COMPARATIVE STUDIES OF SOCIAL CLASSIFICATION SYSTEMS USING RSS FEEDS

Steffen Oldenburg
Chair for Information and Communication Services, Department of Computer Science, University of Rostock
Albert-Einstein-Strasse 21, D-18059 Rostock, Germany

Keywords: Social Classification, Efficient Personal and Collaborative Tagging, Processing and Analysis of Heterogeneous Tag Spaces, Tagging Automation, Tag Space Integration.

Abstract: This paper presents the results of practical studies comparing five well established social classification services for tagging of bookmarks (del.icio.us, BibSonomy bookmarks) and publications (BibSonomy publications, CiteULike, Connotea) in the context of service interoperability and integration. Contrary to most of current research we exclusively focus on the usage of RSS feeds for retrieval of tag-related data. Here we exploit ”recent” feeds, as this method of data retrieval corresponds directly to the way users can retrieve data from these services, e.g. for tag suggestions. We motivate the preferred usage of feeds compared to full site grabbing, and present analysis results of feed data from a period of one month concerning feature distribution, growth, stability and convergence aspects. Furthermore we compare tag spaces and their intersections for potential interoperability and integration of these services, and reveal that tags in practice are not really as freely chosen as often promised.

1 INTRODUCTION

The emerging trend to public sharing of information and knowledge implies a growing demand in lightweight classification with low participation barriers for users. This trend to collaboratively attach any theoretically unrestricted, free-form key words to content - called tagging - has produced a tremendously rising number of non-integrated tag spaces, tagged heterogeneous resources, and isolated tagging services.

However, recent quantitative research raises concern that this growth trend complicates for individual users to efficiently benefit from (resource discovery) and contribute to (resource annotation) social classification over time. This represents a tough challenge in the context of exploring best fitting vocabulary for individual or public resources.

Getting operational with bootstrapping tag spaces by retrieving best fitting tag vocabulary, staying operational over time by evolving best fitting vocabulary as well as staying independent and interoperable by importing and exporting service-specific tag vocabulary and tagged content are essential requirements for long-term user acceptance as well as efficient participation in different-scale social classification systems.

However, focusing on single folksonomies and isolated tagged resources as not being inter-related so far, only little research has been done on these topics.

1.1 Overview and Context

This paper presents work in progress of a PhD thesis in the context of interoperable and integrated social classification systems. As key requirement we determine a transparently extensible, integrative and interoperable tagging service - supporting efficient bootstrapping of tag-related information from heterogeneous social classification systems with different thematic focuses, local or restricted user resources, as well as a dynamic evolution of user-centric tag spaces in an integrated context. Here we present results of comparative studies with well-established classification services over a time span of one month from August 01, 2007 - September 01, 2007, exclusively based on RSS feeds (Really Simple Syndication). Additionally, we compare our results with publicly available full dumps regarding data until Sept. 01, 2007.

Our mid-term target is to analyse requirements to establish a uniform, extensible architecture for a social classification analysis framework and interactive evaluation platform for efficient and integrated personal tagging, deriving relevant tagging features, e.g.

Oldenburg S. (2008).
COMPARATIVE STUDIES OF SOCIAL CLASSIFICATION SYSTEMS USING RSS FEEDS.
In *Proceedings of the Fourth International Conference on Web Information Systems and Technologies*, pages 394-403
DOI: 10.5220/0001523803940403

tag suggestions, from a dynamically evolving individual tag space (assisted tagging).

Our long-term focus is placed on efficient (semi)automated tagging and tag suggestions based on that integrative, interoperable approach, and on analysis and application of the resulting tag spaces for optimized navigability abstracting from the specific tagging services in background. Users should be enabled to work with one consistent, virtual tag space, and not depend on service-specific restrictions.

2 RELATED WORKS

This paper covers specific topics related to comparability, integration and interoperability of social classification and tagging, and analyses leading tagging services with different scales of popularity, growth as well as thematic focuses. For a general overview and research motivation refer to community discussions in (Mathes, 2004), (Shirky, 2005). For recent research of tagging motivations read (Ames and Naaman, 2007) or (Zollers, 2007). Associated quantitative evaluations of static and dynamic features as well as emerging structures in tag spaces are presented in (Cattuto et al., 2006), (Cattuto, 2007), (Golder and Huberman, 2005), or (Lambiotte and Ausloos, 2006). (Zhang et al., 2006) compare the motivations, advantages and drawbacks of traditional top-down and emerging bottom-up semantics concerning Web resources and present results from del.icio.us analysis. A BibSonomy overview is given in (Hotho et al., 2006).

Comparison, Integration and Interoperability Studies. (Gruber, 2005) proposes an approach for defining an ontology that would enable the exchange of tag data and the construction of tagging systems that can compositionally interact with other systems. (Veres, 2006) evaluates semantic intersections and interoperable features between different tagging services (flickr, del.icio.us), but lacks profound quantitative evaluation. The relation between texts from blog posts and tags associated with them are analysed in (Berendt and Hanser, 2007). Inter-relations between different tag spaces are not considered. (Bhagat et al., 2007) analyse how different information networks (e.g. web, chat, email, blog, instant messenger) interact with each other, e.g. correlations between blog - blog, blog - web or blog - messenger. (Schmitz et al., 2007) analyse and compare co-occurrence network properties of del.icio.us data (actual as of 2004-2005) and BibSonomy data (as of July 2006).

Distribution, Growth, and Stability. Feed based analysis using del.icio.us data is exploited in (Shaw, 2005), (Begelman et al., 2006), or on deli.ckoma[1] web site. The last one presents actual statistics derived from recent RSS feeds, and evaluates data retrieval coverage and error probability. (Halpin et al., 2007) analyse whether coherent and stable categorization schemes can emerge from unsupervised tagging, and they evaluate its dynamics over time, including corresponding power-laws in del.icio.us tag distributions for resources with different popularity scale. A brief CiteULike analysis including power-law is given in (Capocci and Caldarelli, 2007).

Tag Space Navigability and Efficiency. (Chi and Mytkowicz, 2007) analyse early data (actual as of 2004-2005) from large-scale del.icio.us with (conditional) entropy concerning efficient navigability, and reveal that efficiency is decreasing over time. Efficiency analysis using entropy measure is also used in (Zhang et al., 2006) and (Li et al., 2007). (Santos-Neto et al., 2007) analyse CiteULike and BibSonomy whether usage patterns can be exploited to improve the navigability in a growing tagsonomy. They analyse the smaller scale services BibSonomy and CiteULike to reveal tagging activity distribution, and define metrics to uncover similarities in user interests. (Brooks and Montanez, 2006) analyse the effectiveness of tags to describe blog contents (technorati[2], REST API). The authors suggest that tags are more useful to assign blogs to broad category clusters than to indicate particular resource content. Hence, they exploit text contents to automatically extract relevant keywords (TF-IDF) for tag usage and compare different combinations of these approaches.

Review of the State of the Art. Existing research approaches introduce metrics and measures for tag related similarities, growth, stability, and efficiency. They apply them on basically comparable data sets - mostly the popular broad folksonomy del.icio.us, in some cases the less frequently used services CiteULike or BibSonomy. However, results from these different research publications cannot be effectively compared due to different time scopes, evaluation targets, amounts of data, data retrieval concepts, and a missing comprehensive analysis architecture following an integrative approach. Thus, chances to evaluate, compare and rank tag or resource spaces, e.g. for efficient tag suggestions, and to deduce conclusions to optimize tagging processes are hard to identify. There is need for an evaluation approach on comparable actual data sets from the same time span, based on uniform data retrieval which is in the scope of this paper.

[1] `http://deli.ckoma.net/stats`
[2] `http://www.technorati.com/`

3 RSS FEEDS AS SOURCE FOR TAG SPACE BOOTSTRAPPING

RSS feeds are offered by many leading social classification services, at least for recent data, in general also for specific tags, users, and resources. This promises a more consistent retrieval of heterogeneous tag data than site-depending methods, e.g. full or random site grabs using Web spiders like `wget`.

```
<item rdf:about="http://code.google.com/">
<title>Google Code – Developer Network</title>
<link>http://code.google.com/</link>
<description></description>
<dc:creator>lhc1111</dc:creator>
<dc:date>2007-08-31T22:02:16Z</dc:date>
<dc:subject>API Code Google ajax</dc:subject>
<taxo:topics><rdf:Bag>
  <rdf:li resource="http://del.icio.us/tag/Google"/>
  <rdf:li resource="http://del.icio.us/tag/Code"/>
  <rdf:li resource="http://del.icio.us/tag/API"/>
  <rdf:li resource="http://del.icio.us/tag/ajax"/>
</rdf:Bag></taxo:topics> </item>
```

Listing 1: Example of RSS feed item.

Past research either fully relied on site grabs or at least initial grabs with further incremental updates using feeds. Grabs are subject to changes in HTML structure, its dynamical generation, as well as site growth, hence need to regard current site properties. Full grabs are not well accepted by many services (site ban warnings, read FAQs), and full dumps are rarely available, e.g. here from CiteULike (direct download), and BibSonomy (acceptance of conditions, download link per mail).

Contrarily, feeds have very similar content structure and XML markup. Service dependency is much lower, though there are minor differences in XML tags or in availability of specific properties for feed items. An example item is given in listing 1.

With feeds we generate less load on service sites, and are unlikely to become subject to site bans. Feed-based growth in tags, users, resources promises statistically relevant data amounts in relatively short time as our analysis will reveal. We can operate without storing site history as we are primarily interested in supporting users with actual tag data, as sites dynamically evolve including interest shifts. Furthermore site growth since service launch has produced such a tremendous amount of data, which cannot be efficiently handled anymore for popular sites, e.g. refer to del.icio.us properties in Table 2.

As we want to support users in uniform tagging with heterogeneous tagging services we need to exploit the same data users have access to. Users do not want to download full dumps or grab histories. Using RSS we benefit from a widely uniform format based on RDF / XML, being UTF8-encoded in most cases, and thus can seamlessly integrate new services complying to this format and services using it, e.g. feeds based on RSS or ATOM standards.

After profound reading of publications emerged in the context of quantitative analysis of tagging services during the last two years we have to pose a basic question: Do we need complete history-scale dumps of tag spaces, or is it sufficient, and more efficient to just evaluate current and future data with less scale, but similar properties concerning distribution, convergence and stability of tag spaces - over some time - to get and stay operational? Interestingly, feeds offer richer semantics in tag data than service backends (read Section 4.6).

4 ANALYSIS AND EVALUATION

In the following section we provide insight into our test environment, and relevant evaluations. Finally we assess our method of data retrieval.

4.1 Test Environment

For our analysis we selected the highly popular site del.icio.us (10 sec interval, fast item updates), the popular sites CiteULike (10 min) and Connotea (3 min, less items per feed), and the less popular site BibSonomy, distinguishing between feeds for bookmarks (Bib1, 10 min) and publications (Bib2, 10 min). Refer to Tables 1 and 2, from now on we will address the services with the given IDs.

Table 1: Service URLs for recent RSS feeds.

Service	URL (http://)	ID
BibSonomy	www.bibsonomy.org/rss	Bib1
	www.bibsonomy.org/publrss	Bib2
CiteULike	www.citeulike.org/rss	Cit
Connotea	www.connotea.org/rss	Con
del.icio.us	del.icio.us/rss	Del

The database schema of our current testbed is illustrated in Figure 1. We use separate schemas for each folksonomy in test. Tags $t \in T = \{t_1, t_2, ..., t_k\}$, users $u \in U = \{u_1, u_2, ..., u_l\}$, and resources $r \in R = \{r_1, r_2, ..., r_m\}$ are stored with time stamps, and tag assignment counters in separate tables and associated to each other in the tag assignments table (TAS) as quadruples $tas = (t, u, r, ts) \in$

$TAS \subseteq T \times U \times R \times TS$ with a time stamp $ts \in TS = TS_{ISO8601} = \{ts_0 \leq ts \leq ts_n\}$ with ts_0 = `01.08.07T16:00:00`, and ts_n =`01.09.07T23:59:59`. Triples (t, u, r) are unique. Items (posts) $i \in I \subseteq U \times R \times TS \times T^*$ extracted from RSS feeds - we only process non-empty items (T^+) - are not directly reflected in the database. They can be retrieved using SQL grouping or MD5 hashes on tag assignment attributes (u, r, ts).

Table 2: Sizes of tags (T), users (U), resources (R), tag assignments (A, in text TAS), co-tags / edges (E), co-tag assignments (C, in text CAS, factor 10^6), and items (I).

	Bib1	Bib2	Cit	Con	Del
T	8716	1664	14282	12215	238047
U	1433	135	1683	2352	213190
R	4726	1529	17912	17032	823411
A	24424	5554	67395	71325	5485163
E	68163	7953	160474	131400	2661505
C	0,114	0,012	3,918	0,299	10,786
I	5285	1570	18221	17440	1822456

The co-tags table stores edges $e = (t_i, t_j) \in E \subseteq T \times T, t_i \leq_{alpha-numeric} t_j, t_i \neq t_j$ of the tag co-occurrence network with usage counters as weights. For each RSS item we sort the local tag list and combine each tag with all its successors (filtering self-co-occurrences), resulting in a local fully connected undirected graph with $n_i * (n_i - 1)/2$ tags (clique) for item i with n tags. Each co-tag assignment $cas \in CAS \subseteq E \times TS$ is stored in the co-tag assignments (CAS) table.

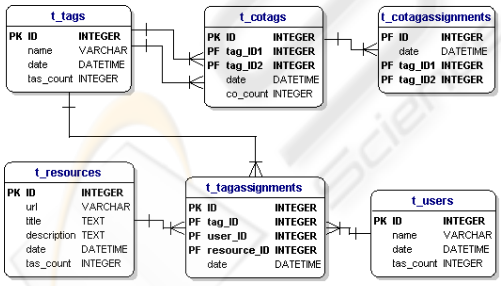

Figure 1: Database schema for test bed.

We requested recent RSS feeds using service specific manually adjusted request intervals. For an overview of data retrieval refer to Figure 2.

Depending on the interval chosen we receive 144 (10 min interval) up to 8640 (10 sec) XML files (feeds) per day, being archived on a daily basis. Item features, e.g. resources and tags, are extracted from archived feeds into CSV tuples using regular expressions. Finally they are propagated to the appropriate database schema. To preserve a maximum of comparability the extraction restricts to use non-empty items (at least one tag), containing only unreserved characters according to RFC 3986, among these at least one character from [`a-zA-Z0-9`]. Space separated word groups are split, tags are unescaped (HTML), we decode UTF-8 %-encoding, and remove [`,;"\`]. For details refer to Tables 2, 5 and Figure 1. The integrity conditions $\sum_{t \in T} tas\_count = \sum_{u \in U} tas\_count = \sum_{r \in R} tas\_count = |TAS|$, and $\sum_{e \in E} co\_count = |CAS|$ are satisfied.

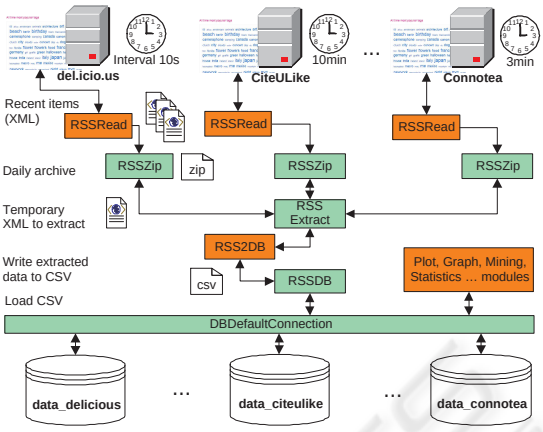

Figure 2: Analysis architecture for test bed.

Tests were conducted on a machine with 1.5 GB RAM, 2 GHz T7200 Dual Core CPU, running Windows XP SP2, MySQL v5.0.27 with large configuration and MYISAM engine, and Sun Java SE v1.6.

4.2 Power-law Analysis

Does RSS feed extracted data reveal typical distribution features? In order to determine whether our data is representative for a folksonomy we need to show that typical distributions comply to a power-law.

Here we present the distributions for tags (see Figure 3), and resources (see Figure 4) per tag assignment. The plots reveal typical power-law behaviour (nearly linear in log-log scale) with small head and big tail at different scales. Del.icio.us represents the most popular folksonomy in test, BibSonomy publications the least frequented one, followed by BibSonomy bookmarks. Connotea and CiteULike reveal very similar properties which is not only visible in these plots. The plots indicate that RSS feeds are an absolutely satisfying data source, as feed data very rapidly establish typical power-law distributions. Subsequently, feeds are satisfactory resources for tag analysis and tag suggestions. We do not have to favour tagging history, but can focus on recent tag re-

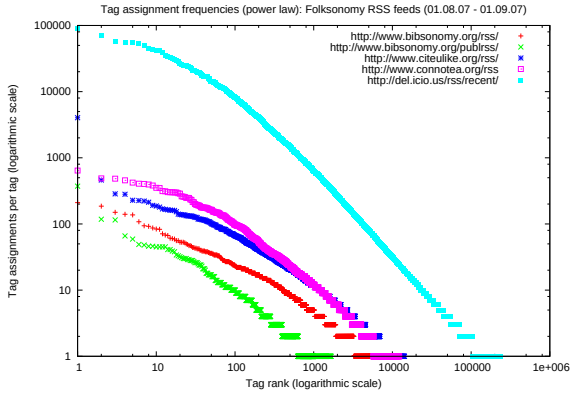

Figure 3: Tag assignments per tag (power-law).

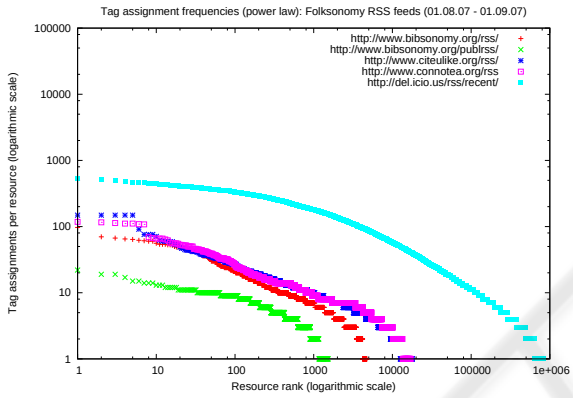

Figure 4: Tag assignments per resource (power-law).

lated information. Less data produces less load on the tagging service, and can be analysed more efficiently. A major drawback is, that we do not retrieve the same number of tail level tags as with site grabs, the part of the distribution bearing most spam, but also less frequently used relevant tags.

4.3 Growth and Convergence Analysis

A further question is whether tag related distributions retrieved from RSS feed data converge quickly enough to get stable after short time, and how long feeds need to be requested to achieve that stability. We provide a general overview about per-day growth and cumulative growth - here tags only - in Figures 5 and 6. Both plots are in log-normal scale to reveal scale (popularity) differences between the services. Per-day tag (resource, user) growth reveals falling trends for del.icio.us, CiteULike, and Connotea, indicating that the longer the studies go the more of the most frequently used tags (actual resources, active users) have been retrieved. Normal-normal scale is nearly linear for cumulative growth. The corresponding distribution of tags concerning number of tags per item is presented in Figure 7. This distribution will become relevant for upcoming tag suggestion research, e.g. for local co-occurrence analysis of 2-, 3-, or n-tag networks.

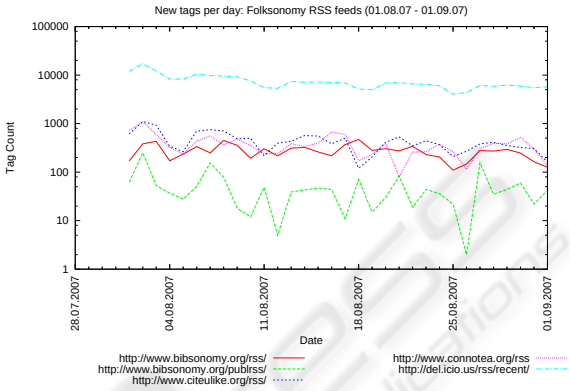

Figure 5: Per day tag growth over time, log-normal scale.

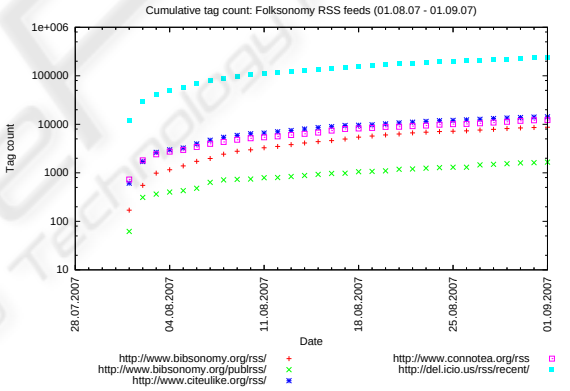

Figure 6: Cumulative tag growth (tags per day) over time, log-normal scale for better visibility due to outscaling of del.icio.us.

How can we assess pairwise similarity between subsequent feeds, e.g. for tag sets? Therefore we retrieved top 1000 tags, users and resources, and computed pairwise similarity between subsequent top sets. For space reasons we only present top 1000 tags for popular del.icio.us site (Figure 8) and top 1000 resources for Connotea (Figure 9), the last one being representative for convergence of all top distributions other than del.icio.us. Del.icio.us distributions stabilize very quickly (90% similarity threshold after two days), the other services need about four days to reach 80% similarity. We use the Jaccard measure for basic set similarity (no regard of rank): $j = sim_{Jaccard}(X,Y) = |X \cap Y|/|X \cup Y|$.

In order to regard an element's rank we introduce a shift distance measure on sets to assess the num-

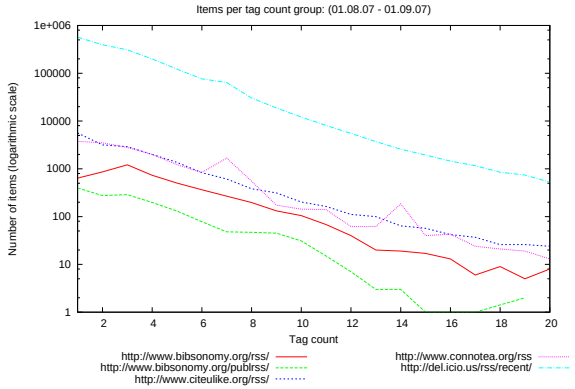

Figure 7: Tagged items grouped according to number of used tags, log-normal scale.

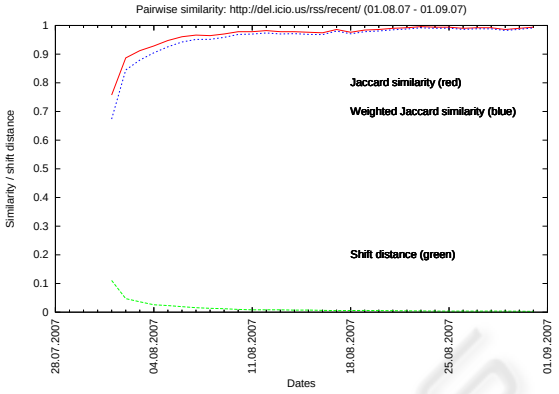

Figure 8: Pairwise similarity of top 1000 tags, del.icio.us.

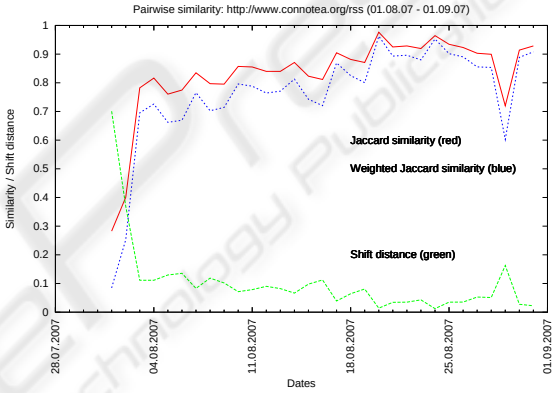

Figure 9: Pairwise similarity of top 1000 resources, Connotea.

ber of position shifts. For two sets X and Y with $n = size(X) = size(Y)$ in order to transform X into Y for all elements e we calculate shift costs $c_{shift}(X,Y) = \sum_{e \in X \wedge e \in Y} |r_X(e) - r_Y(e)|$, insert costs (shift into set) $c_{ins}(X,Y) = \sum_{e \notin X \wedge e \in Y} |n - r_Y(e) + 1|$, delete costs (shift out of set) $c_{del}(X,Y) = \sum_{e \in X \wedge e \notin Y} |n - r_X(e) + 1|$ with rank $r(e) : 1 \leq r(e) \leq n$, and summarize $c_{abs-shifts}(X,Y) = c_{del}(X,Y) + c_{ins}(X,Y) + c_{shift}(X,Y)$. The shift weight s then reads $s = c_{abs-shifts}/c_{max-shifts}$ with $c_{max-shifts} = n * (n+1)$, applied on j: $sim_{Jaccard,weighted} = j * (1-s)$. $c_{max-shifts}$ occurs for two disjoint sets with $n * (n+1)/2$ delete as well as insert shifts, e.g. e_1 shifts down (up) by n positions, e_n by 1 to leave (claim) positions in X (Y).

Assumption 1 is that the sets are equal-sized, otherwise we choose the lower size as reference. Assumption 2 is that only insertion is allowed, assuming only equal-sized or growing sets, which of course is true for our feed-based folksonomy data in test. Hence, between subsequent sets X_i, X_{i+1} holds: $\forall i : |X_i| \leq |X_{i+1}|$. Our shift distance does not perform any reordering, it only looks ahead to assume a measure on it, not taking into account any improved item order after some reordering step. We penalize the initial state of disorder. Other distance metrics may be applied as well, e.g. Levenshtein or Ulam (longest common subsequence) distances.

4.4 Navigability Analysis

Motivated by (Chi and Mytkowicz, 2007), we apply the entropy measure to assess tagging navigability: $H(T) = -\sum_{t \in T} p(t) * log_2(p(t))$ with tag probability $p(t) = |TAS(t)|/|TAS|$, TAS denotes tag assignments, $TAS(t)$ the tag assignments with tag t.

With increasing size of T the entropy will grow as well as with the distribution of $t \in T$ over TAS becoming more uniform. The higher the entropy value grows, the less information is contained, and the less efficient any tag space navigation will become.

Our analysis reveals a similar growth in entropy for all services (at different scales), getting flatter over time (plateau). It is obvious that navigability in del.icio.us is flattening fastest (plateau nearly parallel to x-axis) in comparison with the other services because this service has the highest tag (assignment) growth rate (compare Figure 6). The nearly constant entropy is due to the tag distribution getting less uniform over time. This implies an increasing difficulty for users or tag suggestion algorithms to find unique tags or at least tags with less recall and higher precision. Our analysis result is given in Figure 10. As entropy over time nearly looks the same for all services, this raises the question whether it is the optimal measure to assess efficient navigability.

Entropy in current research is used to evaluate the navigability concerning distribution of single tags only. However, in real life users apply more than one tag to discover a resource. Either the service in use

supports the usage of multiple tags at the same time, or it may offer the opportunity to search in previous search results, e.g. try Connotea in-collection search feature. Thus, it would be more interesting to analyse the entropy of tags in context, e.g. co-tags of 2 and more correlated tags because using more than one tag efficiently reduces search space. Thus, the entropy of tag combinations should result in values indicating better efficiency and specifity. This idea is also motivated by rapidly increasing conditional entropies of documents on tags presented in (Chi and Mytkowicz, 2007) indicating decreasing (single!) tag specifity. This idea will be investigated in future research.

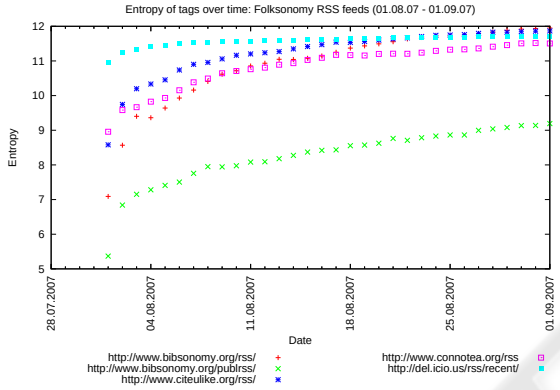

Figure 10: Entropy of single tags over time.

4.5 Intersections between Tag Spaces

Before covering interoperability or integration issues of tag spaces from different folksonomies we have to analyse whether there is a fundamental need. This need arises if tag spaces reveal significantly different thematic focuses with relevant portions of unique tags not being contained in pairwise intersections.

Table 3: Ratio of pairwise intersections between folksonomy tag spaces (row T_i, column T_j): $Ratio[T_i, T_j] = |T_i \cap T_j|/|T_i|$, with $T_i, T_j \in \{Bib1, Bib2, Cit, Con, Del\}$.

	Bib1	Bib2	Cit	Con	Del
Bib1	1,00	0,06	0,19	0,33	0,68
Bib2	0,34	1,00	0,46	0,38	0,68
Cit	0,11	0,05	1,00	0,22	0,44
Con	0,24	0,05	0,26	1,00	0,63
Del	0,02	0,01	0,03	0,03	1,00

Therefore we first explore common parts of tag spaces using pairwise intersections based on tag string equality (see Table 3). It is obvious and it was expected that del.icio.us has the highest coverage of common language words or typical top tags (column Del) used in other tag spaces. However, there are differences in usage of tag assignments for common tags as well as a significant portion of tags not being contained in the intersection (difference between 100% and value given in table). For example the intersection of del.icio.us and Connotea makes up to 67% of Connotea tag space (row Con, column Del) and 3% of del.icio.us tag space (row Del, column Con). This initially motivates a preferred usage of del.icio.us for tagging or tag suggestion retrieval. However, there is a 33% portion of tags in Connotea not being contained in del.icio.us, motivating a preferred usage of Connotea for such topics exclusively covered by these tags, e.g. Connotea and CiteULike have a thematic focus on natural sciences.

As Table 4 reveals, the differing portions of tag spaces are growing over time (decreasing ratios), e.g. see values for full dump intersections. We calculated pairwise intersections, revealing the ratios between tag spaces from full dumps (full long-time data sets) and / or feeds (short time). We use dumps from CiteULike and BibSonomy (both actual as of December 31, 2007), and regard all tag data until the end of our analysis scope. The dump-only intersections more clearly reveal different thematic focuses than feed-only intersections, hence strongly motivate comparative analysis.

Table 4: Ratio of pairwise intersections between folksonomy tag spaces (feeds, full dumps, row T_i, column T_j): $Ratio[T_i, T_j] = |T_i \cap T_j|/|T_i|$, with $T_i, T_j \in \{B1, B2, B1D, B2D, Ci, CiD\}$ with B1=BibSonomy, B2=BibSonomy-Pub., Ci=CiteULike, D denotes a dump. B1D has 33719, B2D 13893, and CiD 197463 tags.

	B1	B2	B1D	B2D	Ci	CiD
B1	1,00	0,06	0,51	0,23	0,19	0,75
B2	0,34	1,00	0,61	0,95	0,46	0,74
B1D	0,13	0,03	1,00	0,16	0,12	0,44
B2D	0,14	0,11	0,38	1,00	0,27	0,66
Ci	0,11	0,05	0,29	0,26	1,00	0,98
CiD	0,03	0,01	0,08	0,05	0,07	1,00

Further research will cover intersection rates between tags according to rank, e.g. top level tags, as well as a comparison of the two co-occurrence networks resulting from each pairwise intersection. We assume to unveil significantly different thematic focuses and co-tag distributions. Other intersection options assume a prior mapping of semantically similar tags, or to filter out spam or irrelevant tags, e.g. applying a tag usage threshold of 2 reduces the tag space significantly by about 50% (long power-law tail, see Figure 3).

4.6 Fairy Tale of Freely Chosen Tags

Why is it useful to analyse and compare tag creation and storage? During our analysis we stumbled upon many slang and spam tags as well as tags with high portion of non-numbers and non-characters. We selected some of these tags and checked, whether these tags retrieve a search result at all, and whether these search results are specific or coincide with those retrieved using normalized tags. All services in test provide a web interface to search for a specific tag as well as a feed interface to request the corresponding recently tagged items.

The idea is also motivated having a look into CiteULike and BibSonomy dumps, revealing that the effective, normalized tags (stored in the service backend) are not equal to those applied in feeds (user intended tags), neither in semantic richness nor (probably) in number. Applying non-normalized tags from feeds either gains no or different search results (exact queries), or the same search result using normalized tags (similarity query, like query). For instance del.icio.us allows usage of unreserved characters (RFC 3986) for tag creation, e.g. `@`, `!`, `#`, `+`. They are used for tags in feeds, but queries skip them using like-queries. They are not used to enforce specific semantics, e.g. query(`c++`) = query(`c`) = query(`c#`) or query(`.net`) = query(`net`). The result feeds contain combinations of tags `c`, `c++`, `c#` or respectively `.net` and `net`, not only the tag being searched for.

Finally this observation contradicts the widely used tagging promise that any freely chosen keyword can be used as a tag. We notice a loss of user specific semantics from feed to backend as well as a much smaller character space to assemble tag words from. Another aspect is that tags are mostly provided in context with other tags (co-occurrences). Even full chapter titles, word groups or sentences are used as tags according to CiteULike and Connotea feeds. In the backend the context between tags is lost due to splitting of word groups, normalizing words in, or eliminating words from them.

This information is not provided by services, e.g. del.icio.us FAQ says that users are allowed to use character `*` in tags to express emotions or ranking, however these characters have no effect. Either they get removed from a tag or the tag is not stored in backend. This cannot be reliably determined using feeds without a dump to compare to. BibSonomy FAQ states that feeds are periodically propagated into backend database, hence does not exclude that effective tags might differ from those applied by users. Users have to know about restrictions in order to adapt their tag spelling and semantic mapping accordingly.

Table 5: Effective tag spaces and queries: F denotes tag feeds, D: tag dumps, e: exact query (q), l: like q., w: wildcard q., b: boolean q., c: in-collection q., r: ranked order. Unreserved (`a-zA-Z0-9_-.~`) / reserved characters, and URL per-cent encoding refer to RFC 3986.

	Bib1	Bib2	Cit	Con	Del
Method	F/D	F/D	F/D	F	F
Query	e/r	e/r	b/e/w	c/e	e/l
Case-sens.	no	no	no	yes	no
Unreserved	yes	yes	yes	yes	yes
Reserved	yes	yes	no	no	yes
äöü	yes	yes	no	yes	yes
%-Enc.	no	no	no	no	yes

Hence, there is a motivation for deeper analysis of intended and effective semantics to evaluate the extent to which different tagsonomies can be compared to and integrated into each other or a separate unified tagsonomy (user based, group based) to support efficient context-based tag suggestions / (semi) automatic classification. For bootstrapping and dynamic evolution (e.g. merging, import, export) of tag spaces it is necessary to know differences and commonalities as well as mutual interpretation of tags and tag properties. Here we provide a brief overview of our observations in Table 5.

4.7 Evaluation of Data Retrieval

For our analysis we requested RSS recent feeds from social classification services, presenting here the results for the continuous feed stream from August 01, 2007 until September 01, 2007 (first day partially). Service-specific item request intervals were manually adjusted and stayed constant from 2nd day on (see Table 1). Table 7 displays the per feed statistics.

A confidential request interval directly depends on the number of items per feed. The lower the number is the more frequently the feed must be requested. A higher number allows for more relaxed intervals. In order to evaluate the confidence in coverage of retrieved data compared to available data, we present an analysis of feed overlapping between subsequent items, for an example see Figure 11, for statistics Table 6. All key figures have been computed using Apache Jakarta Commons Math Statistics[3]. An overlap in our context is defined as follows: be $F = \{f_{ts_0}, \cdots, f_{ts_n}\}$ the stream of feeds with time stamps $ts_0 \leq ts_k \leq ts_n$. An overlap is the item sequence in the intersection $O_{k,k+1} = I(f_{ts_k}) \cap I(f_{ts_{k+1}})$. For all items i we count the occurrences $occ(i)$ denoting

[3] `http://commons.apache.org/math/`

Table 6: Overlap statistics with Item Efficiency $= |Items_{Stored}|/|Items_{Extracted}|$, see table sizes in Table 2. del.icio.us $max = 25$ was a single peak due to a short-time forgotten interval, connotea single peak $max = 226$ due to request issues.

Service	Extr. Items	Avg	Min	Max	Deviation	Kurtosis	Skewness	Efficiency
Bib1	83219	11,66	1	87	11,09	-0,43	-0,38	6,35%
Bib2	61768	34,27	1	297	50,08	-0,64	0,95	2,54%
Cit	228939	12,52	1	62	9,10	-0,99	-0,16	7,96%
Con	139058	7,73	1	226	8,73	-0,76	0,37	12,54%
Del	3838134	1,90	1	25	0,84	-0,22	-0,08	47,48%

Table 7: Items per feed statistics, empty items are filtered.

Service	Avg	Min	Max	Dev
Bib1	18,75	1	19	1,47
Bib2	14,00	14	14	0,00
Cit	51,00	51	51	0,00
Con	9,92	1	10	0,41
Del	15,03	2	28	2,86

the number of feeds including that item, a value of $occ(i) \geq 2$ indicates an overlap. Initial value is 1, otherwise $occ(i)_{ts_k} = occ(i)_{ts_{k-1}} + 1$, if $i \in O_{k-1,k}$. The table shows very small overlap for del.icio.us, 80,3% of items have overlap ≤ 2 (99,9% ≤ 5). The extreme is BibSonomy-Pub. (max 297) with a roughly sporadic item stream with just 62,3% ≤ 20. In between are BibSonomy (82,3% ≤ 20), CiteULike (83,8% ≤ 20), and Connotea (95.8% ≤ 20).

These values could be used to estimate a dynamic back-off for the request interval in order to reduce service load and feed data. Alternatively we could measure item distribution over time based on time stamp differences between subsequent items, but CiteULike RSS feeds do not contain item time stamps.

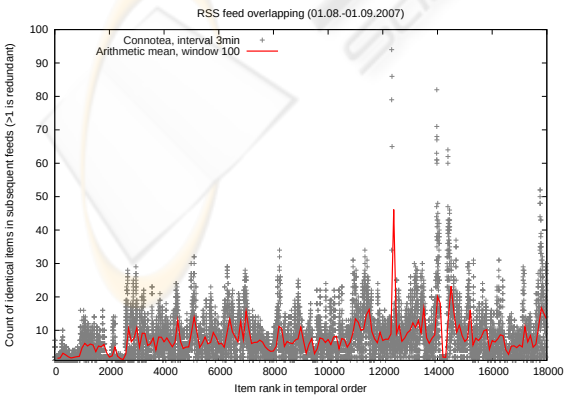

Figure 11: Feed overlapping between subsequent items for Connotea.

5 CONCLUSIONS AND OUTLOOK

We presented comparison results from feed-only analysis of 5 leading social classification services, to the best of our knowledge first work focusing on tag space comparison, integration and interoperability. Our analysis reveals that feed-only data satisfy typical distributions well, stabilize very rapidly concerning top-ranked data, and bear significant focus-dependent pair-wise intersections and thematic differences. Thus, they serve as a promising space saving source for comparative and integrative social network investigations.

Further mid-term research will cover a deeper comparison of feeds and dumps concerning semantic differences and coverage. As indicated in section 4.6 there is need for an analysis of loss in semantics between feeds and backend data. Promising results we also expect from a comparison of co-occurrence networks in order to offer context-specific tag suggestions as well as to unveil network differences for tag space intersections (same tags, different co-tag networks). Currently there is an ongoing master thesis evaluating tag suggestion algorithms concerning efficiency, quality and complexity of resulting tag spaces. A further study scheduled is about the correlations between tag spaces and tagged content using Vector Space Model (VSM, TF-IDF) in order to bootstrap tags for untagged content.

REFERENCES

Ames, M. and Naaman, M. (2007). Why We Tag: Motivations for Annotation in Mobile and Online Media. In *Proceedings of Computer/ Human Interaction Conference (CHI'07)*.

Begelman, G., Keller, P., and Smadja, F. (2006). Automated Tag Clustering: Improving search and exploration in the tag space. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*.

Berendt, B. and Hanser, C. (2007). Tags are not metadata, but ”just more content” - to some people. In *International Conference on Weblogs and Social Media (ICWSM'07)*.

Bhagat, S., Rozenbaum, I., Cormode, G., Muthukrishnan, S., and Xue, H. (2007). No Blog is an Island - Analyzing Connections Across Information Networks. In *ICWSM'2007 Boulder, Colorado, USA*.

Brooks, C. H. and Montanez, N. (2006). An Analysis of the Effectiveness of Tagging in Blogs. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*.

Capocci, A. and Caldarelli, G. (2007). Folksonomies and clustering in the collaborative system CiteULike. http://arxiv.org/abs/0710.2835.

Cattuto, C. (2007). Structure and Evolution of Collaborative Tagging Systems. In *WM 2007: Proceedings of the 4. Konferenz für Professionelles Wissensmanagement*. Workshop on Collaborative Knowledge Management.

Cattuto, C., Loreto, V., and Pietronero, L. (2006). Collaborative Tagging and Semiotic Dynamics. http://arxiv.org/abs/cs.CY/0605015.

Chi, E. H. and Mytkowicz, T. (2007). Understanding Navigability of Social Tagging Systems. In *Proceedings of Computer Human Interaction (CHI'07)*.

Golder, S. and Huberman, B. A. (2005). The Structure of Collaborative Tagging Systems. *Journal of Information Science*, 32:198–208.

Gruber, T. (2005). TagOntology - A way to agree on the semantics of tagging data. Presentation to Tag Camp, Palo Alto, CA,http://tomgruber.org/writing/tagontology.htm.

Halpin, H., Robu, V., and Shepherd, H. (2007). The Complex Dynamics of Collaborative Tagging. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*. Track E* Applications.

Hotho, A., Jäschke, R., Schmitz, C., and Stumme, G. (2006). BibSonomy: A Social Bookmark and Publication Sharing System. In de Moor, A., Polovina, S., and Delugach, H., editors, *Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures*, Aalborg, Denmark. Aalborg University Press.

Lambiotte, R. and Ausloos, M. (2006). Collaborative tagging as a tripartite network. *Springer LNCS*, 3993:1114–1117.

Li, R., Bao, S., Fei, B., Su, Z., and Yu, Y. (2007). Towards Effective Browsing of Large Scale Social Annotations. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*. Track: Web Engineering, Session: End-User Perspectives and measurement in Web Engineering.

Mathes, A. (2004). Folksonomies - Cooperative Classification and Communication Through Shared Metadata. http://www.adammathes.com/academic/computer-mediatedcommunication/folksonomies.html

Santos-Neto, E., Ripeanu, M., and Iamnitchi, A. (2007). Tracking User Attention in Collaborative Tagging Communities. In *Proceedings of International ACM/IEEE Workshop on Contextualized Attention Metadata: personalized access to digital resources*.

Schmitz, C., Grahl, M., Hotho, A., Stumme, G., Cattuto, C., Baldassarri, A., Loreto, V., and Servedio, V. D. P. (2007). Network Properties of Folksonomies. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*. Workshop on Tagging and Metadata for Social Information Organization.

Shaw, B. (2005). Utilizing Folksonomy: Similarity Metadata from the Del.icio.us System. http://www.metablake.com/webfolk/webproject.pdf.

Shirky, C. (2005). Ontology is Overrated: Categories, Links, and Tags. http://www.shirky.com/writings/ontology_overrated.html

Veres, C. (2006). Concept Modeling by the Masses: Folksonomy Structure and Interoperability. *Conceptual Modeling - ER 2006*, pages 325–338.

Zhang, L., Wu, X., and Yu, Y. (2006). Emergent Semantics from Folksonomies: A Quantitative Study. *Journal on Data Semantics VI, Springer LNCS*, 4090:168–186.

Zollers, A. (2007). Emerging Motivations for Tagging: Expression, Performance, and Activism. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*.