investigated intensively in the library and Web literature, e.g. [2, 24]; these methods could be adapted.

(A5.2) Content recognition needs to improve on the mapping to the WordNet domain hierarchy. There are still un-

explored routes to improving the results. For example, our earlier results [1] suggest that going beyond the hierarchical classification system to also reflect that some domains are "related to" others can improve recognition and labelling quality. However, this also increased topic drift; for a labelling method as investigated in the current paper, it might lead to too many (and ultimately meaningless) labels being generated.

The question is what the alternatives are. Approaches that map into folksonomies on the basis of text similarity with already-labelled posts (e.g., `http://www.tagyu.com`) refer to a model class that may be more intuitive (and is certainly more dynamic) than a Dewey-style classification. However, the results of [16] indicate that tag prediction based on text similarity is but a first step in capturing taggers' behaviour (precision and recall at 10 were 0.4 and 0.49).

One direction for future research would be the integration of complementary content-extraction techniques (e.g., from tags or from body nouns) into such classifications, and a user-adaptive personalization based on reading preferences. The question then arises how to combine folksonomy facets on the one hand (such as the open, dynamic, and adaptive nature of the set of available domains / topics of discourse) and traditional taxonomy/ontology facets on the other (such as well-defined terms and inferencing capabilities).

(A5.3) The empirical method of the user study – regarding low inter-annotator reliability as a feature not a bug – is promising, especially in moves towards sophisticated user-adaptivity in open, flexible environments like social media.

Obviously, the present study can be but a start; larger samples both of annotators and of materials are needed. In addition, the conclusions that were derived from the exploratory analysis presented in this paper require confirmation by experimental methods. In particular, we plan to employ psycholinguistic methods in future work in order to validate our conjectures on reading strategies and content assessment.

Further insights could be gained by (quasi-)experimental studies of the impact of features like language, information on general vs. proper nouns[9], author/reader demographics, and tagging system features.

(A5.4) Deterministic text classifiers like the ones used in this paper have the advantage of allowing for a "cold start" (including a cold start of a research question), but they are also known to be rather inflexible and, in general, inferior in performance to classifiers built from machine learning. Machine learning has been employed in tagging systems, e.g., [23] or [15]. Given a larger sample of labelled blogs, ideally a standard data set like Reuters, machine learning approaches could be employed easily. However, for the reasons named above, it does not appear realistic to hope for such a set – for reasons inherent to the blogs domain. Therefore, in future work we aim at utilizing semi-supervised and active learning to both reduce the required size of the manually labelled text sets and to introduce more flexibility and user-adaptivity into the process.

[9] Proper-noun tags did occur frequently in the analyzed corpus, but they were not among the top tags. See `http://www.wiwi.hu-berlin.de/~berendt/Papers/ICWSM07/` .

References

[1] Berendt, B. & Navigli, R. (2006). Finding your way through blogspace: Using semantics for cross-domain blog analysis. In *CAAW: Proc. AAAI 2006 Symposium on Computational Approaches to Analysing Weblogs.* Stanford, CA: March 2006 (pp. 1-8). Technical Report SS-06-03. Menlo Park, CA: AAAI Press. `http://www.wiwi.hu-berlin.de/ berendt/Papers/SS0603BerendtB.pdf`

[2] Bradshaw, S. (2003). Reference directed indexing: Redeeming relevance for subject search in citation indexes. In *Proc. of the 7th ECDL.*

[3] Brooks, C.H., & Montanez, N. (2006). An analysis of the effectiveness of tagging in blogs. In *Proc. CAAW* (pp. 9–14).

[4] Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, 22,* 249–254.

[5] Fellbaum, C. (Ed.) (1998). *WordNet: an Electronic Lexical Database.* MIT Press.

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

[7] Hammond, T., Hanny, T., Lund, B., & Scott, J. (2005). Social bookmarking tools – A general overview. *D-Lib Magazine 11 (4).*

[8] Haveliwala, T., Gionis, A., Klein, D., & Indyk, P. (2002). Evaluating strategies for similarity search on the web. In *Proc. of WWW 2002* (pp. 432–442).

[9] Hotho, A., Staab, S., & Stumme, G. (2003). Wordnet improves text document clustering. In *Proc. of the Semantic Web Workshop at SIGIR-2003.*

[10] Ikeda, D., Fujuki, T., & Okumur, M. (2006). Automatically linking news articles to blog entries. In *Proc. CAAW.*

[11] Krippendorff, K. (1980). *Content Analysis: An Introduction to its Methodology.* Beverly Hills, CA: Sage Publications.

[12] Lewis, D.D., Yang, Y., Rose, T.G., & Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research, 5,* 361–397.

[13] Magnini, B. & Cavaglià, G. (2000). Integrating Subject Field Codes into WordNet. In *Proc. LREC-2000* (pp. 1413–1418).

[14] Marlow, C., Naaman, M., boyd, d., & Davis, M. (2006). HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, ToRead. In *Proc. of the Seventeenth Conference on Hypertext and Hypermedia. HYPERTEXT'06* (pp. 31–40). N.Y.: ACM Press.

[15] Mierswa, I. & Wurst, M. (2005). Efficient Case Based Feature Construction for Heterogeneous Learning Tasks. In *Proc. of ECML 2005* (pp. 641–648). Springer.

[16] Mishne, G. (2006). AutoTag: A collaborative approach to automated tag assignment for weblog posts. In *Proc. of WWW2006* (pp. 953–954).

[17] Oka, M., Abe, H., & Kato, K. (2006). Extracting Topics From Weblogs Through Frequency Segments. In *Proc. of WWW2006 3rd Annual Workshop on the Weblogging Ecosystem.* `http://www.blogpulse.com/www2006-workshop/papers/wwe2006-oka.pdf`

[18] Quintarelli, E. (2005). Folksonomies: power to the people. *ISKO Italy-UniMIB meeting.* `http://www.iskoi.org/doc/folksonomies.htm`

[19] Rosenberg, A., & Binkowski, E. (2004). Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points. In *Proc. of NAACL.*

[20] Rosh, E., Mervis, C.B., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8,* 383-439.

[21] Thelwall, M. (2006). Blogs During the London Attacks: Top Information Sources and Topics. In *Proc. of WWW2006 WS Weblogging Ecosystem.* `http://www.blogpulse.com/-www2006-workshop/papers/blogs-during-london-attacks.pdf`

[22] Toutanova, K., & Manning, C.D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proc. of the 2000 Joint SIGDAT Conf. on Empirical Methods in NLP* (pp. 63–70).

[23] Tremblay-Beaumont, H., & Aïmeur, E. (2005). Feature combination in a recommender system using distributed items: The case of JukeBlog. In *Proc. of Multi-Agent Inform. Retrieval and Recommender Systems WS* at IJCAI-05 (pp. 70–74).

[24] Utard, H., & Fürnkranz, J. (2006). Link-Local Features for Hypertext Classification. In Ackermann, M. et al. (Eds.), *Semantics, Web, and Mining* (pp. 51–64). Springer. LNAI 4289.

[25] Véronis, J. (2001). *Sense tagging: Does it make sense?* Paper presented at the Corpus Linguistics 2001 Conf., Lancaster, UK. `http://citeseer.ist.psu.edu/veronis01sense.html`