Master's Thesis

Pragmatic Analysis of Tagging Motivation in Social Tagging Systems

Hans-Peter Grahsl

grahsl@student.tugraz.at
Matr. No. 0230186

Institut für Wissensmanagement
Technische Universität Graz

Knowledge Management Institute
Graz University of Technology

Head of Institute: Univ.-Prof. Dr. rer. nat. Klaus Tochtermann

Supervisor: Dipl.-Ing. Dr. techn. Markus Strohmaier

Graz, July 2010

For my family again... more than ever.
And for Andrea, my beloved girlfriend.

Abstract

Social tagging systems allow users to annotate different kinds of web-resources (URLs, photos, publications etc.) by means of a freely chosen and unbounded vocabulary of terms, so-called tags. While earlier research has been primarily focused on the analysis of the structure and the dynamics of social tagging systems, research has recently started to study the motivation behind tagging. This master's thesis aims to contribute towards a deeper understanding about the tagging characteristics between two fundamentally different types of tagging motivation, namely categorization versus description. So-called categorizers primarily utilize tags to structure and maintain a navigational aid to their resources. They establish a personal vocabulary of tags which tends to stabilize quickly and exhibits a balanced tag usage frequency. Describers have the predominant goal of annotating their resources in very detailed manner in order to support retrieval activities. Since they are using tags in a descriptive, ad-hoc manner their tag vocabulary typically grows much bigger and also shows an unbalanced distribution of tags. Based on 10 tagging datasets which have been acquired from 6 social tagging systems (BibSonomy, CiteULike, Delicious, Flickr, Diigo and MovieLens) this thesis systematically compares tagging practices of categorizers and describers. For that purpose, a pragmatic analysis has been conducted using a number of statistical measures, which aim to reflect on different intuitions about the tagging characteristics of categorizers and describers. Additionally, the thesis includes empirical results from a human subject study. During a binary classification task, assessing whether users are either categorizers or describers, it has been investigated which of the selected measures come closest to human judgement. The key findings of this thesis are related to a number of tagging characteristics that have been analyzed for categorizers and describers. The results in this thesis demonstrate, that it seems feasible to automatically identify differences in users' tagging pragmatics by means of simple, yet robust statistical measures.

Zusammenfassung

Kooperative Verschlagwortungssyteme erlauben es Anwendern, unterschiedliche Arten von Web-Ressourcen (URLs, Fotos, Publikationen etc.) mittels eines frei wählbaren und offenen Vokabulars, sogenannten "Tags" zu annotieren. Während die Forschung zu Beginn primär auf die Analyse der Struktur und der Dynamik von kooperativen Verschlagwortungssystemen fokusiert war, kam es kürzlich zur Untersuchung von Motivationsstrukturen, die der Verschlagwortung zu Grunde liegen. Die vorliegende Masterarbeit zielt auf ein tieferes Verständnis hinsichtlich der Verschlagwortungscharakteristiken von zwei grundverschiedenen Typen von Motivation ab - Kategorisierung versus Beschreibung. Sogenannte "Kategorisierer" verwenden Tags primär zum Aufbau und zur Pflege einer hilfreichen Navigationsstruktur ihrer Ressourcen. Dazu etablieren sie ein persönliches Vokabular an Tags, das dazu neigt, sich schnell zu stabilisieren und eine gleichmäßige Verwendungshäufigkeit der Tags aufweist. "Beschreiber" haben das vordergründige Ziel, Ressourcen äußerst detailliert zu annotieren, um die Suche möglichst gut zu unterstützen. Da sie ihre Tags ad-hoc und beschreibend einsetzen, wächst ihr Tag-Vokabular typischerweise viel stärker und weist zudem eine ungleichmäßige Verteilung auf. Basierend auf 10 Verschlagwortungsdatensätzen, die von 6 unterschiedlichen kooperativen Verschlagwortungssystemen (BibSonomy, CiteULike, Delicious, Flickr, Diigo und Movielens) akquiriert wurden, werden innerhalb dieser Masterarbeit die Verschlagwortungspraktiken von Kategorisierern und Beschreibern systematisch verglichen. Zu diesem Zweck wurde eine pragmatische Analyse durchgeführt, die auf ausgewählten statistischen Metriken basiert, welche unterschiedliche Intuitionen der Verschlagwortungscharakteristiken von Kartegorisierern und Beschreibern widerspiegeln. Die Masterarbeit beinhaltet überdies noch empirische Ergebnisse einer qualitativen Benutzerstudie. Im Zuge einer binären Klassifikationsaufgabe zur Abschätzung, ob Benutzer eher Kategorisierer oder Beschreiber darstellen, wurde untersucht, welche statistischen Metriken dabei am ehesten der menschlichen Beurteilung entsprechen. Die zentralen Ergebnisse dieser Masterarbeit beziehen sich folglich auf eine Reihe ausgewählter Verschlagwortungscharakteristiken, welche vergleichend für Kategorisierer und Beschreiber analysiert wurden. Die Ergebnisse zeigen, dass es mittels einfachen jedoch robusten statistischen Maßen möglich ist, die Unterschiede in der Verschlagwortungspragmatik von Benutzern automatisch zu identifizieren.

Acknowledgements

If I have seen farther, it is by standing on the shoulders of giants.
– Sir Isaac Newton

First and foremost, I owe a debt of gratitude to Dr. Markus Strohmaier for being an excellent mentor, for his ambitious guidance and for his exceptional assistance during the accomplishments of my thesis. His incessant eagerness to get the most out of a student's work eventually led to the thesis at hand.

For their generous support in the research of my thesis, I would like to further acknowledge the members of the Knowledge Management Institute. In particular, I would like to thank Christian Körner for numerous inspiring discussions as well as criticism and Roman Kern for sharing some of his outstanding programming skills as well as his knowledge in the field of information theory.

I would be remiss if I didn't acknowledge the manifold support of my parents, Ingrid and Johann. Besides showing me endless ways to enjoy my life, they both have inspired my curiosity and passion for life-long learning. I will always be grateful for their patience, understanding and love. It is partly also due to their continous encouragements that I managed to finish my studies, despite the fact, that I have been working full-time for the last five years.

Finally, I wholeheartedly express my gratitude to my beloved girlfriend Andrea Maier. She is an incomparable and adorable woman that supports me in good times and bad. Thanks for your love and numerous heart-warming moments.

Hans-Peter Grahsl, July 2010

Deutsche Fassung:
Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008
Genehmigung des Senates am 1.12.2008

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

Graz, am ……. ……..
(Unterschrift)

Englische Fassung:

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

……. ……..
date (signature)

Contents

List of Figures

List of Tables

Listings

Chapter 1

Introduction

> *What I think is coming instead are much more organic ways of organizing information than our current categorization schemes allow, based on two units – the link, which can point to anything, and the tag, which is a way of attaching labels to links. The strategy of tagging – free-form labeling, without regard to categorical constraints – seems like a recipe for disaster, but as the Web has shown us, you can extract a surprising amount of value from big messy data sets. – Clay Shirky*

1.1 Information organization

1.1.1 Traditional organization strategies

For centuries, library science is striving for well-suited forms to organize information. Enormous efforts have been put into two fundamental strategies - categorization and classification - to bring order to large amounts of information. At first both strategies are briefly introduced and explained. A direct comparison that points out the main differences is presented later on.
According to [Jacob, 2004]

> *Categorization divides the world of experience into groups or categories whose members share some perceptible similarity within a given context. That this context may vary and with it the composition of the category is the very basis for both the flexibility and the power of cognitive categorization.*

At this stage, it is important to mention what is known as the "classical theory of categories" which is based on the following three propositions (cf. [Smith and Medin, 1981]):

- The definition (i.e. intension) of a certain category comprises the union of essential features which identify an entity's membership (i.e. extension) of the category.

- A category's essential features are individually necessary and jointly sufficient to determine (binary) affiliation with the category, causing the boundaries of the category to be fixed and rigid.
- With regard to hierarchies of categories, a member of a category A nested within a superordinate category B must exhibit not only the set of essential features determining membership in its direct category but also all of the essential features necessary for the membership in any superordinate category (B and all ancestors).

[Jacob, 2004] defines classification within the field of library and information science (LIS) as follows:

> *In LIS, the term classification is used to refer to three distinct but related concepts: a system of classes, ordered according to a predetermined set of principles and used to organize a set of entities; a group or class in a classification system; and the process of assigning entities to classes in a classification system. Classification as process involves the orderly and systematic assignment of each entity to one and only one class within a system of mutually exclusive and non-overlapping classes.*

Additionally, [Jacob, 2004] highlights the following two different approaches towards a bibliographic classification scheme.

- **Deductive approach:** Top-down schemes enumerate a set of mutually exclusive classes. An enumerative classification scheme begins with a universe of knowledge and a theory of organization or set of principles that establishes the conceptual structure of the scheme. Whether the universe encompasses all knowledge or is limited to a specific domain, construction of the scheme involves the logical process of division and subdivision of the original universe such that each class, or each level of classes in the structure, is differentiated by a particular characteristic or property (e.g. the property "color" or "shape"). The result is a hierarchical structure of generic (genus / species) relationships wherein each subordinate class is, theoretically, a true species of the superordinate within which it is nested.
- **Inductive approach:** Bottom-up schemes are generated through a process of analysis and synthesis. Construction of the faceted structure begins with analysis of a universe of knowledge to identify the individual elements—properties and features— of the universe. These elements are then organized into mutually exclusive groups on the basis of conceptual similarity, and these groups are, in turn, arranged in successively larger groupings to form facets (aspects) that can be used to represent entities in the universe. In this way, meaningful relationships are established not only between the elements in a group but between the groups themselves. The

result is not a classification scheme but a controlled vocabulary of concepts and their associated labels that can be used, in association with a notation and a prescribed citation order, to synthesize the classes that will populate the classification scheme.

At first glance, classification and categorization seem to be very similar strategies. This "problem" obviously stems from a misconception which is rooted in literature where the terms *classification* and *categorization* are frequently used interchangeably. [Jacob, 2004] blames such imprecision in terminology to disguise the fact that - though being similar - they are nonetheless distinct organization strategies. Both are mechanisms to establish order by grouping according to relatedness and similarity respectively but they are different, for instance, in how that order is carried out. Table 1.1 lists six systemic properties aiming at the comparison of classification and categorization systems. It is actually their differences that have signification implications, both for the constitution of information environments and the design of information systems.

property	**categorization**	**classification**
process	creative synthesis of entities based on context or perceived similarity	systematic arrangement of entities based on analysis of necessary and sufficient characteristics
boundaries	Because membership in any group is non-binding, boundaries are "fuzzy"	Because classes are mutually-exclusive and non-overlapping, boundaries are fixed
membership	Flexible: category membership is based on generalized knowledge and/or immediate context	Rigorous: an entity either is or is not a member of a particular class based on the intension of a class
criteria for assignment	Criteria both context-dependent and context-independent	Criteria are predetermined guidelines or principles
typicality	Individual members can be rank-ordered by typicality (graded structure)	All members are equally representative (ungraded structure)
structure	Clusters of entities; may form hierarchical structure	Hierarchical structure of fixed classes

Table 1.1: Comparison of categorization and classification [Jacob, 2004]

1.1.2 Rethinking information organization

When moving towards the digital domain, a lot of what has been learnt from library and information science with regard to classification and categorization might be revisited. Generally, it can be considered a very difficult task to come up with totally coherent schemes when performing classification or categorization tasks, which is also true for experts. [Shirky, 2005] argues that, even in powerful and almost perfectly appearing categorization schemes it might be that there are some oddities or contextual errors involved (e.g. the category of noble gases in the periodic table of elements). Especially in domains where essence is much less obvious than for instance, in the chemistry example, one might get frustrated quickly and end up with inconsistency problems. He further points out

a partial list of characteristics - for the participants and the domain to be organized - that could help to make ontology a workable classification strategy. The domain itself should exhibit: a small corpus, formal categories, stable entities, restricted entities and clear edges. Additionally, participants should be: expert catalogers, an authoritative source of judgement, coordinated users and expert users. A classical example of a classification scheme that works, because lots of these characteristics hold, is the 4th version of the psychiatrists' Diagnostic and Statistical Manual (DSM-IV) which has the American Psychiatric Association as its authoritative source. It theoretically enables psychiatrists in the U.S. to make the same judgement about mental illnesses, when provided with the same list of symptoms. Other famous examples within the field of library science are the Library of Congress Classification (LCC) or the Dewey Decimal Classification (DDC). [Shirky, 2005] further indicates, that when looking at the Web it becomes quite clear why ontology engineering efforts seem to be a merely imperfect fit. This is due to the fact, that basically all characteristics that are mentioned before, do not apply to the domain of the Internet which is a massive corpus without global authority that primarily serves average users.
Clearly, it seems to make sense to rethink traditional forms of information organization within digital environments. There are no such hard constraints like shelves which impose physically related restrictions on the chosen organization strategy. It therefore seems feasible to leave strict binary categorization behind and move forward to strategies that allow for grouping in a probabilistic way, where only a certain (representative) percentage of people deems entities to be members of a specific category.
Considerations like these, might have led towards a wide-spread introduction and the recently prevalent adoption of (social) tagging systems which are briefly introduced next.

1.2 (Social) Tagging Systems

Generally speaking, *tagging* describes the process of annotating information with keywords or labels, so-called *tags*. Applications that provide users with the possibility of tagging are usually referred to as *tagging systems*. During the last years, there has been a massive adoption of this kind of information organization among different user communities on the web. This was due to an increasing number of launched online systems (e.g. Delicious, Flickr, CiteULike, Diigo and others) for the purpose of organizing and sharing arbitrary content by means of tagging. It is especially the underlying notion of the Web 2.0, where everyone can contribute, that caused tagging actions of individuals to become a *collaborative, social practice*. An important aspect in this regard is the public sharing of particular tags as well as resources.
The basic idea behind tagging is nothing new. However, it fundamentally differs from traditional strategies in two ways. First, instead by an authority or an ex-

pert, the organization of contents is done by ordinary users themselves. Second, users are allowed to freely choose any tags they personally find appropriate, since there is no controlled vocabulary. For this reason, tagging allows much greater flexibility in organizing information than do formal classification schemes, which are the direct opposite. According to [Mathes, 2004] the tagging approach also lowers the barrier of cooperation, since *"groups of users do not have to agree on a hierarchy of tags or detailed taxonomy, they only need to agree, in a general sense, on the 'meaning' of a tag enough to label similar material with terms for there to be cooperation and shared value."* Another important benefit of tagging is pointed out by [Sinha, 2005] who argues that tagging prevents users to suffer from so-called "post activation analysis paralysis". This refers to the fact, that in traditional categorization, the decision is further complicated because users have to instantly decide for a certain (i.e. hopefully the right) category to ensures later retrieval. Tagging, on the other hand, allows users to externalize all activated concepts by virtually assigning an unlimited number of tags to content which should significantly alleviate the fear for making the wrong decisions.
Although it is tempting to think of tagging as the panacea for several problems related to traditional categorization and classification schemes, there are of course a number of drawbacks that come along with the tagging approach. For instance, the ambiguity in the meaning of tags or the redundancy caused by synonym tags. Both are very likely to negatively affect tag-based navigability and retrieval respectively.

1.3 Terminology and Definitions

Related to definitions found in literature, this section introduces a conceptual model used to formally describe the different components which constitute a tagging system.
[Halpin et al., 2007] list the following three main entities of a tagging system: i) the users of the system (who do the tagging), ii) the tags themselves (keywords or labels used for the annotation) and iii) the resources being tagged (generally of arbitrary type). What results from users' collaborative tagging activities are emergent structures comparable to bottom-up categorization schemes, which commonly referred to as *folksonomies*. Actually the term "folksonomy" is a so-called portmanteau word - a blend of the two terms "folk" and "taxonomy" - which has been coined by [Vander Wal, 2004].
According to a slightly simplified version of the formal definition found in [Hotho et al., 2006], a *folksonomy* is usually explained on a tripartite hypergrah graph model and can thus be defined by the tuple $F := (U, T, R, Y)$. U, T and R are finite disjoint sets representing *users*, *tags* and *resources*, while Y is a ternary relation $Y \subseteq U \times T \times R$ between them. The elements $y \in Y$ are called *tag assignments*. *Posts* are tuples in the form (u, T_{ur}, r), where $u \in U$, $r \in R$ and $T_{ur} := \{t \in T | (u, t, r) \in Y\} \neq \emptyset$. In other words a *post* consists of one or more

tag assignments that a certain user has created for a specific resource. They further define what is later referred to as a *personomy*, which can be regarded as a user's personal contribution to the folksonomy F. In fact, the personomy P_u of a certain user is the reduction of F to $u \in U$. It is often convenient to additionally define the *tagging vocabulary* of a certain user as $T_u := \{t \in T | \exists r : (u, t, r) \in Y\}$ representing the set of distinct tags which have been used at least once. Analogously, the set of resources having been tagged by a certain user is given by $R_u := \{r \in R | \exists t : (u, t, r) \in Y\}$.

1.4 Goals and Contributions

The primary goal of this Master's Thesis is to study the tagging motivation of users on different social tagging systems by means of a pragmatic approach. This basically means that users' tagging practices (i.e. their tag usage patterns) are investigated rather than analyzing tagging semantics. Based on statistical measures and concepts of information theory, two fundamentally different tagging motivations - categorization vs. description (see section 3.1.1 and table 3.1 for a detailed explanation and comparsion respectively) - are investigated with regard to differences in their underlying tagging structures and characteristics. For that purpose, this thesis builds upon and further elaborates on findings from recent publications ([Körner, 2009], [Strohmaier et al., 2010a], [Körner et al., 2010b], [Kern et al., 2010], [Körner et al., 2010a]) within our research group (Knowledge Management Institute, Graz University of Technology). During the course of this thesis, the following selected research questions are addressed:

- Are simple statistic measures capable to distinguish categorizers from describers within and across social tagging systems?
- How are suggested measures for the detection of tagging motivation correlated to one another?
- What characteristics of tagging pragmatics are amenable to an automated statistical analysis?
- Would such characteristics differ significantly when comparing tagging activities of individuals adhering to categorization with those who follow the description approach?

Research activities that have been performed include:

- The acquisition of 10 user-centric tagging datasets originating from 6 different tagging systems (see table 3.3 for detailed facts)
- A short survey of existing theories for individual tagging motivation and behavior.

- A qualitative and quantitative analysis of tagging pragmatics based on the acquired datasets.
- The identification and definition of different characteristics and metrics with emphasis on tagging motivation.
- A qualitative and quantitative evaluation of the identified tagging characteristics and metrics.

The thesis aims to make contributions towards a deeper understanding of the different characteristics of tagging motivation. Based on the acquired datasets, it is shown in which aspects and to what degree there are statistically measurable distinctions between a categorization and a description approach to tagging.

Chapter 2

Related Work

> *Information is a source of learning. But unless it is organized, processed and available to the right people in a format for decision making, it is a burden, not a benefit. – William Pollard*

This chapter highlights related literature in the context of this. It starts with relevant work about personal information management by pointing to a number of interesting field studies that have been carried out in the physical as well as digital domain. The following section deals with research work directly focused on aspects of tagging. Finally, selected theoretical concepts that play an important role during the course of this thesis are explained.

2.1 Personal Information Management

Personal information management (henceforth PIM) is an umbrella term, defined in [Jones, 2007] as follows:

> *Personal information management (PIM) refers to both the practice and the study of the activities a person performs in order to acquire or create, store, organize, maintain, retrieve, use and distribute the information needed to meet life's many goals (everyday and long-term, work-related and not) and to fulfill life's many roles and responsibilities (as parent, spouse, friend, employee, member of community, etc.). PIM places special emphasis on the organization and maintenance of personal information collections in which information items, such as paper documents, electronic documents, email messages, web references, handwritten notes, etc., are stored for later use and repeated re-use.*

In a workshop contribution [Jones, 2008] discusses several reasons when to consider information as personal. Table 2.1 summarizes the suggested six senses of personal information listing examples, issues, current search support as well as

potential future improvements towards more personalization. The offered senses, though representing broad distinctions, are not meant to be sharply separated from one another. Instead they shall be holistically understood, as in their union, they actually exclude very little. Furthermore it is suggested to employ them as yardstick for the assessment of existing and upcoming tools and technologies in the field of PIM.

personal relation	**examples**	**issues**	**current support**	**future personalization**
Controlled by, owned by me	Messages in our email accounts; files on our hard drives	Security against break-ins or theft, backups, virus protection, etc.	Desktop search facilities.	Suggest places to keep an item, suggest items to be archived. Identify versions of the same item.
About me	Credit & medical information, tax records, histories of Web browsing & library checkouts	Who sees what when (under which circumstances)? How is information corrected or updated?	"Self-googling" on the Web.	Agents to alert when information about us is, accessed, transferred, modified.
Directed to me	Phone calls, drop-ins, TV ads, web ads, pop-ups.	Protection of us and our money, energy, attention and time.	Junk email filters. Rules and alerts.	Filters on all digital input channels that learn from usage patterns.
Sent (posted, provided) by me	Email, blogs, personal web sites, published reports and articles.	Who sees what when? Did the message get through?	-	Search to track where information we send / post / publish goes and how it's used.
(Already) experienced by me	Email that we've read, web pages we've browsed, billboards we've seen...	How to get back to information again later? Are we getting a "balanced diet" of information?	Integrated desktop & web searches.	Re-finding items no matter which device they are on.
Relevant (useful) to me	Somewhere "out there" is the perfect vacation, house, job, lifelong mate. If only I could find the right information!	If only we knew (had some idea of) what we don't know. How to filter out or otherwise avoid information we don't wish to see? (How to do likewise for our children?)	Content Web filters block access to offensive web pages. Agents to send updates.	Queries expressing persistent interests are derived from and anchored to topic folders.

Table 2.1: Overview of the six senses of personal information [Jones, 2008]

2.1.1 Physical domain studies

- In a study using expert-interviews [Malone, 1983] describes a series of real-world office scenarios focusing on how people are performing so-called "desk organization" within their natural working environments. The interviews were composed of (un)structured questions and ended with a retrieval task, where co-workers chose "probe" documents that later had to be found by the respondents, whose offices varied greatly concerning the level of organization. At one extreme, there were "messy" offices exhibiting miscellaneous piles of paper that represent ill-defined groups of things to do. The other extreme, were "neat" offices featuring precisely characterized information stored in files. This observation is directly related to the major units

of desk organization namely, files and piles, both offering rather different ways to collect groups of information into larger units. While the first are units of elements which are normally explicitly titled and systematically arranged, the latter are usually units of untitled elements having no special order other than haphazard. The major functions of desk organization are finding an reminding. For both these functions, the author lists several implications with regard to the design of computer-aided information environments. Computer systems can help to minimize the mechanical and cognitive load of users throughout the three processes involved in finding tasks: creating classifications (helping to build and maintain multileveled classification systems), classifying information (allowing for multiple and deferred classification and offering automatic classification) and retrieving the information itself (supporting the specification of multiple search dimensions in parallel). Apart from that, computer systems should also cover the reminding function. Thus, users should be informed about pending tasks without them being explicitly requested (e.g. by explicitly classifying documents using a priority scale or by implicitly taking date and time information into account). The author concludes with four major forces leading to the creation of piles in office environments:

  - The mechanical difficulty of creating labeled file folders, binders, and so forth, especially if multiple levels of classification are desired.
  - The cognitive difficulty of creating appropriate categories and deciding how to classify information in a way that will be easily retrievable.
  - The desire to be reminded of tasks to be done.
  - The desire to have frequently used information easily accessible.

- [Lansdale, 1988] investigates the subtle and complex aspects of psychological processes that are involved during the management of personal information. He discusses two major problems affecting PIM by pointing to a lot of relevant research from others. First, there is the common issue of categorising items, both in terms of which categorisations to use and in later recalling the corresponding labels of the introduced categories. Not only does it seem infeasible to generate category names which can be used unambiguously, but information in every-day life often falls into a number of overlapping, fuzzy categories, too. The second problem relates to the fact that humans usually remember far more about documents than what can normally be exploited at the stage of retrieval. Though it appears that human's memory for very detailed information can be poor, recall can be remarkably improved if the same information is holistically understood (i.e. it is embedded in a meaningful context when committed to memory). Besides, it is noted that processes during information retrieval in the human mind totally differ from filing actions (e.g. library systems) where items are accessed by location rather than their meaning. The author further explains

why computer-based information management methods (direct file access, relational databases, first desktop interfaces and spatial data management facilities) that existed until then, failed to broadly support humans in PIM activities. It basically comes down to the fact, that while such systems are capable to aid with handling information, they do not really alleviate retrieval. Unsurprisingly, this can be attributed to a technology driven design, focusing on the automation of existing patterns and strategies of information management, which does not seem to pose an efficient way towards substantial progress in this field. The main contribution of this work is a simple framework for information retrieval which combines two fundamental psychological processes involved: a) recall-directed search (using memory about the required item to narrow down the result set) followed by b) recognition-based scanning (the scanning process to be undertaken when recall does not result in a unique item). The final remarks deal with research challenges that come along with the development of systems based on this framework. Most emphasis is placed on the important dilemma, that the success of retrieval heavily depends upon the efforts people put into the storage process of information.

2.1.2 Digital domain studies

- [Whittaker and Sidner, 1996] address the problem of email overload in PIM which results from the fact that people started to use email for other tasks than is has been originally devised for. Apart from the intended asynchronous communication purpose, email gradually began to serve task management as well as information archiving activities. Their study includes quantitative (collected directly from the corresponding inboxes) as well as qualitative (gathered by semi-structured interviews) evidence that overload in email applications has eventually become a problem that basically manifests itself in cluttered inboxes. This in turn leads to the users' inability to effectively communicate, to backlogs of unanswered mail and to deficiencies in finding information. According to the authors, an inbox should only consist of a few unread messages at any point in time, with the rest of the messages being filed. The quantitative data of their study proofs the opposite to be true, which is most likely related to four specific types of emails that are not discharged at once: to-dos, to-reads, messages with indeterminate status and ongoing correspondence. The authors list a number of potential causes for the filing problem that clearly reflect what [Malone, 1983] and [Lansdale, 1988] found out. During their study [Whittaker and Sidner, 1996] observed three strategies to fight inbox clutter. There are so-called "no filers", "frequent filers" and "spring cleaners". Finally, they propose design guidelines, thereby potential solutions, how future email applications can better assist humans concerning the three major problems involved in email activities:

  - asynchronous communication problem: automatically mark emails belonging to the same conversation with thread IDs to be able to view messages grouped by threads.
  - information archiving problem: support the temporary buffering of incoming messages using techniques to automatically cluster semantically related emails
  - task management problem: means to mark particular inbox items as requiring further / future action, extended by an explicit reminder functionality for postponed tasks

- [Abrams et al., 1998] conducted a user study to analyze the "personal Web information spaces" (i.e. bookmark archives) of Internet users. They describe bookmarks as shortcuts to frequently visited web sites or historical pointers to information that may otherwise be lost. Survey respondents stated to decide on bookmarking actions of web pages as of their following five criteria: general usefulness, quality, personal interest, frequency of use and potential future use. The three main benefits of utilizing bookmarks are to a) reduce the mechanical and physical load of managing URL addresses, b) facilitate the return to related page groups and c) enable users to build their personal Web information space. Moreover, the authors observed four major metaphors for using bookmarks: a) identification (conceptualizing bookmarks as small tags / labels placed on information), b) collection (pulling specific information out of the vast amounts of data on the web), c) movement (traveling through the vast information space implying destinations, landmarks and paths) and d) episodes (describing kind of a chronological navigation history through the web). The authors further describe organization methods for bookmarks as well as different habits of bookmark users, both which can be related to some extent to what has been found in [Whittaker and Sidner, 1996] for the management of personal email. To conclude with, they point out several problems with the handling of bookmarks which directly lead to suggested design considerations for future bookmarking tools. In the following four fundamental areas tools are needed that:

  - organization: scale well, minimize organizational efforts and provide ad-hoc filing mechanisms for users
  - visualization: are capable of visualizing large numbers of bookmarks without hampering users
  - representation: allow for the renaming of bookmarks to something more descriptive and memorable for users
  - integration: fit naturally into the browser environment and support publishing and sharing actions of personal bookmark archives well

- [Alvarado et al., 2003] explored in a qualitative user study (semi-structured interviews) what ways people would find natural for handling personal information. As opposed to previous work from others that mostly focused on single aspects of PIM, the authors of this study chose a broader approach to find out how people generally handle electronic information by simultaneously taking into account email, file and web information. Despite managing rather complex information spaces most respondents felt to be in control of their data. One of the key findings was that people frequently employed a search strategy based on contextual information rather than just performing a keyword search. This is basically a distinction between the two search strategies called "orienteering" vs. "teleporting". A number of identified search goals were collapsed into three main categories of information need: specific information, general information and specific documents. A part from orienteering to specific documents it was surprising though, that people were more likely to orienter to specific information than to general information. This implies that people are obviously maintaining vast amounts of contextual information even about very specific pieces of data. Especially the investigation of email management clearly showed two different groups of people, namely filers and pilers - as already been identified by [Malone, 1983] - which exhibited different search tactics while looking for information.

- While studying PIM in the context of email management [Civan et al., 2008] directly compare two forms of information organization. They set out to explore whether it is better to organize emails traditionally - placing emails into folders - or in the recently prevailing way by tagging them with labels. The study has been conducted on two popular web-based email systems - GMail (Google) and Hotmail (Microsoft) - the first supports tagging while the latter is based on the filing approach. Each participant to the study had to complete three stages: a) an initial interview, b) several information organization tasks in both, GMail and Hotmail which were delivered over five days and c) a final follow-up session. Results of the experiment showed that there are similarities between placing and tagging information. Both models are comparable concerning their retrieval performance, their evolution in mappings between articles and folders or labels over time, and their limitations to fully and explicitly express one's internal conceptualization by sketches afterwards. On the other hand there were a number of key differences (see table 2.2) which could be identified for the input side (i.e. keeping & organizing information) and on the output side (i.e. re-finding relevant information).

differences in organizing information	
cognitive vs. physical effort	there seems to be agreement on higher cognitive efforts needed for the folder based approach, but more physical effort involved with tagging
hiding vs. keeping information seen	folders better support information hiding to avoid clutter, tags can be directly used to emphasize information without workarounds (e.g. special "to-do" folders)
differences in re-finding information	
flexible vs. systematic searching	while tags offer the flexibility to provide multiple paths back to look up an item, folders seem to allow for more systematic routes back to information whose exact location is forgotten
re-finding cues offered by folders vs. labels	while informational cues of labels offer serendipitous encounters, folders provide visual cues allowing for the use of spatial memory and recognition during retrieval

Table 2.2: Key differences in placing vs. tagging information [Civan et al., 2008]

2.2 Social Tagging

2.2.1 Tagging behavior, motivation and incentives

What is the motivation or intention for people when tagging web resources (e.g. URLs, photos, publications etc.)? From 2005 onwards, numerous research studies try to investigate why users are tagging and in what ways they are using tags. While early work is predominantly based on anecdotal evidence, more recent research tries to formulate theoretical models built upon larger datasets. Besides incorporating human subject studies for qualitative evaluation purposes, it became more and more important to include quantitative analysis as well. However, until now, there is neither a broad consensus nor some kind of commonly accepted taxonomy to classify different tagging motivations. Instead there are varying perspectives to look at tagging, all having their specific focus and abstraction level. To get valuable insights into various forms of tagging behavior, motivation and incentives, a chronological overview of important literature from 2005 to 2009 is presented next.

- [Coates, 2005] discusses two strictly opposite tagging approaches. People could understand tags as a modern, flexible replacement to folders and thereby follow some kind of filing behavior ("presumed Delicious approach"). On the other hand they could simply apply any tags that make sense to them in order to characterize web resources ("intentional Flickr approach").

- In [Hammond et al., 2005] the reasons for tagging are rooted in more egocentered or socially oriented tagging activities. The two extreme positions

are a) the so-called "selfish" approach, where people are tagging their own resources to achieve their personal goal and b) the "altruistic" approach, where people are annotating the resources of others for yet others to find them.

- According to [Marlow et al., 2006], motivators to tagging can be roughly classified into organizational and social practices. People motivated by the first use tagging predominantly as an alternative method to well-structured filing. Those that follow the latter mostly represent a highly social and collaborative approach to tagging, by choosing tags mainly to communicate their opinions, feelings and specific qualities of the tagged resources. Additionally, a list of six incentives that express the potential motivations behind tagging is presented which includes: future retrieval, contribution and sharing, attract attention, play and competition, self presentation and opinion expression. The authors further assume a majority of the users to be driven by multiple incentives simultaneously.

- During a human subject study, [Sen et al., 2006] related user tasks, being accomplished by means of annotation activities, to various forms of tagging incentives. User tasks include: self-expression, organizing, learning, finding and decision support. Moreover, they analyzed whether factual, subjective or personal tags are most appropriate to fulfill certain tasks and which of these three types generally results in the highest user satisfaction.

- In a user study comprising twelve Delicious-savvy participants, [Wash and Rader, 2007] discovered producer vs. consumer incentives for Delicious. They investigated the three major activities users engage in when participating in the community - namely bookmarking, tagging and information seeking. For each of these activities they identified related motivators, which include among others: for the activity of bookmarking (to keep track of useful web pages, to share web pages with other people and to achieve social recognition from the Delicious community), for tagging (to organize resources for later retrieval using proper heuristics when selecting tags), for information seeking (to browse Delicious in order to find either novelty, topical or social information).

- [Ames and Naaman, 2007] tackle the question why people annotate mobile and online media by conducting a qualitative study with ZoneTag and Flickr users. From interviews they deduce a taxonomy for tagging motivations along the dimensions of function (organizational or communicational) and sociality (social or self) which results in four possible high-level combinations of tagging motivations. Based on the observations and in order to best support all four categories within the taxonomy, they draw the following implications for the design of tagging systems: make annotations pervasive and multi-functional, make the annotation process as effortless as

possible, do not necessarily force users to annotate, whenever applicable allow annotations in desktop and web-based components for hybrid systems, provide relevant tag recommendations but use them wisely / with caution in general.

- Based on a survey of 142 users that in corporates the systems Flickr, YouTube, Delicious and Connotea [Heckner et al., 2009] propose a model towards information behavior in social tagging systems. The goals of users can be associated to two main functional areas. Users' tagging activities might aim at personal information management (=PIM, cf. [Lansdale, 1988] and [Boardman and Sasse, 2004]) or their main concern is with resource sharing efforts. Users from the first group (strongly driven by information retrieval aspects) use tags to manage their digital assets in order to keep them findable for later use. Those from the second group (partly driven by social reputation aspects) want peers to profit from their own efforts. Additionally, it becomes apparent from their results that Delicious users seem to be mostly interested in PIM while it is the opposite for YouTube users, that predominantly aim at resource sharing.

- [Nov et al., 2009] developed a research model to study possible influence factors related to a user's membership tenure in the photo sharing community of Flickr. They take into account both, motivational factors (measurable using surveys) as well as structural properties (derivable from the underlying system data). They describe four types of individual motivations (extrinsic and intrinsic ones) including enjoyment, commitment, self-development and reputation. Their hypothesis that all four of them are positively correlated to a high photo sharing activity of users could only be partly validated. They found that users showing more commitment and greater structural embeddedness in the community seem to share a higher number of photos. However, no correlations could be detected for enjoyment and reputation factors. Against intuition, they even observed a negative correlation between self-development and the photo sharing activity per year.

2.2.2 Different kinds of tags

- [Golder and Huberman, 2005] try to explain tagging behavior on Delicious by means of different tag kinds. They find the scope of tags to be less interesting than the functions tags serve. Based on concrete functions tags may perform on Delicious, they introduced seven individual types of tags to: identify what (or who) a bookmark is about, identify what a bookmark is, identify who owns a bookmark, refine categories, identify qualities or characteristics, to perform self-referencing and to organize tasks. Finally, they conclude that a significant amount of tagging is done for private use rather than public benefit and that information tagged by others is only partly useful for individuals.

- In order to study tag suggestion techniques based on data from Yahoo's My Web 2.0, [Xu et al., 2006] introduced a taxonomy of tags which incorporates the following five categories
  - content-based tags: either directly describe a resource's content or the categories it belongs to. Usually such tags are specific terms and they are frequently encountered in Yahoo's My Web2.0.
  - context-based tags: aim to provide contextual information to resources (e.g. where or when was the resource created/saved/annotated)
  - attribute tags: represent inherent characteristics of resources which can be explicit but may be implicit, too.
  - subjective tags: are primarily applied to express emotions, feelings, opinions or preferences of individuals.
  - organizational tags: either help to remember (pending) tasks like "to-read" or serve to identify personal assets such as "my-dog".

  Different types of tagging motivation tend to be intrinsically tied to a predominant utilization of tags belonging to any subset of these categories.

- By introducing the notion of purpose tags, [Strohmaier, 2008] addresses a well known problem in the field of human computer interaction. The so-called "gulf of execution" represents the cognitive gap spanning between the content or functionality of a system and the goals that humans have in their mind while using it. Purpose tags should have the potential to bridge this gap, by reflecting on a resource's intent rather than its content. Thus, the main priority of purpose tagging is to explicitly embody the possible goals that resources may serve to achieve. To give an example, a viable purpose tag to annotate the URL www.facebook.com would be "organize a high school reunion". Based on a prototypic purpose tagging application they found out that a) users quickly adapted to this new approach to annotating web resources and generated meaningful purpose tags, b) purpose tagging allows to accurately capture the various goals resources might help to achieve, and c) that the tagging vocabulary produced during purpose tagging significantly differs from the vocabulary consisting of traditional kinds of tags only.

2.2.3 Tagging and (web-)search

- [Hotho et al., 2006] suggested an algorithm called FolkRank which exploits the induced graph structures of folksonomies for ranking. Although being inspired by and based on the seminal PageRank (cf. [Brin and Page, 1998] and [Page et al., 1998]) algorithm, FolkRank needs adaptions in order to be able to apply a weight-spreading scheme on folksonomies. At first, the folksonomy structure itself is converted into an undirected tri-partite

graph structure $G_F = (V, E)$. V is a disjoint union of the set of users, resources and tags while E denotes the total set of edges linking all co-occurrences between different types of vertices (tags-users, users-resources, tags-resources). Analogous to the original formulation of PageRank and similar to motivations found in the HITS (cf. [Kleinberg, 1999]) algorithm, FolkRank employs a ranking scheme where resources become important when they are tagged with important tags by important users. Since the same holds (symmetrically) for tags and users, the graph's vertices are mutually reinforced by each other during the spreading of their weights. Weight-spreading itself is formally defined by $R \leftarrow c(\alpha R + \beta AR + \gamma P)$. R is a weight vector with one entry of each node, A is a row-stochastic version of the graph's adjacency matrix, P is a preference vector whose influence is controlled by constants β and γ, α is a constant damping factor used to avoid oscillation and speed up convergence and c is a constant normalization factor such that $||R||_1 = 1$. What remains is the problem that the graph G_F is undirected which would result in weights that flow back and forth along the same edge during multiple iterations. For that reason, the authors propose a differential approach to compute personalized rankings within the folksonomy. Two weights are calculated, one which incorporates user preferences (P, $\gamma > 0$) and another one which does not consider preferences ($\gamma = 0$). The final weight $R(v)$ of a node v in G_F is called the FolkRank of v and is given by the subtraction of the two weights. What is actually computed are thus the winnners and losers of the mutual reinforcement of resources by comparing the baseline (i.e. no preference vector) with given user preferences. The authors performed an evaluation of FolkRank an a large-scale dataset from Delicious. They could demonstrate, that FolkRank can be used to generate personalized rankings of the items in a folksonomy and that it also allows for the recommendation of users, tags and resources.

- [Bao et al., 2007] investigated the potential of social annotations to improve the quality of web search by taking into account tagging information available on Delicious. For that purpose, they worked out the following two algorithms to enrich current page ranking with:

  - **SocialSimRank (SSR)** is based on their first observation that annotations usually provide effective and multi-faceted summaries of web pages. SSR tries to quantify the similarity between annotations and the search query formulated by users. A naive way to calculate this similarity would be based on simple word matching. However, such approach would not only suffer e.g. from synonyms but also from the sparseness of annotation data for less popular web pages. SSR explores the annotations with similar meanings by building a weighted two-mode network of social annotations and web pages where the user counts represent edge weights. Having N_A annotations and N_P web pages M_{AP} is its $N_A \times N_P$ association matrix whose elements hold the

number of users who applied a certain tag to a specific web page. S_A ($N_A \times N_A$ matrix) and S_B ($N_B \times N_B$ matrix) result from the respective folding operations, where S_A holds similarity scores between annotations and S_B those between web pages. SSR is defined as an iterative algorithm working on these matrices to quantitatively determine the similarity between any pair of annotations until the corresponding $S_A(a_i, a_j)$ values converge.

– **SocialPageRank (SPR)** is proposed to measure how popular web pages are, which is based on their second observation that the total number of annotations per page are able to directly reflect its quality (in the sense of popularity). Again assuming N_A annotations, N_P web pages and N_U users, the corresponding association matrices are M_{PU} (between pages and users), M_{AP} (between annotations and pages) and M_{UA} (between users and annotations). Elements of M_{PU} are assigned with the corresponding counts of annotations applied by a certain user to a specific page - analogically, the elements of M_{AP} and M_{UA} are set. Initially, P_0 is the vector containing randomly assigned SPR scores. During each execution step i SPR then performs six matrix multiplications, thereby updating the intermediate result vectors P_i, U_i and A_i - denoting the popularity vectors of pages, users and annotations in the current iteration - until eventually P_i converges.

Experimental results of two different query sets demonstrated that SSR and SPR have the potential to significantly improve web search.

- The work of [Yanbe et al., 2007] suggests an enhancement of web search, by extending the classical PageRank (cf. [Brin and Page, 1998] and [Page et al., 1998]) approach with the notion of a ranking scheme which can, for example, be derived from social bookmarking services like Delicious. They introduce a measure called SBRank which is simply given by the total number of users having bookmarked a certain page. Their intuition behind this is the fact, that SBRank allows to capture the popularity (quality) of a page by "votes" (i.e. bookmarks) of content consumers rather than focusing on the link structure alone as this largely depends on content producers. Built on SBRank, the authors propose an enhanced search model capable of extending current searching facilities by considering what they call "complex queries". For that purpose this model provides for meta-data search and support for temporal, sentiment as well as controversial queries simultaneously. What results is a rather complex ranking scheme that defines the rank of a specific page as follows: $Rank(p_i) = (1+B(p_i)) * (1+F(p_i)) * (1+V(p_i)) * (1+C(p_i)) * (1+T(p_i,q)) * (1+T^{sen}(p_i,q)) * (1+S(p_i,t_{beg},t_{end}))$ $B(p_i)$ is the popularity estimate using a weighted linear combination of SBRank and SearchRank (i.e. the "classical" rank as provided from search engines). $F(p_i)$ relates to the temporal aspects and represents a freshness

level. $V(p_i)$ is a variance measure of applied bookmarks. $C(p_i)$ considers the controversial aspects expressed as the number of comments. $T(p_i, q)$ results from the similarity between the page tag and the query term vector, while $T^{sen}(p_i, q)$ relates to the similarity of the page sentiment and the query term vector. Finally, $S(p_i, t_{beg} \dots t_{end})$ is the proportion of bookmarks having been added during the period t_{beg} to t_{end} to the overall number of bookmarks for the page in the whole system.
A number of analytical experiments conducted on a prototypic implementation of such a hybrid, enhanced search model led to the following conclusions:

  - page quality measures can be improved by the popularity statistics found in social bookmarking systems
  - precision of relevance estimation can be increased by leveraging user generated meta-data
  - incorporating timestamps allows for both, time-aware popularity measures and temporal queries
  - tags enable page filtering based on sentiment characteristics or controversy levels

- The research work of [Heymann et al., 2008] set out to investigate to what extent user-generated data from the social bookmarking sites Delicious can help to improve web search. Their work is thus primarily focused on the capability of tags to guide users to valuable web content. A number of positive and negative factors (see tables 2.3 and 2.4) are identified and discussed, suggesting why web search might or might not benefit from social bookmarking.

Finding	**Conclusion**
Approx. 120,000 URLs are posted to Delicious each day	The number of posts per day is relatively small; for instance, it represents about $\frac{1}{10}$ of the number of blog posts per day
There are roughly 115 million public posts, coinciding with about 30-50 million unique URLs	The number of total posts is relatively small; for instance, this is a small portion (perhaps $\frac{1}{1000}$) of the web as a whole
Tags are present in the page text of 50% of the pages they annotate and in the titles of 16% of the page they annotate	A substantial proportion of tags are obvious in context, and many tagged pages would be discovered by search engine
Domains are often highly correlated with particular tags and vice versa	It may be more efficient to train librarians to label domains than to ask users to tag pages

Table 2.4: Cons why social bookmarking might not benefit web search [Heymann et al., 2008]

Finding	Conclusion
Pages posted to Delicious are often recently modified	Delicious users post interesting pages that are actively updated or have been recently created
Approx. 25% of URLs posted by users are new i.e. unindexed pages	Delicious can serve as a (small) data source for new web pages and to help crawl ordering
Roughly 9% of results for search queries are URLS present in Delicious	Delicious URLs are disproportionately common in search results compared to their coverage
While some users are more prolific than others, the top 10% of suers only account for 56% of posts	Delicious is not highly reliant on a relatively small group of users (e.g. $< 30,000$ users)
30-40% of URLs and approx. one in eight domains posted were not previously in Delicious	Delicious has relatively little redundancy in page information
Popular query terms and tags overlap significantly (though tags and query terms are not correlated)	Delicious my be able to help with queries where tags overlap with query terms
In a study most tags were deemed relevant and objective by users	Tags are on the whole accurate

Table 2.3: Pros why social bookmarking might benefit web search [Heymann et al., 2008]

In contrast to the work of [Bao et al., 2007] and [Yanbe et al., 2007] discussed above, [Heymann et al., 2008] highlight important limitations of social bookmarking data which corroborate their hypothesis that until then, Delicious may have lacked the size and distribution of tags necessary to effectively augment classical web search approaches.

- [Morrison, 2008] analyzed the information retrieval (IR) effectiveness by studying three different kinds of information retrieval (IR) systems on the web: search engines (Google, Microsoft Live, AltaVista), directory services (Yahoo Directory, Open Directory) and folksonomies (Delicious, Furl, Reddit). For that purpose, he conducted a shootout-style human subject study in which 34 participants manually crafted 103 queries reflecting their personal information needs. The top 20 results to all queries have been collected from each IR system in order for the participants to later judge their relevance in a binary manner (yes or no). The actual comparison of the effectiveness between the different IR systems was done using precision, relative recall and the retrieval rate (i.e. the proportion of documents returned to the maximum number possible). The key findings of this work can be roughly summarized as follows:
  - folksonomies from social bookmarking sites could be effective tools for IR on the web
  - the results from folksonomies overlapped with the results from search engines and directory services at similar rates stated within previous IR studies

  - documents found in result sets of both, search engines and folksonomies were more likely to be judged relevant than those only appearing in search engine results
  - the Delicious system performed better than Open Directory and approx. the same as Microsoft Live (though folksonomies had a lower precision than directory services and search engines in general)
  - there were hardly any statistically relevant differences between the recall of directory services and folksonomies
  - folksonomies may be helpful for certain information needs compared to directory services, but in general search engines were more effective
  - folksonomies performed better for news searches than the directory services but again search engines had much higher performance
  - folksonomies did particularly poorly with exact site searches as well as search with a short, factual answer

2.2.4 Tag generation models - simulating tagging activities

Observations of real world tagging activities have shown that there are several factors which influence tag choices of users to different degrees. One of the most obvious factors is directly related to past user activities (i.e. previous tag assignments). Besides, tag selection may also be affected by e.g. collaborative aspects in form of the whole tagging community, the content of resources or the amount of effort required for tagging. In the following, three widely-known tag generation models are presented, all which have its roots in a simple stochastic urn model.

- During their analysis of Delicious, [Golder and Huberman, 2005] discovered that the collective tags for resources which are applied by lots of users exhibit stable patterns over time - that is their proportions become almost fixed. They attribute their observation directly to the Pólya urn model. In its most basic form, this stochastic model consists of an urn which only contains two balls of different colors (e.g. say a blue and a red one) at the beginning. In each step of the experiment, one ball is randomly drawn and then put back along with one additional ball having the same color. After a number of draws a specific pattern starts to emerge as the fraction of balls having a given color starts to stabilize over time. Though, this fraction of balls converges to random limits, implying that for each separate run of the experiment the outcome is expected to be different.

- An improvement to the basic Pólya urn model explained above is the so-called Yule-Simon model (cf. [Yule, 1925] and [Simon, 1955]) which allows new tags to be added to the tagging vocabulary over time. According to [Cattuto et al., 2006] the model in its original version can be described as the process of generating text from scratch as follows

> *At each discrete time step one word is appended to the text: with probability p the appended word is a new word, never occurred before, while with probability* $1-p$ *one word is copied from the existing text, choosing it with a probability proportional to its current frequency of occurrence. This simple process yields frequency-rank distributions that display a power-law tail with exponent* $\alpha = 1 - p$*, lower than the exponents we observe in actual data. This happens because the Yule-Simon process has no notion of "aging", i.e. all positions within the text are regarded as identical.*

Clearly, this model follows the rules of a behavior that is also better known as preferential attachment or "the rich get richer" respectively.

- [Halpin et al., 2007] propose a generative tagging model that does not only capture preferential attachment but incorporates tag selection based on a tag's information value, too. They specify the information value (IV) of a tag x as the probability $P(I(x))$. The IV of a hypothetical tag equals 0 in two extreme cases: a) for a tag which is used during search that would retrieve all resources or b) for a tag that has never been applied and thus would retrieve no resources during search at all. On the other hand, the IV of a tag is 1 if it would exclusively select the resource in question. The authors estimate $P(I(x))$ empirically by retrieving the total number of resources from the Delicious website and converting it to a probability accordingly. Apart from the IV, preferential attachment should ensure to produce a power-law distribution. There is a baseline probability $P(a)$ that represents the likelihood of a user to add a tag to a resource. $P(o)$ is the constant probability that a user reinforces an old (i.e. previously used) tag. If an old tag is added, this happens with a probability of $P(\frac{R(x)}{\sum R(i)})$, where $R(x)$ is the number of tag selections for a particular tag x in the past and $\sum R(i)$ is the total number of all previous tag applications so far. The final generative probability of a tag x is given by the linear interpolation of the preferential attachment and the IV using λ as its weighting factor: $P(x) = \lambda * P(I(x)) + (1 - \lambda) * P(a) * P(o) * P(\frac{R(x)}{\sum R(i)})$

2.2.5 Work in our group KMI - Graz University of Technology

- In an ACM graduate student research challenge [Körner, 2009] suggests that it seems feasible to automatically detect different motivations behind tagging. He presents statistical measures aiming to identify between two fundamentally different types of tagging (cf. [Coates, 2005]): a) so-called categorizers who primarily annotate resources to categorize them for later browsing activities and b) so-called describers who tag resources in a verbose, descriptive manner in order to support their goal of later searching activities. The author proposes to identify describers by using a measure

called "orphaned tags" (and others) which relates a user's infrequently used tags to the total number of tags found in the tagging history. Categorizers can be detected using a measure called "tag entropy" that borrows from information theory. This measure tries to quantify the efficiency of tagging which can be viewed as an encoding process of resources using annotations.

- [Strohmaier et al., 2010a] later elaborated on the preliminary findings of [Körner, 2009]. In particular, they proceed the work to investigate three essential questions: 1) In what ways is user motivation amenable to quantitative analysis? 2.) Does users' motivation for tagging vary within and across social tagging systems? and 3) How does the variability in user motivation influence resulting tags and folksonomies? For that purpose, the authors suggested a measure that combines two statistical measures proposed so far and applied it to several real-world tagging datasets. Most importantly, they discovered that tagging motivation varies not only across but within tagging systems. It was further shown that there is a significant difference between the tag agreement levels of categorizers and describers. The key findings of this work are:
  - not all tags are equally useful for tasks such as information retrieval due to different levels of descriptiveness and agreement
  - it is possible to influence tagging behavior by proper system design which should enable operators of tagging systems to guide their users (at least to some extent)
  - tag recommendation might benefit from augmenting current techniques with the knowledge about users' motivation and thereby providing better support for the tagging activities of individuals
- [Körner et al., 2010b] continued the line of research which was started by [Körner, 2009] and [Strohmaier et al., 2010a] by systematically defining and evaluating the usefulness of different measures aiming at the distinction of categorizers and describers in social tagging systems. Previous work has been largely based on the investigation of different intuitions. In their line of work though, the authors especially focused on both, a qualitative evaluation (backed by a human subject study) as well as a quantitative evaluation (based on analytical experiments). Five measures are introduced aiming to quantify various forms of users' behavior in order to derive knowledge about their underlying motivations. The results of the qualitative evaluation show that a very basic measure (i.e. the ratio of a user's number of tags to the number of resources) appears to best capture human judgement. The key finding of the quantitative evaluation, using simple recommendation techniques to simulate latent system-dependent influence, reveals that tagging motivation noticeably affects users' tagging behavior. All in all, the main contributions of this body of research are:

  - a distinction between categorizers and describers based on a number of intuitions about their corresponding tagging behavior
  - five statistical measures for the automatic detection of categorizers and describers in social tagging systems
  - a qualitative and quantitative evaluation of all investigated measures
  - an interpretation of evaluation results suggesting what measures are indicative of which kind of user motivation

- [Kern et al., 2010] performed a quantitative study to investigate whether different types of tagging behavior - description vs. categorization - influence the quality / performance of basic tag recommendation techniques. They implemented two simple tag recommender systems, namely a personomy-based and a folksonomy-based one. While the first exclusively selects tag suggestions originating from the personal tagging history of a user the second suggests the most frequently used tags from other users belonging to the group of describers. The user separation itself has been performed according to statistical measures as already proposed in earlier research work (cf. [Körner, 2009] or [Strohmaier et al., 2010a]). The baseline for the evaluation of the resulting performance of the two recommender systems was set by randomly formed user groups. All calculations have been done for different splits of the user base (i.e. splitting users into 10% categorizers and 90% describers up to 90% categorizers vs. 10% describers) on a dataset sampled from Delicious. The authors could show that the personomy-based recommender particularly benefits the tagging activities of categorizers while describers primarily tend to tag similarly to other like-minded taggers within the folksonomy. Additionally, they identified the threshold where it would be wise for tag recommenders used in production to switch over from a personomy-based recommendation approach to a folksonomy-based recommender system. This threshold can be found at the intersection of the two recommenders' relative improvements over the random baseline.

- Together with colleagues from the Knowledge and Data Engineering Group at the University of Kassel, the work of [Körner et al., 2010a] addressed the hypothesis whether the quality of emergent semantics within social tagging systems is dependent on the tagging pragmatics of users. The authors discuss four different measures which are capturing only usage patterns of users' tagging activities and are thus totally independent of tagging of semantics. This allows for the exploration of a potentially existing link between tagging pragmatics and semantics. Starting out with groups of extreme categorizers and describers as detected by the respective measures, their strategy is to analyze the suitability of each of the pragmatic measures to assemble a subset of users which provides a sufficient context to harvest emergent tag semantics. As more and more users are gradually added to the

corresponding groups, it is assessed at each step how well the intermediate folksonomy (as defined by the current user subset) serves as a basis to compute semantically related tags. Semantical relatedness is determined using a measure called "tag context similarity"(cf. [Jiang and Conrath, 1997]) which operates in a vector space spanned by tags and is based on cosine similarity. This measure has shown to produce valid results during an analysis performed on WordNet[1] [Miller, 1995] data and further supports the assumption to yield more closely related tags when better implicit semantic structures are present in the investigated folksonomy. In order to validate their hypothesis, several experiments have been carried out on a very large dataset [2] crawled from the Delicious bookmarking system during November 2006.
A recap of their key findings comprises among others:

- the verbose tagging style of describers obviously provides a better basis for harvesting meaningful tag semantics in general
- the most verbose taggers are likely to be spammers who tend to negatively affect the overall semantic accuracy
- the presented pragmatic measures can be used to identify a relatively small subset of users which may induce an equal or even better quality of semantic relations than that of the whole population
- the same measures could also be utilized to detect and filter users that would otherwise cause "semantic noise" concerning the global semantic precision

2.3 Relevant theory

The following sections aim to briefly explain basic concepts of graph, network and information theory which are relevant in the context of this thesis. R-partite graphs - in particular bipartite and tripartite ones - play an important role since they provide the theoretic means needed to formalize the personomy as well as the folksonomy model which have been introduced in (1.3). The resulting graph structures from users' tagging activities represent large real-world networks which are usually investigated and discussed by means of network theoretic concepts. Finally, it is important to understand some fundamental concepts of information theory because parts of the statistical analysis in this work incorporate entropy-based calculations of the corresponding probability distributions.

2.3.1 Relevant concepts from graph theory

[Easley and Kleinberg, 2010] abstractly explain graphs as structures that allow to specify relationships between items. For that purpose, a graph consists of a

[1] `http://wordnet.princeton.edu`
[2] available at `http://www.kde.cs.uni-kassel.de/benz/papers/2010/www.html`

set of objects and a set of relations between pairs of these objects. Two certain objects are said to be neighbors if there is a relation between them. Relations itself can either be symmetric or asymmetric which causes the resulting graphs to be undirected and directed respectively.
A formal definition of graphs can be found, for instance, in [Diestel, 2005] who defines a graph as a pair $G = (V, E)$ of sets such that $E \subseteq [V]^2$ and E contains 2-element subsets of V. A graph's vertex set is given by $V(G)$ whereas $E(G)$ represents its set of edges. The vertex count of a graph G is its order $|G|$ while the number of edges is denoted by $||G||$. A vertex v is incident with an edge e if $v \in e$ (e is then an edge at v). Two vertices x, y of G are adjacent, or neighbors, if xy is an edge of G. Two edges $e \neq f$ are adjacent if they have an end in common (i.e. starting or ending at the same vertex). The set of neighbors of a vertex v is denoted by $N_G(v)$. The degree $d_G(v)$ of a vertex v is the number of edges $|E(v)|$ at v which is equal to the number of the neighbors of v in this definition. A vertex having degree $d_G(v) = 0$ is isolated (a.k.a an orphan). [Diestel, 2005] further defines the r-partite ($r >= 2$) graph whose set of vertices V admits of a partition into r distinct classes such that every edge has its ends in different classes. In other words, vertices from within the same partition class r must not be adjacent.

2.3.2 Relevant concepts from network theory

As opposed to graph theory, which primarily deals with theoretical aspects, research in the field of network theory is focused on the analysis of large-scale networks occurring in the real world. [Easley and Kleinberg, 2010] present a non-exhaustive list containing five of the main categories of large-scale network data that have frequently appeared in recent research:

- **Collaboration graphs:** record who works with whom in a specific setting (co-authorships by researchers, co-appearances by actors etc.)
- **Who-talks-to-whom graphs:** capture communication structures within large communities (e.g. the Microsoft instant messaging graph)
- **Information linkage graphs:** hold data of less specific information networks whose data often stand out both in scale and diversity (central examples are snapshots of the world wide web)
- **Technological networks:** are oriented towards the representation of physical devices and connections (e.g. the interconnections of computers or a power grid network)
- **Networks in the natural world:** represent graph structures in natural sciences such as biology (e.g. who-eats-whom relationships among species, neural connections within an organism's brain, networks making up a cell's metabolism)

In general, these categories are not to be meant exclusively distinct from one another. Since, for instance, the datasets that are used during the course of this thesis contain information originating from different social tagging systems, there is no clear affiliation to any single category. Instead there are several ways to reasonably relate them not only to different, but even to more than one category simultaneously. Thus, it is one viable way to categorize tagging datasets as *collaboration graphs* (people collaboratively generate meta-data), *information linkage graphs* (the web contains massive amounts of tagging data that is not focused towards specific resources) as well as *who-talks-to-whom graphs* (tagging can be regarded as means to (in)directly communicate with others by expressing one's opinions or emotions within a community).
A very important aspect of network theory that is relevant in the scope of this thesis is related to the general notion of popularity. [Easley and Kleinberg, 2010] provide a very intuitive explanation by discussing popularity as a phenomenon which is characterized by extreme imbalances:

> *"... while almost everyone goes through life known only to people in their immediate social circles, a few people achieve wider visibility, and a very, very few attain global name recognition. The same could be said of books, movies, or almost anything that commands an audience."*

Searching for reasons why and how such imbalances arise, people have done several experiments to measure the distribution of web links - all which basically led to very similar findings. The fraction of web pages having a number of k incoming links is roughly proportional to $1/k^c$ with c being approximately 2 (cf. [Broder et al., 2000]). There are also several examples from other domains exhibiting similar behavior which are mentioned, for example, in [Newman, 2003] or [Albert and Barabasi, 2002]. Generally, distributions of the form $f(k) = a/k^c$ (k is decreasing to some fixed power c and a is a proportionality constant) are called power-law distributions. [Easley and Kleinberg, 2010] also point out a well-known method to quickly check, whether a dataset follows a power-law behavior. Taking the logarithm of both sides of the equation yields $log(f(k)) = log(a) - c * log(k)$. Plotting this log-log relationship should then show a very close approximation of a straight line, given that the investigated data indeed exhibits a power-low distribution. The slope of the line is given by $-c$ while $log(a)$ defines its y-axis intercept. Just as the famous central limit theorem is used to explain the normal distribution which is frequently found in natural sciences, the often recurring power-law distribution is rooted in an underlying model of preferential attachment (a.k.a. "rich-get-richer").

2.3.3 Relevant concepts from information theory

This section aims to quickly introduce the reader to closely related, yet fundamental concepts of information theory - namely entropy, joint entropy, conditional

entropy, relative entropy and mutual information. The following basic definitions are taken from [Cover and Thomas, 2006].

Entropy

Entropy represents a quantity to measure the uncertainty of a random variable. In other words it is the number of bits needed on average to describe a random variable. Let X be a discrete random variable and $p(x)$ the corresponding probability mass function then the entropy $H(X)$ is defined as

$$H(X) = -\sum_{x} p(x) log p(x) \tag{2.1}$$

In this thesis, the logarithm is always to the base 2 which expresses the resulting entropy in bits. Since entropy is a function of the distribution on the random variable X, it does only depend on the probabilities of its different states rather than the actual values taken by X. The entropy $H(X)$ can also be interpreted as the expected value of $log(\frac{1}{p(X)})$ where X is drawn according to the probability mass function $p(x)$ and E denotes expectation. Thus

$$H(X) = E_p log \frac{1}{p(X)} \tag{2.2}$$

Joint entropy and conditional entropy

A natural extension of the previous definition to a pair of random variables directly leads to the definitions of joint entropy and conditional entropy. Let (X, Y) be a pair of discrete random variables and $p(x, y)$ be the corresponding joint probability mass function then the joint entropy $H(X, Y)$ is defined as

$$H(X, Y) = -\sum_{x} \sum_{y} p(x, y) log p(x, y) \tag{2.3}$$

Conditional entropy $H(Y|X)$ is defined as the entropy of a random variable, given another random variable. Let (X, Y) be a pair of discrete random variables, $p(x, y)$ the corresponding joint probability mass function and $p(y|x)$ the conditional probability of y given x then the conditional entropy $H(Y|X)$ is defined as

$$H(Y|X) = -\sum_{x} \sum_{y} p(x, y) log p(y|x) \tag{2.4}$$

Relative entropy (a.k.a. Kullback-Leibler divergence)

Given two probability distributions p and q the relative entropy quantifies how closely they are related to one another. In other words it expresses the inefficiency of assuming a certain probability distribution to be given by q when its true

distribution is given by p. Let $p(x)$ and $q(x)$ be two probability mass functions then the relative entropy or Kullback-Leibler distance $D(p||q)$ between them is defined as

$$D(p||q) = \sum_x p(x) log \frac{p(x)}{q(x)} \tag{2.5}$$

The following are four important properties of relative entropy. $D(p||q)$

1. is always non-negative
2. equals 0 iff both distributions match exactly
3. is not symmetric in general
4. does not satisfy the triangle inequality

It is further important to note, that relative entropy is not a true distance metric between probability distributions which is due to properties 3 and 4. Though, it is often helpful to think of relative entropy as a "distance".

Mutual information

Mutual information measures the amount of information that one random variable contains about another random variable. In other words it represents the reduction of uncertainty of one random variable due to the knowledge of another one. Let X and Y be two random variables having a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$ then the mutual information $I(X; Y)$ is defined as

$$I(X; Y) = \sum_x \sum_y p(x, y) log \frac{p(x, y)}{p(x)p(y)} \tag{2.6}$$

Obviously, this can be considered a special case of the relative entropy, which measures the dependence of two random variables as the relative entropy between their joint distribution $p(x, y)$ and their product distribution $p(x)p(y)$.

Chapter 3

Experimental Setup

This chapter explains the approach that has been taken to tackle the underlying research questions in the context of this thesis. It starts with a comparison of the categorization versus the description approach to tagging. Measures that are capable of detecting different types of tagging motivations are presented next. Additionally, this chapter provides an in-depth look to the datasets which have been used for the statistical analysis. Namely, ten different tagging datasets acquired from six varying social tagging systems are presented together with basic aspects of the crawling strategies, some high-level aspects of the data acquisition process as well as any (self-)imposed or system-dependent constraints and restrictions. Finally the chapter concludes with some privacy considerations about online user-generated data in general.

3.1 Approach

3.1.1 Categorization vs. description

From now on, and in relation to the thoughts mentioned in [Coates, 2005], this thesis will concentrate on two radically different types of tagging motivation, namely the categorization vs. the description approach to tagging. Users of social tagging systems that follow one of these approaches will be referred to as categorizers or describers respectively, whose tagging activities are characterized next (cf. the preliminary results within the same context of [Körner, 2009] who performed first experiments on the automatic detection of tagging motivation which have been later elaborated on and refined by [Strohmaier et al., 2010a]).

- **Categorizers:** try to organize their resources similarly to classical filing of documents. They utilize tags to build and maintain a resource repository which is structured in a way that allows for easy navigation (browsing) of contained resources later on. For this purpose, they establish their own personal vocabulary which typically tends to stabilize relatively quickly. Further, they try to avoid both, semantically and syntactically similar tags.

What results from a tagging structural point of view can be regarded as a replacement to a taxonomy - some form of shared or personal high-level ontology to navigate resources.

- **Describers:** have the predominant goal of describing their resources in a painstakingly detailed and precise manner. Doing so, requires them not to control or restrict their tagging vocabulary in order to be able to always use new (i.e. the most accurate) words for their tagging activities. The annotations of describers intentionally aim at the description of the content of resources which is the reason why they concentrate on tags that presumably facilitate later search and retrieval. Additionally and in contrast to categorizers, the tag vocabulary of describers contains lots of rarely used tags as well as a number of synonyms.

The presented distinction is based on tagging pragmatics rather than its semantics. Apart from that, the characterization is ideal and theoretic by nature as it explains the opposite ends of a spectrum. Chances are high that the motivation of real-world taggers is to be found somewhere in between. Users, though primarily describing their resources, could be simultaneously concerned with the maintenance of a few well-established categories, too.

Table 3.1 borrowed from [Körner et al., 2010b] contrasts categorizers with describers based on a number of intuitions about these two tagging approaches.

	Categorizer	**Describer**
Goal	later browsing	later retrieval
Change of vocabulary	costly	cheap
Size of vocabulary	limited	open
Tags	subjective	objective
Tag reuse	frequent	rare
Tag purpose	mimicking taxonomy	descriptive labels

Table 3.1: Intuitions about Categorizers and Describers

3.1.2 Measures to detect different types of tagging motivation

To automatically detect whether users show a stronger tendency towards being categorizers or describers, measures are needed that concentrate on the users' tagging practices by operating on "low-level features" of the underlying tagging structure. This section introduces several measures that are basically capable to perform such distinction to a certain degree. To begin with, a few simple measures are discussed and related to the categorization and description approach accordingly. These measures and/or variants thereof have already been partly addressed within different contexts and scopes by e.g. [Körner et al., 2010a] and [Körner et al., 2010b].

- **Size of tagging vocabulary:** The tag vocabulary size $|T_u|$ (see equ. 3.1) equals the unique number of tags that can be found within a user's personomy. As listed in table 3.1, describers are expected to exhibit a virtually

unlimited tag vocabulary while categorizers try to create an elaborated tag set, limited in size to best serve their individual needs.

$$Tvoc_u = |T_u| \tag{3.1}$$

- **Tags-to-resources ratio:** The tags-to-resources ratio Trr_u (see equ. 3.2) relates the tag vocabulary size $|T_u|$ to the total number of resources $|R_u|$ within a user's personomy. Compared to categorizers, describers are likely to score higher values due to their characteristic to apply many distinct tags to their resources, which is not the case for taggers that primarily employ tags for categorization purposes.

$$Trr_u = \frac{|T_u|}{|R_u|} \tag{3.2}$$

- **Mean tags per resource:** The $Tmean_u$ measure (see equ. 3.3) detects the average number of tags that users apply to resources, thereby reflecting on different levels of tagging verbosity. Since categorizers are interested to keep their resources efficiently browsable they can be expected to keep this value relatively low. Conversely, describers would probably apply all reasonable tags that come to their mind in respect of later search and retrieval activities. Thus, they are supposed to score higher values.

$$Tmean_u = \frac{\sum_{r \in R_u} |T_{ur}|}{|R_u|} \tag{3.3}$$

- **Orphans-to-tags ratio:** The orphans-to-tags ratio Otr_u (see equ. 3.4) captures the number of tag orphans $|T_{u_{freq=1}}|$ produced by their respective users and relates it to the size of the tagging vocabulary $|T_u|$. So-called tag orphans are tags which have been used only once throughout a user's tagging history. This measure is an indicator whether taggers are mainly reusing already existing tags or introducing lots of new tags. A high number of tag orphans would likely be to the disadvantage of a categorizer, while the describer's approach is generally not penalized and instead may even benefit thereof.

$$Otr_u = \frac{|T_{u_{freq=1}}|}{|T_u|} \tag{3.4}$$

Table 3.2 contains some remarks on potential drawbacks and limitations that may come along with the naive measures introduced so far.

Measure	Drawback / Limitation
size of tagging vocabulary	unable to consider the evolutionary process of tagging - most users typically start with few tags
tags-to-resources ratio	mutually dependent to some extent - the resource and tag count both are gradually increasing up to some point
mean tags per resource	unable to distinguish different verbosity levels of tagging - users tend to assign either fewer or more tags which is not necessarily related to their main underlying motivation
orphans-to-tags ratio	very strict measure susceptible to typos or different spellings - possibly also biased by further syntactic aspects

Table 3.2: Drawbacks and limitations of naive measure to detect tagging motivation

The development of more sophisticated measures might allow to overcome some of the aforementioned limitations. With this in mind, [Körner, 2009] and [Strohmaier et al., 2010a] proposed two measures for the detection and separation of categorizers and describers which are introduced next:

- **Tag orphaniness:** The M_{desc} measure (see equ. 3.5) is an elaborated version of its simple counterpart, namely, the orphans-to-tags ratio Otr_u. Instead of using the strict definition of a tag orphan $|T_{u_{freq=1}}|$, it defines a fuzzy variant which is not directly susceptible to issues such as typos or different spellings. Describers can be expected to have a high fraction of so-called orphaned tags (i.e. infrequently used tags) within their tagging vocabulary whereas the effectiveness of the resource organization of categorizers would suffer from lots of orphaned tags. The minimum frequency n, up to which tags are considered orphaned tags, is individually determined for every user by $\lceil \frac{|R(t_{max})|}{100} \rceil$, which always causes the first-percentile of the tag histogram to be cut-off.

$$M_{desc} = \frac{|\{t : |R(t)| \leq n\}|}{|T|}, n = \lceil \frac{|R(t_{max})|}{100} \rceil \tag{3.5}$$

- **Normalized conditional tag entropy:** Based on concepts from information theory, the M_{cat} measure (see equ. 3.6) tries to quantify the tagging behavior of users by thinking of tagging as an encoding process. The conditional entropy of resources given tags $H(R|T)$ allows to assess the effectiveness of a user's encoding quality when tagging resources. For normalization purposes, it is necessary to calculate the conditional entropy of an optimal categorizer $H_{opt}(R|T)$, which is determined by considering a user's individual $Tmean_u$ (see equ. 3.3) value and adapting the joint probability $p(r,t)$ accordingly. Extreme categorizers, behaving almost the same like their perfect counterparts, would reach values around 0, whereas extreme describers may even score values above 1, since this measure has not been strictly

normalized to the interval $[0, 1]$.

$$H(R|T) = -\sum_{r \in R} \sum_{t \in T} p(r,t) log_2(p(r|t))$$
$$M_{cat} = \frac{H(R|T) - H_{opt}(R|T)}{H_{opt}(R|T)} \tag{3.6}$$

- **Final combination:** The resulting M_{comb} measure (see equ. 3.7) incorporates both of the aforementioned aspects simultaneously and is simply defined as the arithmetic mean of the normalized conditional entropy M_{cat} and the tag orphaniness M_{desc}.

$$M_{comb} = \frac{M_{cat} + M_{desc}}{2} \tag{3.7}$$

Both measures were developed independently from one another and are based on different intuitions referring to the characterization from table 3.1. As pointed out in [Strohmaier et al., 2010a], the proposed measures have the following important properties. They are:

- *content-agnostic* (i.e. suitable across different media)
- *language-independent* (i.e. applicable to users of different language)
- *user-centric* (i.e. only information from the personomy is needed)

3.2 Description of Data

Recent studies dealing with research questions in the field of tagging show evidence that the behavior and motivation of users producing tagging data is varying a lot. Thus, it is very important to have manifold tagging data available in order to be able to address broader research questions and objectively measure results. To get a quick overview, table 3.3 shows the basic composition of the generated personomy datasets for all six tagging systems.
The presented datasets within this report, though smaller in overall scale than some other available datasets (section 3.3), aim to provide the following benefits:

- tagging data which is focused towards individual users instead of the crowd
- a detailed dataset description containing basic parameters and plots showing selected metrics as well as any important restrictions or applied constraints
- easy access to tagging data due to a common XML-based file format for every different system

System	$\mathbf{R}_{type}$	|U|	|R|	|T|	$\mathbf{R}_{min}/U$	|T|/|R|	P	$\mathbf{T}_{assigns}$
BibSonomy	bookmarks	82	91,343	28,567	500	0.3127	102,139	364,547
	publications	26	24,551	11,006	500	0.4483	26,951	88,449
CiteULike	publications	581	545,535	148,396	500	0.2720	570,130	2,207,087
Delicious	bookmarks	895	1,087,316	184,164	1,000	0.1694	1,581,174	5,149,753
Diigo	lists category	155	126,263	3,650	500	0.0289	145,248	165,800
	lists tag	138	107,931	52,531	500	0.4867	127,936	524,323
	bookmarks	131	152,944	64,826	500	0.4239	182,651	660,640
Flickr	photos	451	951,077	212,902	1,000	0.2239	951,077	5,946,222
	sets	1,419	1,966,269	49,298	500	0.0251	1,966,269	2,267,703
MovieLens	movies	99	7,078	9,983	100	1.4104	31,186	59,271

Abbreviations used in the table heading: U . . . user, R . . . resource,T . . . tag,
min . . . minimum number of elements, P . . . total posts (i.e. number of complete tagging events
consisting of one or more tag assignments), assigns . . . total number of single tag assignments to resources

Table 3.3: Basic statistics of the personomy datasets from six tagging systems acquired during the course of this thesis

Apart from those aspects above, the datasets have been generated with the following four major goals in mind:

1. **Completeness:** Tagging data should be complete for every single user contained within the corresponding dataset. This means that all of the user's publicly available tag assignments (i.e. from the very first to the most recent one) are contained. At the time of writing, there were systems where it was not possible to get all tag assignments for arbitrary users because public access was restricted to the last n user posts (e.g. Delicious n = 4000). Thus, as soon as a user exceeds this limit one would loose this user's tagging information from the very beginning, which depending on the addressed research question(s), might not be acceptable.

2. **Activity:** Tagging data should consist only of users showing a relatively high tagging activity. With activity meaning that the datasets do not contain any users that have only very few tag assignments in their history as this would not allow for significant results in statistical processing. Note however, that the applied lower bounds for tag assignments had to be individually adapted for the incorporated tagging systems. This is due to the fact that there are systems where it is very hard to identify a high number of users having lots of tag assignments each. Thus, the users' tagging activity levels differ across the six chosen tagging systems.

3. **Diversity:** Tagging data should reflect diversity across and even within tagging systems. This basically means that the individual tagging style of users may differ with regard to characteristics such as the motivation (what to achieve by tagging), the verbosity (the detail level of tagging), the vocabulary (how to achieve the goal by tagging) or the addressed resources (bookmarks, photos etc.). Having diverse tagging data allows to get immediate insights into those manifold tagging styles and to analyze certain

overall tendencies of users' tag assignments within and across tagging systems. The data diversity may also help to quantify the applicability of a specific system's research results to others and/or the general case.

4. **Chronology:** All tagging activities of users should contain time stamp information, in order to be able to maintain chronological order of the tag assignments. This allows to study evolutionary differences with regard to the tagging pragmatics.

Especially when doing tagging related research that aims to investigate user specific measures and tries to characterize individual tagging style and/or anomalies, the datasets acquired in the course of this thesis proofed to be a very valuable basis for other research activities as well. To give an example, parts of these datasets have been quite useful to Christian Körner's (currently PhD student in the research group of Ass. Prof. Dr. Markus Strohmaier) work [Körner, 2009] for the ACM Student Research Competition during the Hypertext 2009 conference. Apart from that, this research group has recently finished work [Strohmaier et al., 2010a] on a somewhat broader topic which is based on the presented datasets.

3.2.1 Criteria for selecting social tagging systems

As soon as tagging studies try to explore different tagging behavior and motivation aspects for a multitude of resources, there is a need for datasets that are both, large enough in size and manifold. The selection of appropriate social tagging systems was primarily driven by the following considerations:

- convenient API access to the system or at least a consistent HTML page layout that allows for HTML screen scraping techniques
- reasonable amount of tagging data (detailed system-specific thresholds stated in the next section) that can be publicly viewed and fetched using unauthenticated data access
- tagging data that needs to be focused towards the individual user (i.e. its personomy)
- focus on different resource types, not only tagging data of textual resources
- featuring at least several hundred users that have enough tagged resources to make statistical distinction of different user behaviour and motivation possible
- a mixture between very popular (mature) systems and new (experimental) ones that are not yet heavily used

After extensive reviews of several tagging systems considering the above mentioned aspects, six have been chosen for data acquisition purposes which are listed alphabetically in table 3.4:

System	Resources
BibSonomy	bookmarks, publications
CiteULike	publications
Delicious	bookmarks
Diigo	bookmarks, lists
Flickr	photos, sets
MovieLens	movies

Table 3.4: Selection of six social tagging systems

3.2.2 Crawling strategies, constraints and restrictions

To address the different characteristics of social tagging systems, specific strategies had to be developed for crawling the datasets. For each system relevant to the study, individual implementations had to be worked out, all of which basically performed three general steps:

1. Identifying a list of relevant users of a certain system
2. Acquiring public content of users from the corresponding system
3. Transforming the acquired data into the uniform data representation based on XML

Table 3.5 provides a quick high-level overview to get a basic idea about each step needed for the personomy generation procedure of every system:

System	User list Generation	Data Acquisition	Personomy Creation
BibSonomy	querying DB snapshot	querying DB snapshot	direct XML encoding
CiteULike	filtering provided dataset	filtering provided dataset	direct XML encoding
Delicious	HTML screen scraping	HTML screen scraping	XSL transformation
Diigo	HTML screen scraping	HTML screen scraping	XSL transformation
Flickr	HTML screen scraping	REST API calls	XSL transformation
MovieLens	filtering provided dataset	filtering provided dataset	direct XML encoding

Table 3.5: High-level overview for the three basic data aquisition steps

In the following, the individual strategies as well as any imposed constraints used for the data acquisition of each dataset are described in detail. Note that for all systems any untagged resources that may have occurred in a user's tagging history were simply ignored since such resources are not essentially conducive the tagging studies at hand.

- **BibSonomy:** For the generation of BibSonomy personomies there was neither the need for building a user list nor for crawling the data from the website. Instead, the officially provided dataset[1] from the Knowledge and Data Engineering Group at the University of Kassel (released for the ECML

[1] `http://www.kde.cs.uni-kassel.de/ws/dc09/dataset`

PKDD Discovery Challenge 2009) was used. As stated on their website, this dataset consists of an almost complete dump of "all public bookmarks and publication posts of BibSonomy until (but not including) 2009-01-01. Posts from the user dblp (a mirror of the DBLP Computer Science Bibliography) as well as all posts from users which have been flagged as spammers have been excluded."
User personomies for both, bookmark and publication data could be directly constructed by querying the database that has been reconstructed using the provided dump. For the bookmark as well as publication resource type the number of available tagged resources per user had to be between 500 and 4000 inclusive. Using these constraints 82 user personomies of tagged bookmark data as well as 26 user personomies of tagged publications could be extracted from the database and were directly encoded into the XML-based data representation.

- **CiteULike:** The CiteULike team offers complete "who-posted-what" snapshots[2] of their database on a daily basis from 2007-05-30 onwards. In this study the snapshot from 2009-08-04 has been used in order to extract CiteULike personomies. The imposed constraints were set to a minimum of 500 and a maximum of 4000 tagged articles per user. Besides that, the most frequently occurring tag in the dataset is an automatically assigned system tag named `no-tag` which is used in case a users do not specify any tags when saving a resource. After ignoring all article posts that only contained the tag `no-tag`, 581 potential users remained that additionally fulfilled the restrictions concerning the number of tagged resources. For all of them the resulting CiteULike personomies could be directly generated from the provided database snapshot.

- **Delicious:** The first step involved generating a list consisting of several thousand user names. This was done by starting with a set of the eight most popular tags[3] on delicious. At the time of writing these were the tags `design`, `blog`, `video`, `software`, `tools`, `music`, `programming`, `webdesign` (in descending ordered by number of occurrence). For each of these tags the 100 most recent bookmarks have been gathered and all user names that have saved any of those bookmarks have been collected. This eventually led to a large list of unique usernames. Based on that list the tagging data of every specific user that fulfilled the imposed constraints (to have at least 1000 but at most 4000 tagged bookmarks) was gathered by scraping all the HTML pages and merging them into one big XHTML conforming page that represents a user's complete bookmark history. Finally, all XHTML pages for every user have been appropriately transformed by XSLT which resulted in 895 Delicious personomies.

[2] `http://www.citeulike.org/faq/data.adp`
[3] `http://delicious.com/tag?sort=numsaves`

Note that the upper bound of 4000 resources is due to a system limitation that did not allow to publicly view more than the 4000 most recent resources for any given user name at the time of writing. Since it was necessary for the study to build a complete and chronologically ordered personomy (i.e. fetching the bookmarks from the very first one onward) this limitation had to be respected.

- **Diigo:** For the Diigo system special interest was given to the investigation of users that are organizing their bookmarks in so called bookmark lists. Besides normal tagging functionalities, those named lists which are assigned to a predefined set of 18 different list categories (including "Others" and "Not Categorized" as unspecific containers), allow users to manage their bookmarks in an additional way which, by intuition, most probably supports the categorization behaviour.
  By exhaustive browsing through all bookmark lists[4] of every category a user list was constructed by means of HTML screen scraping. For all users that fulfilled a lower bound of at least 500 items and an upper bound of at most 4000 items within bookmark lists all publicly available data was crawled from the Diigo website. Apart from that, the "normal" bookmarking history of the same users was also crawled (again using the same constraints). Doing so allows to directly oppose the two different organization approaches of bookmarks for one and the same user.
  In the end, three different types of Diigo personomies could be created from the collected data by XSLT transformation. Firstly, 155 Diigo personomies could be generated for the lists data. Secondly, 138 Diigo personomies were extracted for the tags data of the corresponding users' resources within bookmark lists since those items are almost entirely tagged in a traditional way, too. When extracting tagged resources from within lists, we had to filter out any posts that been automatically assigned with the tag `no_tag`. Thirdly, the crawled data of users' normal bookmark histories resulted in another 131 Diigo personomies.

- **Flickr:** Doing iterative calls to the Flickr web page that shows recently uploaded photos[5], several thousand user ids have been collected by means of traditional HTML screen scraping techniques. Since there is a system limitation for users without a paid "pro account", which restricts access to the 200 most recent uploaded photos, only "pro" users could have been taken into account for further crawling activities of Flickr data. With regard to the imposed restrictions, users had to have either between 1000 and 4000 tagged photos and/or 500 to 4000 photos within sets. Using Flickr's Rest API, the complete photo streams as well as all photos within sets (i.e. albums) were saved into corresponding XML documents for all users meeting

[4] `http://www.diigo.com/list/home`
[5] `http://www.flickr.com/photos`

the requirements. In the last step, these documents were XSLT-transformed into two different types of Flickr personomies. One type representing tag related data which led to 451 Flickr personomies and the other type representing sets related data which resulted in 1419 Flickr personomies. Note that due to the possibility in Flickr that users may also tag photos of others, we filtered the tagging data to incorporate solely personal tags.

- **MovieLens:** The creation of a user list as well as crawling data from the MovieLens website was not necessary because a publicly available dataset[6] of the GroupLens Research Team from the University in Minnesota could be used. Unfortunately, this tagging dataset is relatively small since it was already released in October 2006 when the system has not yet been that heavily used. Nevertheless, it was possible to directly generate 99 MovieLens personomies from this dataset after weakening the lower bound constraint to at least 100 tagged movies per user. No upper bound has been defined for this dataset since there was no problem with overly active users that might have tagged thousands of movies.

3.2.3 Significance and relevance of the dataset samples

It is always important to consider quantitative as well as qualitative aspects of datasets that are involved in any kind of statistical processing. While the presented datasets should be large and diverse enough within the scope of this thesis, they would certainly be not significant enough to be able to draw any direct conclusions on the whole population of the particular tagging systems. Any observations and results found during this thesis are thus not necessarily valid to the same extent for the general case.
To get an impression about the qualification of the presented dataset samples, the relevance with regard to the tagging activities has been assessed for three of the bigger datasets, namely CiteULike, Delicious and Flickr Photos. For that purpose, three lists containing the 100 most popular tags within each of these datasets are generated. For each of these tags, the set of the top 10 co-occurring tags is calculated by a simple graph folding operation. In this concrete case, the folding operation allows for the reduction of bipartite tag-resource graphs into unipartite tag-tag graphs. Formally, [Wasserman and Faust, 1994] define folding by means of a matrix multiplication of the affiliation matrix (i.e. the adjacency matrix of resources and tags) with its transpose. All determined top 10 co-tag sets that resulted from the folding operation are then intersected with the corresponding related tag sets that are actually published online for the particular tagging systems. Taking the intersection is based on the reasonable assumption that the related tag sets there are calculated in the same manner. Knowing the resulting co-tag overlap values roughly expresses the comparability of the underlying tagging structures - at least for the popular parts - between the dataset

[6] `http://www.grouplens.org/node/73`

samples and the complete tagging systems they are acquired from.
Table 3.6 holds an excerpt of the total results. It lists the overlap coefficient (which equals the dice coefficient cf. [van Rijsbergen, 1979] in this particular case) of each of the co-tag and related tag sets for the 25 most popular tags.

CiteULike		**Delicious**		**Flickr Photos**	
top 25 tags	co-tag overlap coefficient	top 25 tags	co-tag overlap coefficient	top 25 tags	co-tag overlap coefficient
review	0.3	design	0.6	nature	0.4
humans	0.1	tools	0.6	japan	0.3
animals	0.1	software	0.8	france	0.1
evolution	0.3	webdesign	0.7	art	0.1
support	0.0	programming	0.6	flowers	0.4
human	0.0	blog	0.4	spain	0.2
male	0.1	web	0.6	nikon	0.1
models	0.0	web2.0	0.3	flower	0.5
research	0.0	reference	0.7	city	0.0
female	0.1	css	0.7	canon	0.1
model	0.2	video	0.5	germany	0.3
analysis	0.2	tutorial	0.8	usa	0.1
protein	0.1	music	0.7	australia	0.1
theory	0.5	javascript	0.8	garden	0.4
methods	0.2	howto	0.8	uk	0.2
statistics	0.5	free	0.6	europe	0.2
adult	0.1	art	0.5	barcelona	0.5
network	0.5	inspiration	0.7	travel	0.3
of	0.1	development	0.7	vintage	0.4
psychology	0.1	tips	0.8	macro	0.4
simulation	0.2	linux	0.8	blue	0.5
software	0.2	photography	0.7	beach	0.2
molecular	0.1	opensource	0.7	london	0.3
gene	0.2	flash	0.7	paris	0.1
data	0.1	business	0.3	car	0.0

Table 3.6: Overlap coefficients between co-tags and related tags

The average overlap coefficients for all of the 100 most popular tags are: 0.24 (for CiteULike), 0.57 (for Delicious) and 0.23 (for Flickr Photos). Despite the small size of the datasets, these values show that samples exhibit some tagging-structural similarity for Delicious and at least little tagging-structural similarity for CiteULike as well as Flickr.

3.2.4 Limitations

An important limitation of the presented datasets is the missing information with regard to any users' tag gardening [Weller and Peters, 2008] activities like "weeding" (i.e. removing tags, replacing/renaming tags) for example. Quite evidently, this problem is inherent to all datasets that are acquired from primary or secondary sources and could only be reasonably tackled by requesting this missing data directly from the corresponding system operators, if available at all. Additionally, the datasets do not contain other than public user information and

no content filters (neither to resources nor to tags) have been applied. The only exception relates to the filtering of the automatically assigned tag `no-tag` for any CiteULike and `no_tag` for any Diigo Lists posts. Since those posts have not been explicitly (i.e. manually) tagged by their corresponding users, they have been removed.

3.3 Other available datasets

While it would have been possible to directly use other publicly available datasets for the research activities during this thesis, I try to point out some drawbacks that may have been involved when doing so.
There are already several popular tagging datasets available for research, which are listed in table 3.7. All these datasets are relatively large in their overall size and aim to support research activities such as performing and analyzing collaborative filtering techniques. In particular, they are helpful when doing large-scale folksonomy analysis. Nevertheless, it is often hard to find complete, chronological data that is primarily focused towards individual users instead of the crowd. Furthermore, the available datasets mostly lack a detailed description, which makes it hard for researches to quickly assess, whether or not such datasets are adequate to their studies. More often than not, another disadvantage of available datasets is to neither explicitly mention any constraints that may have been applied on the given data, nor to point out any inherent limitations that might influence later analysis. Lastly, while trying to work on multiple tagging datasets in parallel, it might be quite laborious to deal with totally different data formats as well as encoding issues. Besides being smaller in scale, the datasets which have been crawled for this thesis should allow to overcome most of these issues.

System	**Year**	**Availability**	**Contact**
CiteULike	daily snapshots	via email & download	Richard Cameron
Bibsonomy	periodic (semi-annual)	license agreement	Andreas Hotho
MovieLens	2009	via download	GroupLens Info
GiveALink	current via API	via API	Filippo Menczer
ESP Game	2006	via download	Luis von Ahn
Delicious	2007/2008	via email	Alan Said
Delicious, Stumble Upon, Wikipedia	2008/2009	via download	Arkaitz Zubiaga
Delicious, Flickr, Last.fm, zexe.net	2006, 2007, 2008	via download	Vittorio Loreto

For links, details and references see `http://kmi.tugraz.at/staff/markus/datasets/`

Table 3.7: Non-exhaustive list of social tagging datasets available for research

3.4 Privacy issues regarding user-generated data

Whenever dealing with user-generated data, in this case real-world tagging data created by particular users on the chosen tagging systems, care must be taken not to violate privacy. During the last decades, simple anonymization of personally identifiable information (PII) (e.g. full names, social security numbers and the like) has often been considered the panacea for all privacy related issues. Apart from that, one of the fundamental difficulties in respecting the secrecy of PII is definitely related to technical advances. A concrete example is given by the Netflix movie rental service. At first glance, probably nobody would classify movie rental and rating data as PII. Though, researchers have proven different in showing that more than 80% of Netflix users could be identified only by knowing when and how they rated any three of their rented movies so far [Narayanan and Shmatikov, 2008]. Another famous example shows that having simple demographic information about the U.S. population - a combination of ZIP code, birth date and sex - might allow to uniquely identify 61% U.S. citizens (of the population in 1999) and 63% U.S. citizens (of population in 2000) respectively [Golle, 2006]. One of the most renowned examples concerning privacy violations was given by the release of the AOL search query dataset. [Barbaro and Zeller, 2006] re-identified a person based on its anonymized user id by deeper investigation of the related search session, which was found and published within this very query log. Studies like these clearly show that it needs ways to protect privacy beyond anonymizing or removing PII from datasets.
To quote Paul Ohm (Associate Professer at University of Colorado Law School) at this point

> *. . . we can derive from reidentification science two conclusions of great importance: First, the power of reidentification will create and amplify privacy harms. Reidentification combines datasets that were meant to be kept apart, and in doing so, gains power through accretion: every successful reidentification, even one that reveals seemingly nonsensitive data like movie ratings, breeds future successes. Accretive reidentification makes all of our secrets fundamentally easier to discover and reveal. Our enemies will find it easier to connect us to facts that they can use to blackmail, harass, defame, frame, or discriminate against us. Powerful reidentification will draw every one of us closer to what I call our personal "databases of ruin". Second, regulators can protect privacy in the face of easy reidentification only at great cost. Because the utility and privacy of data are intrinsically connected, no regulation can increase data privacy without also decreasing data utility. No useful database can ever be perfectly anonymous, and as the utility of data increases, the privacy decreases.* [Ohm, 2009]

Especially from a research perspective's point of view, the usefulness of data always seems to be at odds with the anonymity of data. However, the presented

datasets within this thesis do not really suffer from this dilemma since there is no data involved, that could not be publicly accessed anyway. One could argue though, that users have not explicitly permitted their tagging data to be incorporated within this study. In respect thereof and whenever possible, the study at hand tries not to directly expose any user-specific information which might immediately uncover their profiles on the respective tagging systems.

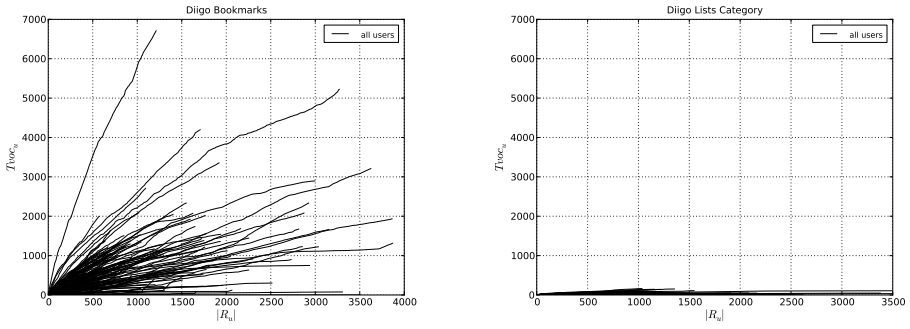
Chapter 4

Results

4.1 Basic tagging characteristics of the datasets

A short tabular overview of all ten datasets has already been provided in table 3.3 within section 3.2. To get better insights into the high-level tagging characteristics of each of the datasets, the following four aspects related to tagging characteristics of users will be presented first:

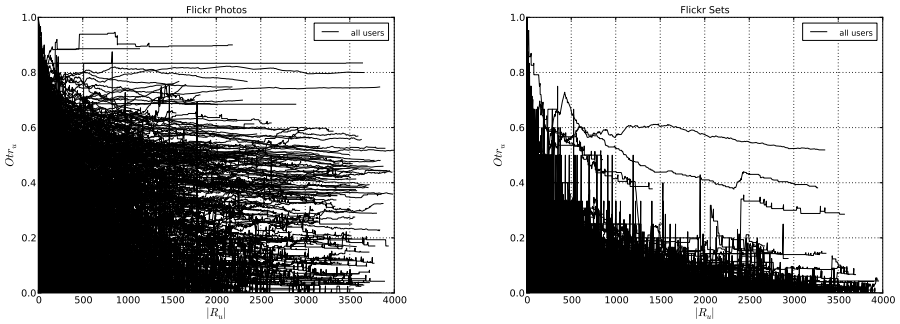
- **Growth of tagging vocabulary:** This measure simply relates the size of the resource set ($|R_u|$) to the size of the tag set ($|T_u|$) for a user's tagging history after every posting activity.



(a) Diigo Bookmarks (b) Diigo Lists Category

Figure 4.1: Growth of tagging vocabulary for selected datasets. This figure illustrates that there are significant differences in the growth of tagging vocabulary among users. Especially, it shows that in tagging datasets where a categorization approach is predominant, the tagging vocabulary stabilizes quickly and is generally much smaller. Further plots can be found in appendix B.1.1

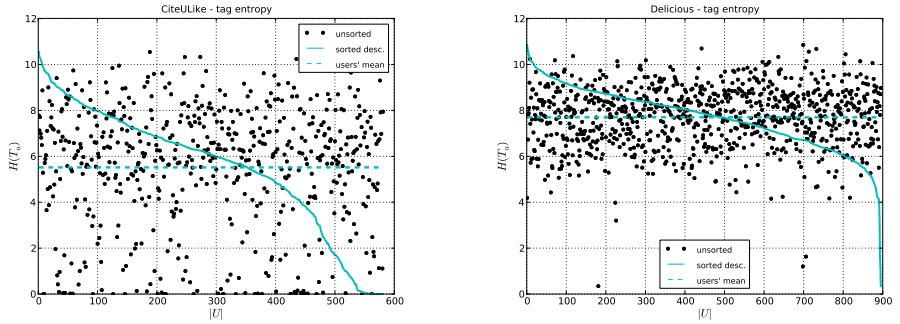
- **Evolution of tag orphan ratio:** The orphan ratio (Otr_u see equ. 3.4) is given by the number of tags having a frequency of one related to the number of all tags. Again this is done along a user's complete tagging history.



(a) Flickr Photos (b) Flickr Sets

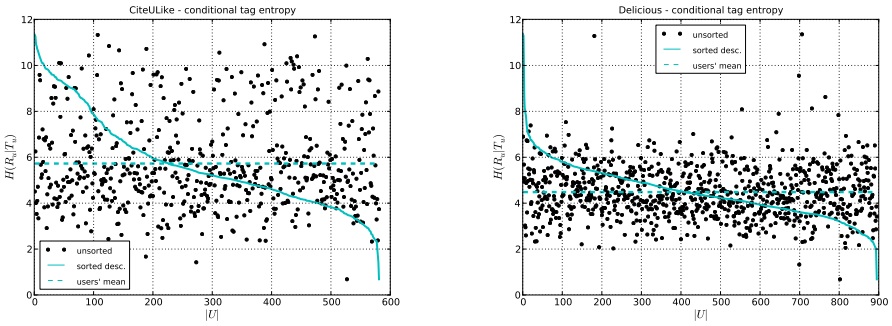
Figure 4.2: Tag orphan ratio for selected datasets. This figure contrasts the two Flickr datasets. The tag orphan ratio varies a lot, though there is a clear tendency towards low tag orphan ratios in the Flickr Sets datasets, where tags are utilized to organize photos into "virtual albums". Further results are depicted in appendix B.1.2

- **Entropy and conditional tag entropy:** According to the standard measures of information theory and as defined in section 2.3.3 the entropy as well as the conditional entropy of a user's tag histogram distribution are calculated and plotted.



(a) CiteULike (b) Delicious

Figure 4.3: Tag entropy for selected datasets (see appendix B.1.3 for the remaining datasets). These figures allow for the following quick estimations about the datasets. First, they show that the entropy of users' tag distributions is varying a lot within and across different tagging systems. Second, the figures depict noticeable differences with regard to the standard deviation σ of the entropy values. In this particular case, the population sample from Delicious exhibits smaller σ values. Third, one can quickly estimate how many of the users of a certain tagging systems achieve higher or lower entropy values than the average of the corresponding population sample. While for CiteULike roughly two thirds of the users lie above average, it is only approximately half of the users for Delicious.



(a) CiteULike (b) Delicious

Figure 4.4: Conditional tag entropy for selected datasets (see appendix B.1.3 for the remaining datasets). These figures allow for the following quick estimations about the datasets. First, they show that the conditional entropy of users' tag distributions is varying a lot within and across different tagging systems. Second, the figures depict noticeable differences with regard to the standard deviation σ of the conditional entropy values. In this particular case, the population sample from Delicious exhibits smaller σ values. Third, one can quickly estimate how many of the users of a certain tagging systems achieve higher or lower conditional entropy values than the average of the corresponding population sample. While for CiteULike only about one third of the users lie above average, it is almost half of the users for Delicious.

4.2 Statistical analysis

4.2.1 Correlations of suggested measures

Table 4.1 shows the pairwise correlations of the 7 potential measures to detect tagging motivation. Both, pearson's product moment correlation coefficient as well as spearman's rank correlation coefficient are listed for four out of ten datasets. Correlation results for the other six datasets can be found in table B.1 in appendix B.2.

4.2.2 Overlap in tagging vocabulary / likelihood of shared tags

The intuition, listed in table 3.1 that describers are using more objective and descriptive tags than categorizers gives reason to assume that they also produce a higher word overlap in their corresponding tagging vocabularies. The following analysis makes this hypothesis quite obvious. By taking groups of different sizes (i.e. the top-n% using the M_{comb} measure see equ. 3.7) of either categorizers or describers, the amount of shared tags between every possible pair of users within a certain group has been calculated over all datasets. A tag is considered a shared tag if it was used by at least one pair of users (i.e. by two different users from the same population). The results are listed in tables 4.2 and 4.3 respectively.

	CiteULike		Delicious		Flickr Photos		MovieLens	
measures	pearson	spearman	pearson	spearman	pearson	spearman	pearson	spearman
$Tvoc_u \vert Trr_u$	**0,806**	**0,945**	**0,824**	**0,904**	**0,879**	**0,955**	0,627	**0,899**
$Tvoc_u \vert Tmean_u$	0,676	**0,778**	0,660	0,695	0,379	0,593	0,640	**0,817**
$Tvoc_u \vert Otr_u$	0,324	0,396	0,600	0,725	0,576	0,748	0,317	0,678
$Tvoc_u \vert M_{desc}$	0,395	0,453	0,605	0,706	0,551	**0,790**	0,378	0,683
$Tvoc_u \vert M_{cat}$	0,345	0,559	0,610	0,735	0,647	**0,808**	0,164	0,526
$Tvoc_u \vert M_{comb}$	0,382	0,522	0,624	0,738	0,625	**0,808**	0,283	0,609
$Tvoc_u \vert P_{overlap}$	0,526	0,722	0,573	0,691	0,388	0,541	0,613	**0,787**
$Trr_u \vert Tmean_u$	**0,773**	**0,795**	0,698	0,710	0,378	0,588	0,694	**0,786**
$Trr_u \vert Otr_u$	0,350	0,409	0,671	**0,787**	0,641	**0,774**	0,608	**0,774**
$Trr_u \vert M_{desc}$	0,371	0,400	0,544	0,631	0,550	0,749	0,550	0,706
$Trr_u \vert M_{cat}$	0,312	0,538	0,612	0,737	0,668	**0,807**	0,147	0,491
$Trr_u \vert M_{comb}$	0,351	0,489	0,596	0,703	0,636	**0,791**	0,352	0,597
$Trr_u \vert P_{overlap}$	0,569	0,743	0,603	0,707	0,400	0,537	0,725	**0,754**
$Tmean_u \vert Otr_u$	0,095	0,067	0,246	0,296	0,145	0,321	0,244	0,471
$Tmean_u \vert M_{desc}$	0,231	0,152	0,419	0,491	0,258	0,431	0,278	0,461
$Tmean_u \vert M_{cat}$	0,071	0,208	0,331	0,418	0,152	0,346	0,098	0,435
$Tmean_u \vert M_{comb}$	0,135	0,187	0,383	0,464	0,207	0,386	0,200	0,459
$Tmean_u \vert P_{overlap}$	0,630	**0,988**	**0,808**	**1,000**	0,582	**0,986**	**0,860**	**1,000**
$Otr_u \vert M_{desc}$	**0,840**	**0,828**	**0,781**	0,705	**0,831**	**0,840**	**0,957**	**0,940**
$Otr_u \vert M_{cat}$	**0,752**	**0,801**	**0,863**	**0,873**	**0,859**	**0,888**	0,671	0,698
$Otr_u \vert M_{comb}$	**0,824**	**0,831**	**0,846**	**0,814**	**0,877**	**0,880**	**0,860**	**0,845**
$Otr_u \vert P_{overlap}$	-0,128	-0,058	0,304	0,293	0,247	0,279	0,365	0,400
$M_{desc} \vert M_{cat}$	**0,796**	**0,925**	**0,896**	**0,893**	**0,861**	**0,954**	0,726	**0,753**
$M_{desc} \vert M_{comb}$	**0,914**	**0,972**	**0,971**	**0,967**	**0,958**	**0,983**	**0,914**	**0,906**
$M_{desc} \vert P_{overlap}$	0,019	0,032	0,485	0,489	0,329	0,385	0,367	0,387
$M_{cat} \vert M_{comb}$	**0,973**	**0,986**	**0,976**	**0,976**	**0,971**	**0,991**	**0,943**	**0,950**
$M_{cat} \vert P_{overlap}$	-0,012	0,091	0,401	0,415	0,241	0,296	0,199	0,356
$M_{comb} \vert P_{overlap}$	-0,001	0,070	0,453	0,461	0,292	0,338	0,311	0,382

Table 4.1: Pairwise measure correlation results incorporating four selected datasets. Pearson's product moment correlation coefficient and spearman's rank correlation coefficient have been calculated over all users within the corresponding dataset. The heat-map style uses the following four correlation intervals: no correlation $|x| <= 0.25$ (white), low correlation $0.25 < |x| <= 0.50$ (yellow), medium correlation $0.50 < |x| <= 0.75$ (orange) and high correlation $|x| > 0.75$ (red). The table at hand allows to quickly assess, if and to what degree the discussed measures are correlated to one another. It also expresses whether or not correlations are consistently present and stable across all datasets. Unsurprisingly, there are very high correlations among the M_{cat}, M_{desc} and M_{comb} measures throughout all datasets. This due to the fact that, although being based on different intuitions, M_{cat} and M_{desc} basically operate on similar tag distributions and M_{comb} itself is just a linear combination of the other two. Additionally, one can see see that, despite being the simplest measure, T_{voc_u} seems to be a "versatile" measure concerning its expressiveness. This is due to the fact that it exhibits medium to high correlation with all other measure across the investigated datasets in the majority of the cases. B.1 for the other six datasets.

	n=5%	n=10%	n=15%	n=20%	n=25%	all users
BibSonomy Bookmarks	1,50	3,75	6,58	11,64	16,07	85,62
BibSonomy Publications	0,00	0,00	0,67	1,90	1,33	30,49
CiteULike	0,12	0,35	0,63	1,18	1,75	18,53
Delicious	7,99	14,12	18,99	20,14	24,27	128,75
Diigo Bookmarks	0,00	9,72	14,67	16,58	17,97	116,46
Diigo Lists Category	0,00	0,02	0,09	0,07	0,09	0,24
Diigo Lists Tag	0,00	4,19	10,32	13,27	19,77	78,14
Flickr Photos	0,13	0,41	0,63	0,93	1,25	26,31
Flickr Sets						
Movielens	0,00	0,14	0,13	0,11	0,20	6,16

Table 4.2: Top-n% categorizers - absolute mean shared tags. By taking groups of different sizes of categorizers, the absolute amount of mean shared tags has been calculated by exhaustive comparison of any two categorizers within the respective group. It is clearly visible, that the resulting values are very low when compared to the ones determined for the whole user population (i.e. the right most column). A direct comparison of the resulting values for different groups of describers can be found in the table 4.3 below.

	n=5%	n=10%	n=15%	n=20%	n=25%	all users
BibSonomy Bookmarks	53,50	124,43	130,77	176,28	191,85	85,62
BibSonomy Publications	0,00	26,00	25,67	40,40	30,47	30,49
CiteULike	39,09	28,81	28,85	33,73	40,87	18,53
Delicious	255,88	284,93	302,15	279,31	283,87	128,75
Diigo Bookmarks	249,60	230,28	260,19	221,05	224,85	116,46
Diigo Lists Category	0,24	0,17	0,38	0,32	0,40	0,24
Diigo Lists Tag	213,80	193,37	154,14	161,36	151,50	78,14
Flickr Photos	100,82	79,87	101,54	113,66	117,20	26,31
Flickr Sets						
Movielens	1,00	3,31	8,09	8,92	12,83	6,16

Table 4.3: Top-n% describers - absolute mean shared tags. By taking groups of different sizes of describers, the absolute amount of mean shared tags has been calculated by exhaustive comparison of any two describers within the respective group. It is clearly visible, that the resulting values are relatively high throughout all different datasets and user groups. In fact, they are significantly higher in the majority of the cases when compared to the ones determined for the whole user population (i.e. the right most column). A direct comparison of the resulting values for different groups of categorizers can be found in the table 4.2 above.

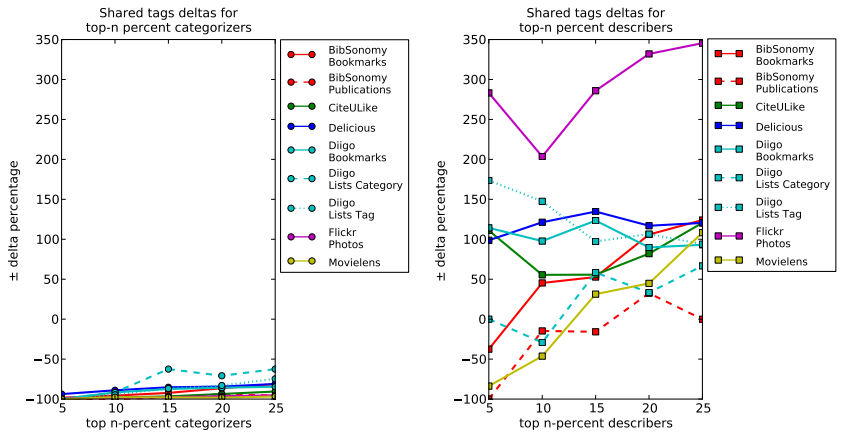


Figure 4.5: Shared tags deltas in relation to the mean of the whole datasets. This figure illustrates to what extent the number of shared tags differs between the corresponding top-n% categorizers (left plot) or describers (right plot) and the baseline, which is given by the mean value of the respective dataset's whole population (i.e. 100% of the users). One can immediately see high negative Δ values (i.e. all Δ values around -100%) for categorizers in contrast to high positive Δ values for different groups of describers. This observation lets conclude that describers tend to share a higher absolute number of tags than categorizers.

In order to account for the general differences with regard to the size of the tagging vocabularies between categorizers and describers the absolute values above have been related to the mean size of tagging vocabulary of the corresponding user groups. The outcome of the calculations can be found in tables 4.4, 4.5 and figure 4.6 below.

	n=5%	n=10%	n=15%	n=20%	n=25%
BibSonomy Bookmarks	1,31%	3,60%	5,65%	8,14%	9,68%
BibSonomy Publications	0,00%	0,00%	1,09%	2,00%	1,50%
CiteULike	1,78%	0,69%	0,92%	1,26%	1,59%
Delicious	6,08%	7,72%	8,15%	7,98%	8,34%
Diigo Bookmarks	0,00%	4,75%	4,48%	4,59%	4,11%
Diigo Lists Category	0,00%	0,26%	0,99%	0,55%	0,69%
Diigo Lists Tag	0,00%	2,30%	3,83%	3,95%	5,08%
Flickr Photos	0,46%	0,61%	0,55%	0,72%	0,88%
Flickr Sets					
Movielens	0,00%	0,38%	0,46%	0,37%	0,65%

Table 4.4: Top-n% categorizers - relative mean shared tags. Different levels of tagging verbosity have been considered by relating the absolute values from above (see table 4.2) to the mean size of the tagging vocabulary of the corresponding group of categorizers.

	n=5%	n=10%	n=15%	n=20%	n=25%
BibSonomy Bookmarks	6,64%	11,66%	12,14%	13,01%	14,27%
BibSonomy Publications	0,00%	2,64%	3,48%	7,88%	6,68%
CiteULike	4,69%	4,16%	3,67%	3,78%	4,29%
Delicious	14,09%	16,60%	17,35%	17,40%	18,07%
Diigo Bookmarks	10,88%	11,74%	11,69%	11,31%	12,41%
Diigo Lists Category	0,49%	0,34%	0,81%	0,72%	0,89%
Diigo Lists Tag	14,27%	9,81%	9,37%	10,47%	10,86%
Flickr Photos	3,83%	3,75%	4,66%	5,24%	5,64%
Flickr Sets					
Movielens	0,85%	1,59%	2,56%	2,97%	4,35%

Table 4.5: Top-n% describers - relative mean shared tags. Different levels of tagging verbosity have been considered by relating the absolute values from above (see table 4.3) to the mean size of the tagging vocabulary of the corresponding group of describers.

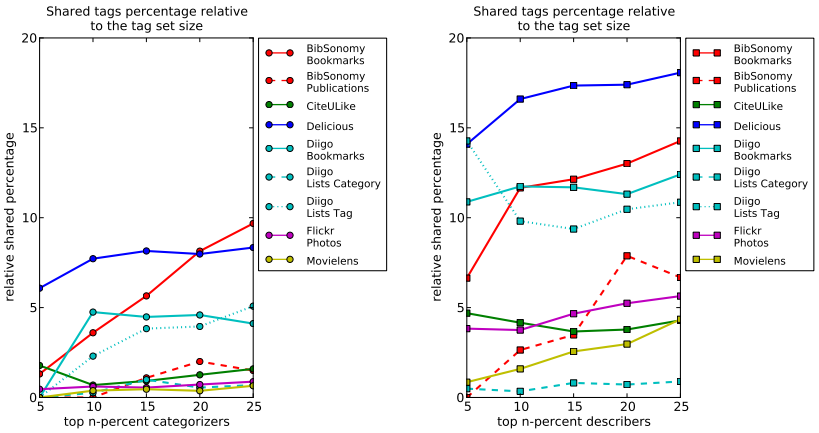


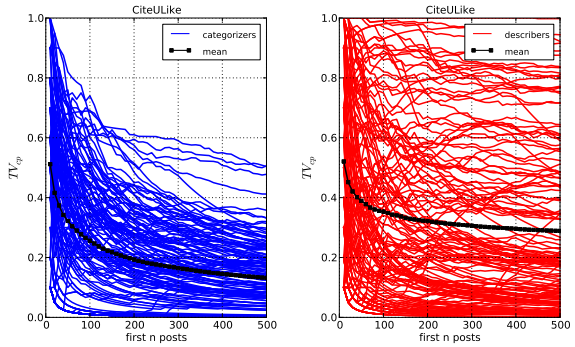
Figure 4.6: Shared tags percentage relative to the size of the tagging vocabulary. These two plots clearly show that different groups of describers always produce a higher level of shared tags. The only exception is the group of the top-15% of categorizers for the Diigo Lists Category dataset in which by nature, all users follow a very strict categorization approach and thus, share almost no tags anyway. These results should therefore remedy the objection that describers share more tags by pure likelihood.

4.2.3 Evolution of tagging vocabulary

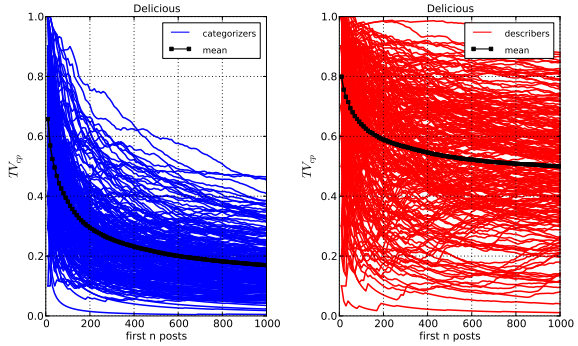
This section tries to explore the influence of tagging motivation on the evolution of a user's tagging vocabulary. This is done by analyzing a) the probability of changing the tag set and b) the mean relative change rate of the tag set when adding posts to a user's personomy. One can speculate, that describers are not only more likely to introduce new tags, but also exhibit higher relative change rates to their tagging vocabulary when doing so. This directly refers to both, the *change of vocabulary* and *tag reuse* intuitions listed in table 3.1. In the following, two measures are introduced to quantify these characteristics.

- **Tag vocabulary change probability:** For every resource's tag set, it is determined whether or not the tagging vocabulary is changed. This is done by intersecting the tags of the i-th resource R_i with the total tag set, that has been built up to this point (T_{i-1}). The tag vocabulary change probability TV_{cp_n} after n resources (see equation 4.1 is then given by the number of resource tag sets that change the vocabulary related to the total number of n resources. Figures 4.7 and 4.8 as well as those listed in appendix B.3.1 depict the TV_{cp_n} measure's evolution for the first n (= dataset's R_{min}) resources of the top 25% categorizers and describers (according to their M_{comb} measure see equ. 3.7) together with the mean of the corresponding population for all datasets.

$$\begin{aligned} T_0 &= \emptyset \\ T_n &= \bigcup_{i=1}^{n} T_{R_i} \\ TV_{cp_n} &= \frac{\sum_{i=1}^{n} T_{R_i} \cap T_{i-1} \neq \emptyset}{n} \end{aligned} \tag{4.1}$$

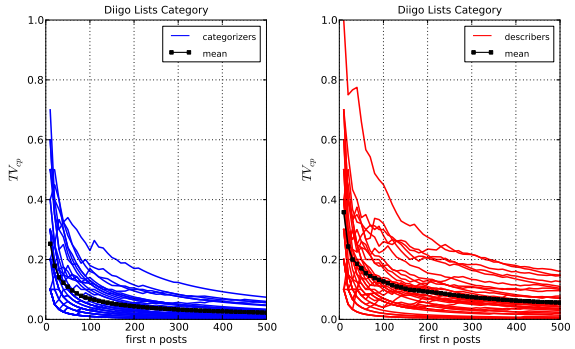


(a) Tag vocabulary change probability CiteULike dataset

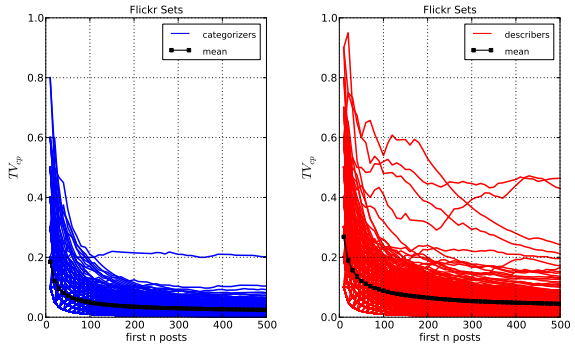


(b) Tag vocabulary change probability Delicious dataset

Figure 4.7: Looking at the TV_{cp_n} mean values only, it is immediately visible that describers are more likely to change their tagging vocabulary at any point in time. No matter what starting values, the mean TV_{cp_n} of describers is still at least twice as high as the categorizers' after n resources which can be observed for 7 out of 10 datasets: BibSonomy Bookmarks (figure B.17), CiteULike (figure 4.7(a)), Delicious (figure 4.7(b)), Diigo Boomarks (figure B.19), Diigo Lists Tag (figure B.20), Flickr Photos (figure B.21) and Movielens (figure B.22). The Bibsonomy Publications dataset is included for the sake of completeness, but cannot be considered significant to this kind of analysis due to its small size.



(a) Tag vocabulary change probability Diigo Lists Category dataset



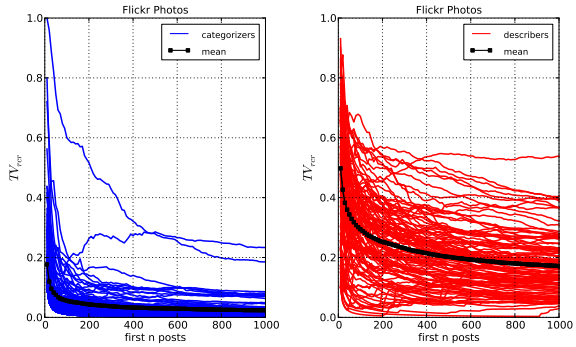
(b) Tag vocabulary change probability Flickr Sets dataset

Figure 4.8: As expected, there is not much difference between the mean tag vocabulary change probability (TV_{cp_n}) of categorizers and describers within the two datasets, that are innately dominated by users adhering to the categorization approach (Diigo Lists Category and Flickr Sets). The reason for this is that lists (Diigo) or sets (Flickr) are used for categorization purposes by nature. It would definitely mean a conceptual misuse if they were used primarily for a description approach. However, it can be observed from these two datasets, that for extreme categorizers the TV_{cp_n} values are considerably lower than for the average identified categorizers in other datasets (see figures B.17 to B.22 in appendix B.3.1)

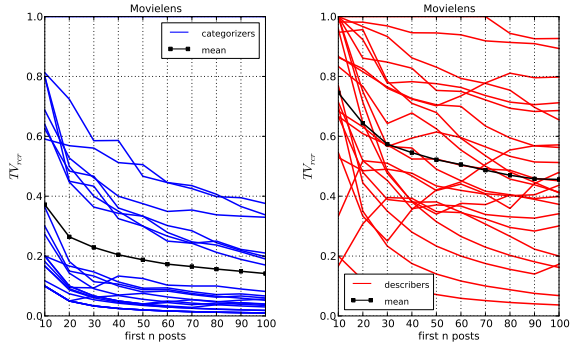
- **Tag vocabulary change rate:** First, the absolute change rate of the tagging vocabulary is determined. For all resource's tag sets T_{R_i}, the size of the intersection with the total tag set generated so far (T_{i-1}) is added and divided by the the total number of resources. For normalization purposes, the absolute change rate is related to the mean tags per resource ($mtpr_n$) at the same point in time which results in the relative change rate of the tagging vocabulary TV_{rcr_n} after n resources (see equation 4.2). The evolution of the TV_{rcr_n} measure is illustrated for all datasets in figures 4.9 and 4.10 as well as those listed in appendix B.3.2, again for the first n (=

dataset's R_{min}) resources of the top 25% categorizers and describers (according to their M_{comb} measure see equ. 3.7) together with the mean of the corresponding population. This measure addresses the relative change rate of the tagging vocabulary. In other words it quantifies how many new tags are introduced on average along the tagging history of users. Just taking the absolute change rate would definitely favor description behavior due to a higher tagging-specific verbosity in general. Thus, a normalization based on the mean tags per resource has been applied.

$$\begin{aligned} mtpr_n &= \frac{\sum_{i=1}^{n} |T_{R_i}|}{n} \\ TV_{acr_n} &= \frac{\sum_{i=1}^{n} |T_{R_i} \cap T_{i-1}|}{n} \\ TV_{rcr_n} &= \frac{TV_{acr_n}}{mtpr_n} \end{aligned} \tag{4.2}$$

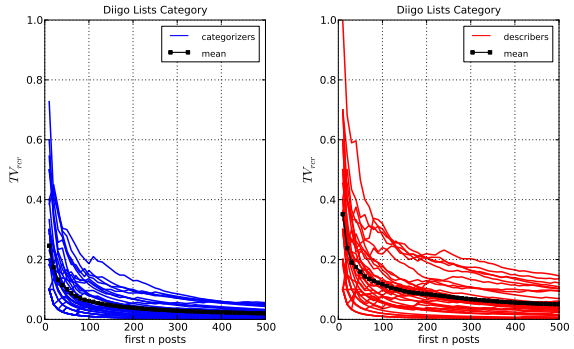


(a) Tag vocabulary change rate Flickr Photos dataset

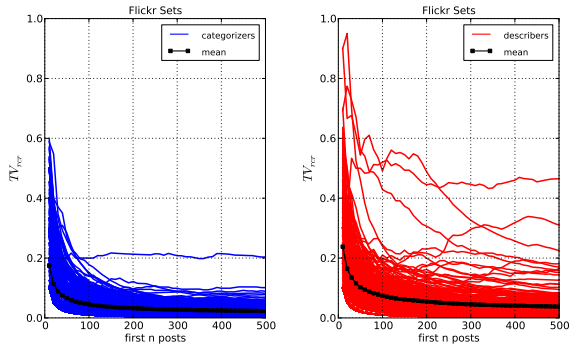


(b) Tag vocabulary change rate Movielens dataset

Figure 4.9: The outcome shows, that describers score higher mean TV_{rcr_n} values throughout all datasets. It is interesting to note though, that the evolution of this measure is very similar across the different datasets within the two user groups. For describers this means that the results are within the same tenth. To be precisely, the mean end-values after n resources lie between $0.2 <= TV_{rcr_n} <= 0.3$, for 6 out of 10 datasets: BibSonomy Bookmarks (figure B.23), BibSonomy Publications (figure B.24), CiteULike (figure B.25), Delicious (figure B.26), Diigo Boomarks (figure B.27) and Diigo Lists Tag (figure B.28). For Flickr Photos (figure 4.9(a)) the mean end-values are a little lower. This might result from an overly strong normalization, occurring when the mean tags per resource are higher than for the other datasets. It is opposite for Movielens (figure 4.9(b)) where end-values are somewhat higher. Most likely, this is due to the fact that the evolution had to be already cut-off at n=100 resources. Nevertheless, it is reasonable to assume the end-values at n=500 resources to be within the same range, like for the others. Observations akin to the ones just mentioned for describers partly hold for categorizers, too. The measure's evolution is even more consistent for the group of categorizers. For 8 out of 10 datasets mean values are $TV_{rcr_n} <= 0.1$. The only two exceptions are given by the Diigo Bookmarks (figure B.27) and Diigo Lists Tag (figure B.28) datasets respectively, whose mean values are $TV_{rcr_n} \approx 0.2$ and thus relatively high in comparison to those of describers. Still, describers score considerably higher values in both these datasets ($+\Delta \approx 0.1$ on average).



(a) Diigo Lists Category dataset



(b) Flickr Sets dataset

Figure 4.10: As with the TV_{cp_n} measure, the two datasets focused on categorization (Diigo Lists Category figure 4.10(a) and Flickr Sets figure 4.10(b)) exhibit significantly lower TV_{rcr_n} values (somewhere around 0.05), which also do not differ much between categorizers and describers within these datasets.

4.2.4 Tag cloud related properties

Users' tagging pragmatics directly affect the characteristics of tag clouds and tag histograms respectively. Thus, only by looking at the tag clouds of users that clearly behave according to one of the two fundamentally different tagging approaches, it should be possible to predict which tagging motivation they are primarily following. To assess this hypothesis, tag clouds have been generated for all personomies that originate from the absolute top or bottom 20 users according to different measures. When trying to perform a classification task for tag clouds based on naive measures (e.g. Trr_u or $Tmean_u$ see equ. 3.2 and 3.3), one quickly realizes that it might be very difficult. This is due to the fact, that even for extreme users determined by simple measures, the tag cloud characteristics may appear very similar. This is shown in figures 4.11 and 4.12 as well as those in appendix B.4.

Figure 4.11: Tag cloud of an assumed categorizer that originates from the Delicious top-20 users according to the $Tmean_u$ measure (see equ. 3.3). The tag cloud does not reflect the intuitions (see table 3.1 about the tagging behavior of categorziers. For instance, there are lots of rarely used tags as well as subjective terms. Although lower values of $Tmean_u$ seem to be necessary, they do not seem to be sufficient to correctly identify categorizers.

Figure 4.12: Tag cloud of an assumed describer that originates from the Delicious top-20 users according to the $Tmean_u$ measure (see equ. 3.3). The tag cloud shows intuitions (see table 3.1) about the tagging practices of describers. This might be an indication, that the $Tmean_u$ measure could be better suited for the identification of describers rather than categorizers. At least, high values seem to be a good indicator towards the identification of a describer's motivation.

When using more elaborate measures like M_{comb} (see equ. 3.7) for such a classification task, one will find that it becomes easier to do so. Tag clouds resulting from personomies out of the absolute top 20 categorizers and describers according to elaborate measures, show clear differences in their composition. Again, this is illustrated in figures 4.13 and 4.14 as well as those in appendix B.4 showing tag clouds of potential categorizers versus potential describers out of the absolute top 20 of the selected datasets according to their M_{comb} values.

Figure 4.13: Tag cloud of a potential categorizer originating from the Flickr Photos top-20 users according to the M_{comb} measure (see equ. 3.7). Tag clouds of categorizers have a very uniform tag distribution (i.e. the frequency among tags is balanced) which is in contrast to a describer's tag cloud which is presented below in figure 4.14. This can be regarded as a potential indicator for categorizers to use their tag clouds as an aid for later navigation. Besides, it is often the case that extreme categorizers have a distinctly smaller tagging vocabulary than the average describer.

Figure 4.14: Tag cloud of a potential describer originating from the Flickr Photos top-20 users according to the M_{comb} measure (see equ. 3.7). Tag clouds of describers are showing some high-frequency tags while at the same time containing lots of low-frequency tags. Tag distributions of that kind can be expected when users are tagging in an ad-hoc manner for descriptive purposes to support later retrieval. As a result, extreme describers often have a distinctly larger tagging vocabulary than the average categorizer.

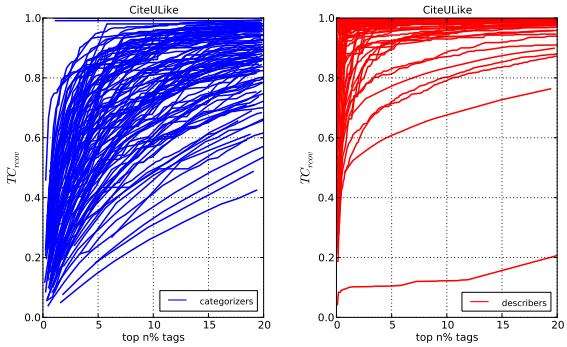
These observations give reason to the assumption that further analysis of tag cloud related properties may show interesting results. For that purpose, two measures taken from [Strohmaier et al., 2010b] are introduced, which directly allow to quantify tag cloud related characteristics that are studied next.

- **Resource coverage** (see equations 4.3 and 4.4): Informally, the absolute resource coverage t_{acov} is defined as the set of resources R_t that can be reached from a specific tag. Given a tag cloud $TC \subseteq T$ (i.e. composed of a certain subset of the complete tag set) the absolute resource coverage is the number of unique resources that can be reached from all of its tags. Relative coverage values t_{rcov} and TC_{rcov} respectively, are given by division with the total number of resources $|R|$.

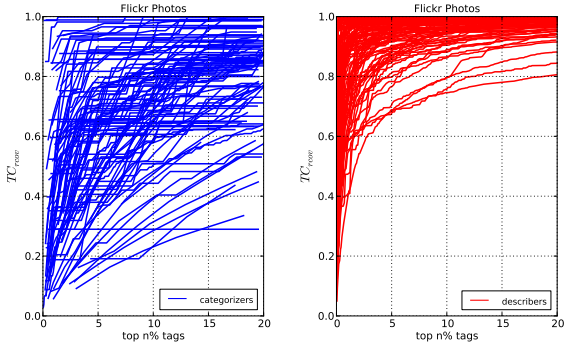
$$\begin{aligned} t_{acov} &= |R_{ut}| \\ t_{rcov} &= \frac{t_{acov}}{|R_u|} \end{aligned} \tag{4.3}$$

$$TC_{acov} = |\bigcup_{i=1}^{|TC|} R_{ut_i}|$$
$$TC_{rcov} = \frac{TC_{acov}}{|R_u|} \tag{4.4}$$

Concerning the evolution of resource coverage, one may expect noticeable differences between categorizers and describers. Figure 4.15 as well as those in appendix B.4.1 depict the evolution of the resource coverage of all datasets.



(a) CiteULike dataset



(b) Flickr Photos dataset

Figure 4.15: CiteULike and Flickr Photos resource coverage for the top-20% most frequent tags of the top-25% categorizers and describers, who were separated by the M_{comb} measure. It can be seen that the Pareto principle (i.e. more than 80% of the resources can be reached with less than 20% of the most frequent tags) clearly holds for describers of all datasets (see appendix B.4.1 for the others), an observation which is generally much less distinct for categorizers. In fact, describers already reach even more than 80% coverage at around 5% of their top-tags only, in the majority of the cases. At first sight, this might not be too surprising given the differences in the distributions of the tag histograms between categorizers and describers. It is important to note though, that tag histograms cannot directly express the evolution of resource coverage since it is not apparent at all, whether the resource sets behind the tag frequencies are highly overlapping or mainly disjoint.

- **Resource overlap:** (see equations 4.5 and 4.6): A further indication to distinguish between a categorization and a description approach may be given by the resource overlap of a personomy's tag cloud. Frequently, resource sets of different tags within a tag cloud are not disjoint. This happens as soon as users start to assign more than one tag to any single resource. The resource overlap allows to quantify this "redundancy", by determining the intersections of resource sets throughout the tags in a tag cloud

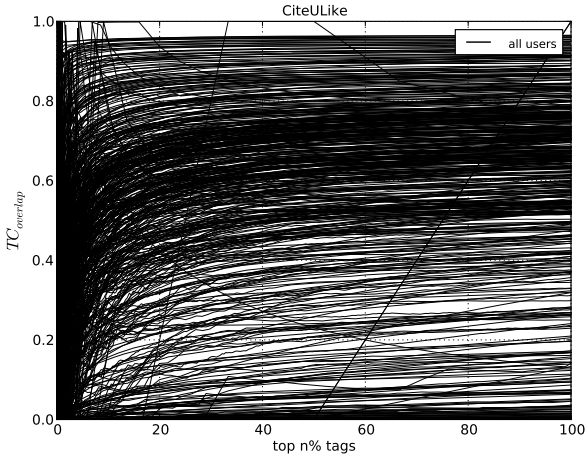
$TC \subseteq T$. Formally, the overlap for tag clouds TC containing at least two tags is defined in equation 4.5

$$TC_{overlap} = \frac{|TC| * \left(\sum_{i=1}^{|TC|} t_{i_{acov}} - TC_{acov} \right)}{(|TC| - 1) * \sum_{i=1}^{|TC|} t_{i_{acov}}} \tag{4.5}$$

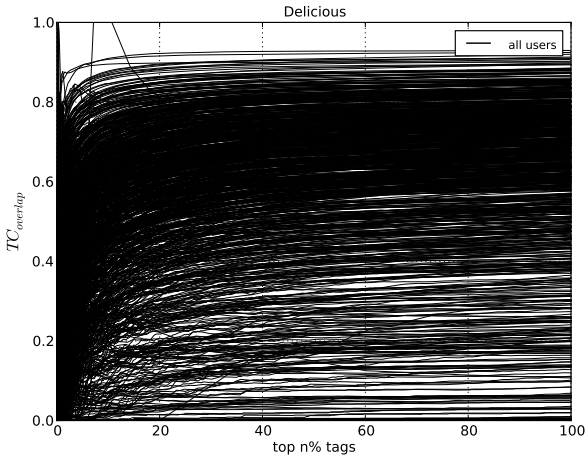
For a user's whole personomy P (i.e. $TC = T$), the overlap can be easily calculated according to equation 4.6.

$$P_{overlap} = \frac{|T_u|}{|T_u| - 1} * (1 - \frac{|R_u|}{|TAS_u|}) \tag{4.6}$$

Concerning the evolution of the resource overlap, one may also expect clear differences between categorizers and describers. Figures and as well as those listed in appendix B.4.2 depict the evolution of the resource overlap along the top n% tags.

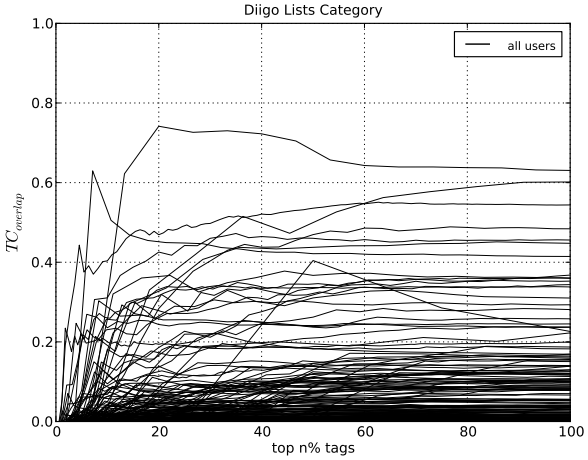


(a) Resource Overlap CiteULike dataset

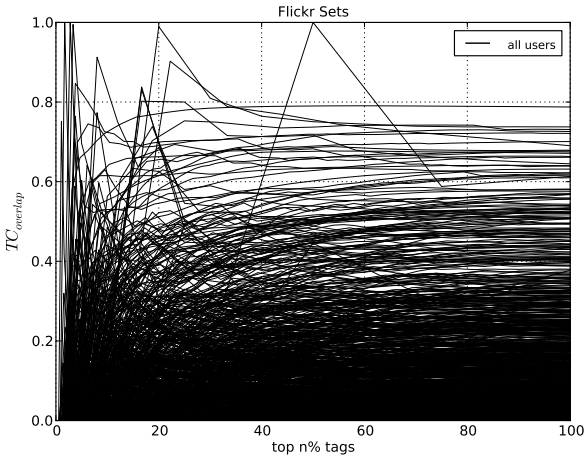


(b) Resource Overlap Delicious dataset

Figure 4.16: Complete resource overlap evolution of CiteULike and Delicious datasets. It can be seen that the $TC_{overlap}$ values are distributed rather uniformly. This is not only true for these two datasets but also, to some extent, for a number of others listed in appendix B.4.2. Generally speaking, this analysis allows to conclude that the resource overlap of users is varying a lot between and within different tagging systems. However, based on the intuition, that categorizers want to achieve little resource overlap values in order to support more efficient navigation, they should be primarily found within the lower half of the diagrams. Indeed, this hypothesis is corroborated by the following figure 4.17. Conversely, describers could even benefit of high resource overlap, which would allow them to find resources more easily by specifying multiple tags during retrieval. This is depicted, for instance, in figure B.43 in appendix B.4.2, where most users exhibit relatively high resource overlap values. They seem to prefer a descriptive tagging approach in the photos dataset, since the Flickr tagging system allows them to simultaneously maintain categories using sets.



(a) Diigo Lists Category



(b) Flickr Sets

Figure 4.17: Complete resource overlap evolution of Diigo Lists Category and Flickr Sets datasets. It is clearly visible, that for tagging systems in which users are primarily driven by categorization behavior - Diigo Lists Category (figure 4.17(a)) and Flickr Sets (figure 4.17(b)) - there is a strong tendency towards low resource overlap ($TC_{overlap} <= 0.3$). This observation corroborates the belief that it might also be feasible to separate categorizers and describers using their corresponding resource overlap. A reasonable threshold to start with might be taken out of the interval $0.3 <= TC_{overlap} <= 0.4$.

4.2.5 Differences in tag agreement

This section deals with the question whether or not tag agreement is affected by different types of tagging motivation. For that purpose, the personomies within each dataset are split into to groups of equal size. Precisely, the group of cat-

egorizers is given by the first half of the users sorted ascendingly according to their corresponding M_{comb} values, while the group of describers is given by the complementary user set (i.e. the other half of users sorted descendingly).
For each of the overall top-n resources within all datasets, two tag distributions are generated - one for the group of categorizers and one for the group of describers. This allows to determine the tag agreement T_a (see equ. 4.7) within each of the two tag distributions. As defined in [Strohmaier et al., 2010a], T_a is given by the number of tags that at least k-percent of users agree on.

$$T_a = t \in T_r, \frac{|U_t|}{|U_r|} > k \tag{4.7}$$

One can expect, that describers achieve a higher level of tag agreement than categorizers because they are focused towards an accurate description of resources, ideally by using objective tags (see intuitions about categorizers and describers in table 3.1. The tag agreement values have been calculated for the top $n = 1000$ resources for each dataset except the BibSonomy Publications dataset, where there were only 134 shared resources among the groups of categorizers and describers respectively, due to the small number of 26 personomies in total only. Note that tag agreement could not be calculated for the Flickr Photos and Flickr Sets datasets, due to the fact that all photos are represented by unique resource identifiers. Table 4.6 holds all tag agreement results for different k-percent user agreement.

k	10	20	30	40	50	60	70	80	90	mean
BibSonomy Book.										
wins categorizers	27.50	26.80	24.00	24.40	25.00	19.80	24.20	25.80	26.10	24.84
wins describers	**67.30**	**68.30**	**69.60**	**67.20**	**66.00**	**62.30**	**52.30**	**48.60**	**45.90**	**60.83**
ties	5.20	4.90	6.40	8.40	9.00	17.90	23.50	25.60	28.00	14.32
BibSonomy Pub.										
wins categorizers	32.84	32.84	32.09	30.60	30.60	22.39	22.39	22.39	22.39	27.61
wins describers	**49.25**	**49.25**	**50.00**	**51.49**	**51.49**	**61.19**	**61.19**	**61.19**	**61.19**	**55.14**
ties	17.91	17.91	17.91	17.91	17.91	16.42	16.42	16.42	16.42	17.25
CiteULike										
wins categorizers	**71.40**	**65.80**	**45.50**	24.10	21.30	18.00	12.30	11.80	10.80	31.22
wins describers	23.00	27.50	42.20	**59.30**	**62.20**	**49.80**	**52.80**	**54.60**	**55.40**	**47.42**
ties	5.60	6.70	12.30	16.60	16.50	32.20	34.90	33.60	33.80	21.36
Delicious										
wins categorizers	14.20	2.20	0.60	1.30	1.60	0.60	0.80	0.60	0.10	2.44
wins describers	**76.00**	**94.00**	**95.00**	**91.10**	**82.90**	**61.70**	**40.30**	**20.60**	**5.80**	**63.04**
ties	9.80	3.80	4.40	7.60	15.50	37.70	58.90	78.80	94.10	34.51
Diigo Book.										
wins categorizers	24.80	26.80	24.40	13.80	17.70	12.00	14.30	16.40	21.00	19.02
wins describers	**70.10**	**65.40**	**67.10**	**79.50**	**74.50**	**77.80**	**70.20**	**59.40**	**36.70**	**66.74**
ties	5.10	7.80	8.50	6.70	7.80	10.20	15.50	24.20	42.30	14.23
Diigo Lists Cat.										
wins categorizers	**53.80**	**43.20**	26.30	20.60	19.80	6.30	4.80	4.80	4.80	20.49
wins describers	35.00	40.90	**45.00**	**31.00**	**30.10**	**10.40**	**8.70**	**8.80**	**8.80**	**24.30**
ties	11.20	15.90	28.70	48.40	50.10	83.30	86.50	86.40	86.40	55.21
Diigo Lists Tag										
wins categorizers	18.00	19.20	18.10	13.60	17.00	10.30	15.00	21.40	28.00	17.84
wins describers	**76.00**	**73.10**	**74.10**	**78.60**	**74.70**	**80.60**	**71.60**	**59.60**	**41.50**	**69.98**
ties	6.00	7.70	7.80	7.80	8.30	9.10	13.40	19.00	30.50	12.18
Movielens										
wins categorizers	32.50	39.60	28.90	18.40	14.40	**6.10**	**2.50**	**2.30**	**2.00**	16.30
wins describers	**62.20**	**47.00**	**36.90**	**24.00**	**13.50**	6.00	1.70	0.80	0.20	**21.37**
ties	5.30	13.40	34.20	57.60	72.10	87.90	95.80	96.90	97.80	62.33

Table 4.6: Relative tag agreements results for all of the applicable datasets for different k-percent user agreement values, where bold values signal higher tag agreement of describers over categorizers. Obviously, regardless of the chosen k-percent value, there is a distinct tendency towards describer wins for the majority of the datasets - BibSonomy Bookmarks, BibSonomy Publications, Delicious (figure B.47), Diigo Bookmarks and Diigo Lists Tag. For the Movielens dataset there is less difference between wins of categorizers and describers. Firstly, this might result from the relatively small size of the dataset. Secondly, tagging movies can be considered a very subjective task and thirdly, it's the only dataset where users are allowed to use several words (sometimes even whole sentences) for a single tag application, which innately makes tag agreements occur much less frequently and often by sheer chance only. Besides the Diigo Lists Category dataset, where the motivation for categorization is predominant, the sole exception is given by the CiteULike dataset, in which up to k=30% of user agreement level, categorizers outperform the tag agreement levels of describers. Nevertheless, for the other percentages describers clearly score more wins than categorizes. Furthermore, the mean percentages (over all different k-percent values) of describer wins are higher for all investigated datasets. Figures B.45 to B.49 hold the equivalent graphical representations of the actual tag agreement results for every single dataset and can be found in appendix B.5.

4.2.6 Influence of resource titles

Directly related to the subjectiveness vs. objectiveness of tags (referring to table 3.1), this section examines whether or not resource titles exert influence on the tagging pragmatics of users. This question is tackled on two different levels of

abstraction. First, on macro-level, by determining to what degree the tagging vocabulary overlaps with the resource's title words within the whole personomy. Second, on micro-level, it is assessed a) how many posts contain at least one tag originating from the corresponding resource's title (normalized to total number of posts) and b) how many tag assignments exhibit a tag derived from the resource's title words (normalized to total number of tag assignments). Whenever tag-title intersections are calculated, all resource titles within a user's personomy are first tokenized to build the set of title words TW_u. Both, the tags and title words are filtered for stop-words according to the stop-word list used by the Snowball[1] stemmer.
Since this kind of analysis relies on the availability of resource title data, it is limited to seven out of the ten datasets, namely the BibSonomy Bookmarks, Bibsonomy Publications, Delicious, Diigo Bookmarks, Diigo Lists Category, Diigo Lists Tag and Flickr Photos datasets. For the others, there is either no title information contained (CiteULike and Movielens) in the datasets, or it is not possible to intersect tags and title words, which is the case for the Flickr Sets dataset, where tags are represented by numerical IDs.

Starting with the macro-level, two measures are introduced (equations 4.8 - cf. [Körner et al., 2010b] and 4.9). Both measures are based on the intersection between the set of tags and the set of title words within a user's personomy, but are normalized differently.

$$ttir_u = \frac{|T_u \cap TW_u|}{TW_u} \tag{4.8}$$

$$tvoc_u = \frac{|T_u \cap TW_u|}{T_u} \tag{4.9}$$

On micro-level, again two measures are defined. One which operates on post level (equation 4.10) and another which focuses on the tag assignment level (equation 4.11. Finally, a simple combination incorporating both characteristics simultaneously is evaluated by taking the arithmetic mean of these two measures (equation 4.12).

$$R_{tti_u} = \frac{|R_{u_{|T_r \cap TW_r| \geq 1}}|}{|R_u|} \tag{4.10}$$

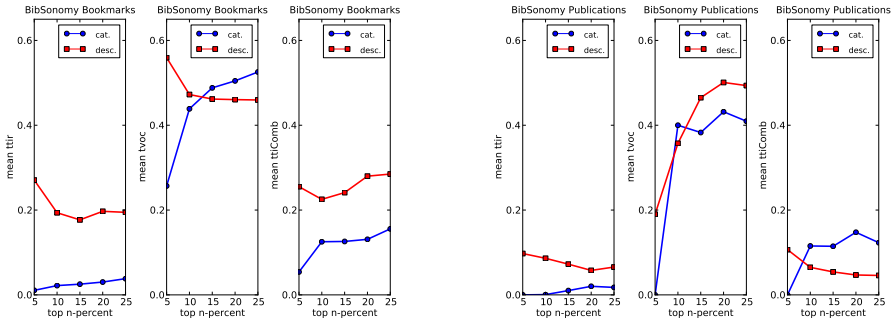
$$TAS_{tti_u} = \frac{|TAS_{u_{(t \cap TW_r) \neq \emptyset}}|}{|TAS_u|} \tag{4.11}$$

$$ttiComb_u = \frac{R_{tti_u} + TAS_{tti_u}}{2} \tag{4.12}$$

What follows is a comparison for different top-n% groups of categorizers and describers for the seven datasets these three measures ($ttir_u$, $tvoc_u$ and $ttiComb_u$)

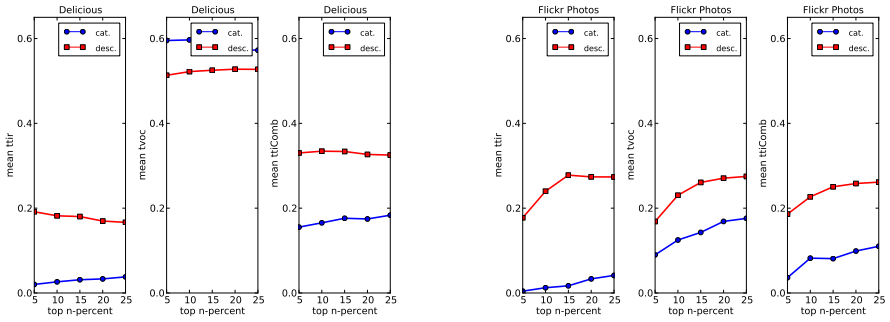
[1] `http://snowball.tartarus.org`

are applicable to. The outcome is depicted in figures 4.18 to 4.20. On the one hand, there is no clear trend for the mean $tvoc_u$ measure (equation 4.9) between the top-n% categorizers and describers, while on the other hand, higher values are scored by describers for the mean $ttir_u$ measure (equation 4.8) throughout all datasets. The reason for this must somehow arise from the different normalization since both measures are based on the same numerator and only vary in their corresponding denominator. Indeed, further investigation shows that the larger the set of tags within a personomy, the higher may the values for the intersection of the tags and title words be expected. While the $tvoc_u$ measure (equation 4.9) tries to balance this by normalizing to the size of the tag set itself, it is an indication that the normalization based on the title words favors the description approach to a certain extent. This hypothesis is reinforced by the fact, that the $ttir_u$ measure (equation 4.8) is highly correlated to the size of the tagging vocabulary $|T_u|$, which in turn shows medium correlation to the M_{comb} measure applied for the separation of users. On the contrary, no such correlations exists for the $tvoc_u$ measure (equation 4.8). Apart from that, as can be seen for the Diigo Lists Category dataset (figure 4.20(a)), values for the $ttir_u$ measure are very close to zero for both, the top categorizers and describers. This might also be a general issue for other datasets that predominantly contain personomies oriented towards a categorization approach. It therefore seems to be hard to derive the influence of resource titles on the tagging pragmatics of users on a macro-level. On the micro-level though, it's obvious that describers exhibit higher values for the mean $ttiComb_u$ measure (equation 4.12) in the majority of the datasets, apart from two exceptions. One is the BibSonomy Publications dataset (figure 4.18(b)) where results are indeed contrary to what is expected. Since this is a very small dataset it might be heavily influenced by outliers and cannot be considered significant. The other one, the Diigo Lists Category dataset (figure 4.20(a)) contains almost no describers anyway which might explain why the resulting values are so close to each other.



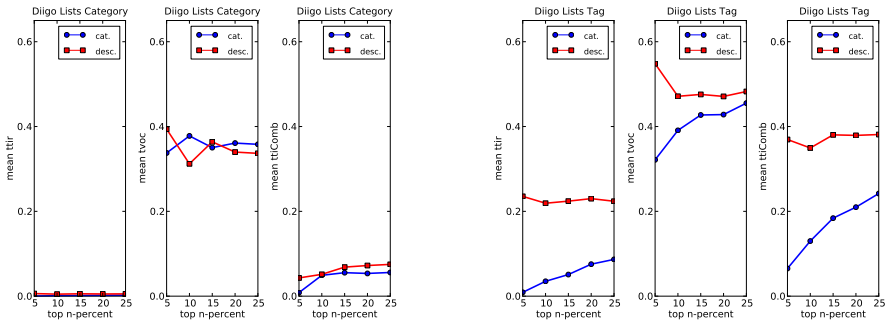
(a) BibSonomy Bookmarks (b) BibSonomy Publications

Figure 4.18: Title influence measures for BibSonomy



(a) Delicious (b) Flickr Photos

Figure 4.19: Title influence measures for Delicious and Flickr Photos



(a) Diigo Lists Category (b) Diigo Lists Tag

Figure 4.20: Title influence measures for Diigo

Summarizing the presented results from above gives reason to believe, that resource titles influence the tagging pragmatics of all users (i.e. regardless of their primary motivation) to some degree. Actually, this is in accordance with the results of [Lipczak and Milios, 2010], who found out, that there is indeed a relation between the occurrences of terms as tags and as title words. Concretely, their experiments showed that about 15% of the tags from their Delicious dataset could be found in the resource titles.
However, based on the incorporated datasets in this thesis, no substantial differences concerning the usage of tags that originate from title words could be identified when comparing different groups of categorizers and describers. Nevertheless, it seems feasible to detect slight differences when performing a micro-level analysis. The micro-level calculations allow to conclude, that title words of resources tend to affect the tag choices of describers somewhat more than that of categorizers. Again, this is in accordance with what has been recently found by [Lipczak and Milios, 2010]. They conclude, that *"users are much more willing*

to be influenced by the title if they are not planning to use the tag frequently", which in turn complies with the intuitions about the tagging characteristics of describers.

4.2.7 KL-Divergence of tag vs co-tag distributions

Based on established concepts of information theory, the Kullback-Leibler (KL) divergence, as defined for discrete random variables in section 2.3.3, is used to assess potential differences within tag-cotag distributions between categorizers and describers. Given a probability distribution the KL divergence is able to express in bits how closely it is related to a model distribution. According to [Cover and Thomas, 2006] the KL divergence can technically be interpreted as the associated "coding penalty" when choosing a candidate distribution p to approximate the model distribution q. Against a common intuition though, it cannot be considered a "distance measure" in a strict mathematical sense.
In order to measure potential differences between categorizers and describers on a macro-level, users are separated according to their M_{comb} measure. For the top 100 tags within a particular dataset, the tag vs. co-tag distributions for the dataset's whole user population are calculated which serve as model distributions. The candidate probability distributions compared to these model distributions are given by the collective tag vs. co-tag distributions based on different top-n percentages (15, 20, 25, 30, 35) of categorizers and describers respectively. Finally, the $D(p||q)$ values between these model and candidate distributions are determined. Since the KL divergence can only be calculated for probability distributions with the same number of events, the model distributions were truncated accordingly to match the length of the corresponding candidate distributions (thereby comparing the most significant co-tags). Table 4.7 lists the outcome by means of an aggregated overview.

	top 15%		top 20%		top 25%		top 30%		top 35%	
	cat.	desc.	cat.	desc.	cat.	desc.	cat.	desc.	cat.	desc.
BibSonomy Bookmarks	**81**	19	**64**	36	**66**	34	**52**	48	**63**	37
CiteULike	**62**	38	**71**	29	**80**	20	**84**	16	**82**	18
Delicious	35	**65**	47	**53**	**53**	47	**52**	48	**56**	44
Diigo Bookmarks	**57**	43	**69**	31	**74**	26	**74**	26	**77**	23
Diigo Lists Tag	50	50	**58**	42	**70**	30	**68**	32	**68**	32
Flickr Photos	**77**	23	**85**	15	**82**	18	**83**	17	**83**	17

Table 4.7: Kullback-Leibler divergence ($D(p||q)$) of tag vs. co-tag distributions for different groups of categorizers vs. describers (separated by the M_{comb} measure) for a dataset's top-100 tags. The table shows for how many tags out of the top 100, either categorizers or describers (in form of collaborative groups of varying size) achieve higher $D(p||q)$ values, where values in bold font signal wins for the respective group. There is a clear domination of higher KL divergence within the tag vs. co-tag distributions of categorizers throughout all investigated datasets, regardless of the different top-n percentage of users. The Delicious dataset is the only exception, where the results are closer and describers even score higher $D(p||q)$ values for more than half of the tags within two groups (top 15% and top 20%). These results corroborate the belief, that describers exhibit more coherent tagging activities since there seems to be a consistently higher similarity among their co-tag distributions of frequently used tags.

Chapter 5

Evaluation

This chapter deals with the qualitative evaluation of five selected measures to automatically detect whether users follow a categorization or description approach to tagging. The contents are based on the contribution to the work of [Körner et al., 2010b].

5.1 Qualitative Evaluation

In order to assess the ability of the introduced measures to detect different types of tagging motivation 3.1.2, five selected measures are further investigated. During a qualitative human subject study which has been performed on the Delicious dataset, participants were confronted with a classification task. Based on a subset of posts taken from 50 users' personomies they had to decide whether a given personomy either reflects the tagging activities of a categorizer or a describer. For that purpose, participants were shown 25 random pairs of Delicious users and had to compare their tagging records based on a subset of information that had been extracted from the correpsonding personomy using the following two strategies:

- **Resource Alignment (Table 5.1):** Posts are shown which refer to the same resources allowing participants to directly oppose the users' tag annotations given a subset of shared resources.

Resource	Tags User A	Tags User B
URL 1	Tag 1_A,...,Tag n_A	Tag 1_B,...,Tag m_B
⋮	⋮	⋮

Table 5.1: Resource alignment to compare tagging records for shared resources

- **Tag Alignment (Table 5.2):** Posts are shown which refer to the same tags allowing participants to contrast postings given a subset of shared tags

and thereby providing them with insights into several tag co-occurrences.

Posts User A - Tag 1		Posts User B - Tag 1	
resources	tags	resources	tags
URL 1_A	$t1_A,\ldots,tn_A$	URL 1_B	$t1_B,\ldots,tm_B$
⋮	⋮	⋮	⋮
	⋮		⋮
Posts User A - Tag n		Posts User B - Tag n	
	⋮		⋮

Table 5.2: Tag alignment to compare tagging records for shared tags

5.1.1 Sampling

The following five measures have been selected to be used during the human subject study: tag orphaniness (M_{desc} - equ. 3.5), normalized conditional tag entropy (M_{cat} - equ. 3.6), tag-title-intersection-ratio ($ttir_u$ - equ. 4.8), tags-to-resources-ratio (Trr_u - equ. 3.2) and personomy overlap ($P_{overlap}$ - equ. 4.6).
For each of these five measures, five user pairs from the Delicious dataset were randomly drawn out of the measure's top (i.e. potential group of describers) and bottom (i.e. potential group of categorizers) 25% respectively. Users themselves were also chosen randomly, allowing a pair of users to be drawn from either of the two groups or from one and the same group. In order to avoid a skewness towards any of the two user groups within the sample, it was guaranteed that the resulting user pairs were close to evenly distributed among their possible origins (top-top, top-bottom, bottom-top, bottom-bottom). With regard to the resource and tag alignment explained above, all resulting user pairs had to meet constraints of at least 25 shared resources and tags respectively.

5.1.2 Setup

Before starting the test, all participants were instructed about the fundamentally different tagging approaches of categorization and description using the respective intuitions listed in table 3.1. Moreover, they were provided with illustrative examples of at least two sample user pairs to get acquainted with the classification task.
During the actual task, participants were presented with the 25 user pairs, resulting from the data sampling, one at a time. To prevent from informational / cognitive overload the task had been simplified a little. Data of the resource alignment part had been restricted to show a random sample of 15 shared resources. For the tag alignment part, 5 shared tags together with at most 5 posts for each of them, are used. The resulting subsets of the respective personomies provided the basis for the human evaluation.

5.1.3 Participants

Three male as well as female participants at an average age of 28.5 years - all from an academic background - took part in this evaluation. Their tagging experience was varying and composed as follows: while four out of six stated to have some tagging practice, one subject reported much tagging experience and another one quoted to have low experience. Referring to their self-assessment only one participant would characterize himself to be a potential describer while the other five said to follow the tagging approach of categorizers.

5.1.4 Results

Inter-rater Agreement

The inter-rater agreement for all six participants was calculated using both, Fleiss' Kappa as well as pairwise Cohen's Kappa which is listed in table 5.3.

	P2	P3	P4	P5	P6
P1	0.40	0.43	0.72	0.44	0.56
P2		0.56	0.44	0.32	0.60
P3			0.49	0.45	0.62
P4				0.56	0.68
P5					0.40

Table 5.3: Inter-rater agreement among six participants (pairwise Cohen's Kappa)

In either case, Fleiss' Kappa and pairwise Cohen's Kappa are both $\kappa = 0.51$. According to (cf. [Landis and Koch, 1977]) this level of inter-rater agreement can be regarded as moderate agreement ($0.41 \leq \kappa \leq 0.60$). Given the fact that the evaluation task is at least to some extent subjective and complex by nature, the resulting kappa values appear adequate. Participants have to classify users on a small subset of a personomy, which sometimes makes it hard to recognize the correct type of the underlying tagging motivation. Such cases of indecision may have led to subjective outcomes.

Confusion Matrices

In order the quantify which of the selected measures came closest to the human ratings for 50 Delicious personomies, separate confusion matrices have been calculated. The classification results of each measure served as potential ground truth. The absolute outcomes are listed in table 5.4 while figure 5.1 holds a visual representation of the relative results. Its important to note the removal of all human classifications that ended in a draw. This enables more obvious results that have been achieved by the participants.

	True	False
tags-to-resources ratio		
Categorizers	106	26
Describers	114	12
tag orphaniness		
Categorizers	90	36
Describers	104	28
personomy overlap		
Categorizers	86	10
Describers	130	32
normalized conditional tag entropy		
Categorizers	85	41
Describers	99	33
tag-title-intersection-ratio		
Categorizers	102	36
Describers	104	16

Table 5.4: Confusion matrix results for selected measures during the user study (absolute). It interesting to see, that for all measures, there are more true describers than categorizers. Conversely, there are also less false describers than categorizers in general (apart from the personomy overlap measure). Thus, there seems to be higher accordance between the judgements of humans and the classification results of the measures for users that are motivated by description.

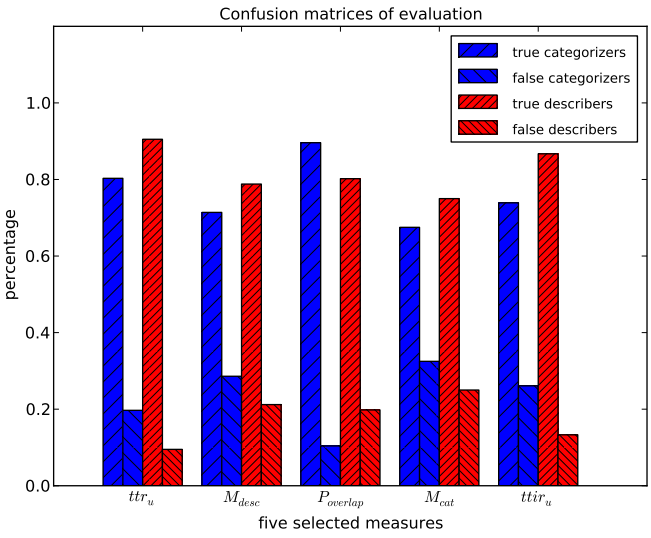


Figure 5.1: Confusion matrix results for selected measures during the user study (relative)

Accuracy

Figure 5.2 illustrates the accuracy (equ. 5.1) values that resulted from the classifications of the selected measures. The relative improvement is given by putting the results in relation to a randomly determinded baseline.

$$accuracy = \frac{\#TC + \#TD}{\#TC + \#FC + \#TD + \#FD} \tag{5.1}$$

TC ... true categorizers, TD ... true describers

FC ... false categorizers, FD ... false describers

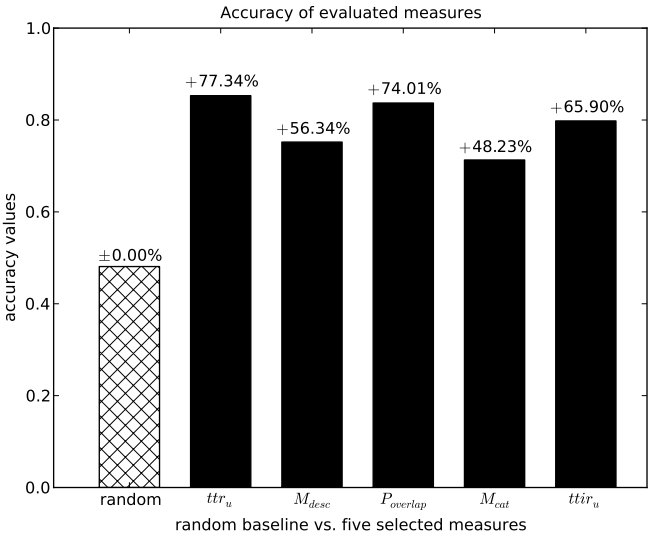


Figure 5.2: Accuracy for the five evaluated measures. With values of approximately 0.8, the three best performing measures are the tags-to-resources-ratio (Trr_u - equ. 3.2), the personomy overlap ($P_{overlap}$ - equ. 4.6) and the tag-title-intersection-ratio ($ttir_u$ - equ. 4.8). Interestingly though, the two more elaborated measures - tag orphaniness (M_{desc} - equ. 3.5) and normalized conditional tag entropy (M_{cat} - equ. 3.6) - exhibited the lowest accuracy values. This might be due to the fact, that the participants in the study primarily made their judgements based on observations that are more closely related to simpler measures. In fact, this seems natural since it would be very hard to reflect on the characteristics of elaborate measures (i.e. the tag orphaniness or the normalized conditional tag entropy) in a human subject study - especially when there is only a small excerpt of a user's personomy available during the task.

Chapter 6

Findings

This chapter summarizes key findings of this thesis. It is important to bear in mind, that each of the following observations might not be appropriate or valid to the same extent for datasets other than the ones investigated here.

6.1 Summary of the key findings

Correlations of suggested measures: There is a consistently high correlation between the M_{cat} and M_{desc} measures throughout all investigated datasets. The theoretical explanation for this is the fact, that both measures focus on similar tagging aspects. While M_{desc} directly operates on the tag distribution, M_{cat} is indirectly based on it since it operates on the resource distribution conditional to the tag distribution. The high correlation basically means that they are equally well-suited to capture the tagging motivation of users based on their respective underlying intuition Both measures should return similar results when discriminating between categorizers / describers and the remaining users in a dataset. Additionally, their consistently high correlation justifies the simple linear combination resulting in the finally suggested M_{comb} measure (see correlation tables 4.1 and B.1).

Overlap in tagging vocabulary: The results presented in tables 4.4, 4.5 and figure 4.6 show that different groups of describers, besides a single negligible exception, exhibit a higher relative overlap in their corresponding tagging vocabulary for the analyzed datasets. This remedies the objection that describers share more tags by sheer chance and at the same time reinforces the intuition that describers are to use (more) objective tags while categorizers predominantley choose (more) subjective tags.

Tag vocabulary change probability: The mean tag vocabulary change probability (TV_{cp_n}) values of describers are at least twice as high as the categorizers' after having tagged n=dataset's R_{min} number of resources for the majority (i.e.

7 out of 10) of the investigated datasets. Furthermore, TV_{cp_n} values in datasets where a categorization approach is predomiant are considerably lower than those for the average identified categorizers within other datasets (for details see section 4.2.3, figure 4.7 and 4.8).

Tag vocabulary change rate: After n=dataset's R_{min} number of resources, describers score higher mean TV_{rcr_n} values throughout all datasets. In fact, the evolution of this measure is very similar across different datasets within the two user groups. For describers the mean-end values lie mostly between $0.2 <= TV_{rcr_n} <= 0.3$ while for categorizers they are $TV_{rcr_n} <= 0.1$ in the majority of the cases (for details see section 4.2.3, figure 4.9 and 4.10).

Resource coverage of tag clouds: For the top-20% most frequent tags of the top-25% categorizers and describers (separated by the M_{comb} measure) it has been shown that the Pareto principle clearly holds for describers of all datasets. Thus, generally more than 80% of the resources can be reached with less than 20% of a describer's top-tags only. In fact, even the top 5% of the tags are sufficient in the majority of the cases. This rule usually does not apply to categorizers to the same extent (for details see section 4.2.4 figure 4.15).

Resource overlap within tag clouds: Is has been shown that $TC_{overlap}$ values are distributed rather uniformly and thus varies a lot among the users across different tagging systems. However, there is a clear tendency towards relatively low resource overlap values ($TC_{overlap} <= 0.3$) within datasets where users inherently adhere to a categorization strategy (for details see figures 4.16 and 4.17 in section 4.2.4).

Differences in users' tag agreement: Tag agreement is calculated for the top n=1000 resources for equally sized groups of categorizers and describers. Regardless of the variable k-percent user agreement value, there is a clear tendency towards describer wins for all seven datasets that are amenable to this kind of analysis. The biggest differences between categorizer and describer agreement has been identified within the Delicious dataset (for details see table 4.6 in section 4.2.5).

Influence of resource titles: The analysis for the overlap between the tags and the resources' title words shows that it is hard to measure consistent differences on a macro-level (i.e. using the $tvoc_u$ and $ttir_u$ measures on complete personomies). The $tvoc_u$ measure does not show a clear trend at all, while the $ttir_u$ measure favors describer behavior due to its high correlations with the tag vocabulary size as well as the M_{desc} and M_{comb} measures respectively. On micro-level though (i.e. using the $ttiComb_u$ measure operating on posts and tag assignments), there are slight differences showing that describers exhibit somewhat higher values in the majority of the cases. However, no substantial differences concerning the usage of

tags that originate from title words could be identified when comparing different groups of categorizers and describers (for details see figures 4.18 to 4.19(b) in section 4.2.6).

KL-Divergence of tag vs. co-tag distributions: For the top-100 tags in a particular dataset, the tag vs. co-tag distributions for the whole user population (models) are compared with the collective tag vs. co-tag distributions resulting from different top-n percentages of categorizers and describers (candidates). Calculating the KL-divergence between these model and candidate distributions shows, that apart from a single dataset, there is a clear domination of higher KL-divergence within the group of categorizers within the other datasets (for details see table 4.7 in section 4.2.7).

Qualitative Evaluation: Surprisingly, the qualitative evaluation (see chapter 5) revealed that rather simple measures (Trr_u equ. 3.2, $P_{overlap}$ equ. 4.6, $ttir_u$ equ. 4.8) reached higher accuracy than more elaborate ones (M_{cat} equ. 3.6, M_{desc} equ. 3.5). For the binary classification task of assigning users into the group of categorizer or describers this does, however, not necessarily mean that elaborate measures are worse - instead it just teaches us that simple measures obviously reflect the natural judgement of humans more closely (for details see table 5.4, figure 5.1 and figure 5.2 in section 5.1.4).

Chapter 7

Conclusion

This final chapter provides a brief summarization and discussion of the results and gives some concluding remarks as well as an outlook on potential future work.

7.1 Goals and contributions

The overall goal of this thesis was to study differences in the tagging pragmatics of individuals within and across collaborative tagging systems. By building upon selected findings from recent publications within our research group, special emphasis has been placed on contrasting two fundamentally different types of tagging motivation, namely categorization versus description. For that purpose, a number of statistical relations have been defined that aim to capture varying tagging characteristics of individual users. Based on ten acquired datasets from six different tagging systems, the thesis set out to explore whether or not there are automatically detectable differences between categorizers or describers with regard to the following tagging related aspects:

- the overlap and the evolution of users' tagging vocabulary
- properties directly derivable from the tag clouds of users
- differences in users' tag agreement
- the potential influence from titles of tagged resources
- the divergence of tag versus co-tag distributions of varying user groups

The following list briefly describes selected key findings (in relation to the aforementioned investigated aspects) that contribute towards a deeper understanding of the differences in tagging pragmatics between categorizers and describers:

- Describers exhibit a higher relative overlap in their corresponding tagging vocabulary. They also achieve higher values concerning the likelihood of

modifications to the tag vocabulary along their tagging history. Whenever describers modify their tagging vocabulary they add more tags on average than categorizers.

- With regard to the resource coverage based on the top-20% of the tags present in a user's tag cloud, it has been shown that the Pareto principle clearly holds for describers. This observation is generally less distinct for categorizers. The resource overlap within a user's tag cloud is varying a lot within and across tagging systems and thus, does not show a clear trend between categorizers and describers in principal. However, a strong tendency towards relatively low resource overlap has been detected throughout all users within datasets that are innately dominated by categorization behavior.

- Within the group of describers, there is a substantially higher tag agreement, which has been measured for the most popular resources of each dataset. This indicates that describers not only seem to have a more coherent understanding about one and the same resource, but obviously also tend to use more inter-subjective as well as (syntactically) similar tags than categorizers.

- Concerning the influence of resource titles, there are no big differences between the tagging activities of categorizers and describers on a macro-level (i.e. analyzing a user's tagging history on the whole). However, differences become noticeable when performing a micro-level analysis which operates on single tag assignments and posts respectively. It shows that describers tend to exhibit more overlap between tags and resources' title words than categorizers.

- The divergence within tag vs. co-tag distributions has been calculated for groups of users of varying size. Relating this to the resulting divergence of the whole population showed, that categorizers show higher divergence than describers in almost all investigated cases.

Apart from these findings, the thesis contributes all acquired datasets in a common XMl-based file format that can be easily utilized and thereby may provide a consistent groundwork for other researchers dealing with tagging related questions of individuals.

7.2 Limitations and future work

Besides a number of valuable insights that have been identified during the course of this thesis, there are some limitations to should be kept in mind. First, directly related to the datasets, it has not been possible to study users' tag gardening techniques like "weeding" (i.e. removing tags, replacing/renaming tags) for example.

This is simply due to technical limitations because such information cannot be acquired by means of data-crawling (neither by screen-scraping techniques nor by direct API access). Additionally, because of the varying and sometimes relatively small size of the data samples, the presented findings might not be appropriate or valid to the same extent in the general case. Especially, it should be pointed out that the incorporated datasets are too small to make any reasonable claims about the whole user population of the respective tagging system. Second, with regard to the performed analysis, there are no commonly accepted thresholds concerning the separation of users into categorizers and describers based on a certain measure so far. Therefore, whenever users had to be separated for subsequent analysis, this has occurred by simply taking different percentages of the top and bottom users, which have been identified according to the applied measure.

Finally, an outlook on both theoretical and practical future work is given:

- **Theoretical:** Until now, no hard or soft thresholds have been identified, that would specify where the discriminative quality of a certain measure starts or ceases respectively. Whenever presented measures produce indifferent results (i.e. there is a fuzzy range where it is hard to tell whether a users belongs to the group of categorizers or describers) there should be other (possibly combinable) strategies that might produce clearer outcomes. Another aspect concerns the stability of measures of over time (i.e. along a user's tagging history) which could be assessed by intensive time series analysis. Apart from that, it would be interesting to analyze differences between the bipartite and unipartite graph structures of categorizers and describers by solely using established concepts from classic network theory and put these findings into relation with the current statistical results.

- **Practical:** User interfaces of tagging systems might exploit the knowledge about users following either a categorization or description approach to tagging. For example, in addition to tag clouds, categorizers can be supported by alternative (possibly even customizable) methods and ways that help them the browse their resource repository more efficiently. Describers on the other hand, could benefit from individually adapted (tag-based) search interfaces that support the retrieval process of resources analogous to faceted search. Both groups of users can be assisted with proper tag recommendation strategies that better reflect their overall goal while tagging resources. Categorizers should benefit from tag suggestions out of their own vocabulary (or from the vocabulary of other similar categorizers) while the verbosity of describers is probably better supported with tag suggestions originating from the whole population of other describers on the system.

Appendix A

Data Structure for Personomies

A.1 Defining a data format

In order to be able to cope with all the gathered data originating from different social tagging systems, a clearly structured data format had to be designed. The decision to use XML for storing personomy data of users has been made due to the following reasons:

- Parsers and APIs available for any popular programming language
- Flexibility with regard to extensions of the contained information
- Easily transformable to other data representations by use of XSLT
- Relatively effortless data validation capabilities by means of XML Schema

Of course there are drawbacks involved with choosing XML, such as a relatively high amount of storage overhead and a little loss in data processing speed due to the parsing overhead of the XML structure. Nevertheless, those issues can be neglected because the focus in this work is primarily on offline statistical processing of individual users' data and is thus not directly affected by scalability problems (i.e. concerns about storage space or processing speed).

A.1.1 XML-Schema

Figure A.1 and the corresponding source listing below illustrate the XML-Schema that represents the simple, generic data structure used to store personomies. It has been designed to enable common access to user specific tagging data of any arbitrary (social) tagging system. By means of appropriate XSLT transformations, all user data of the six investigated tagging systems has been transformed to follow this schema.

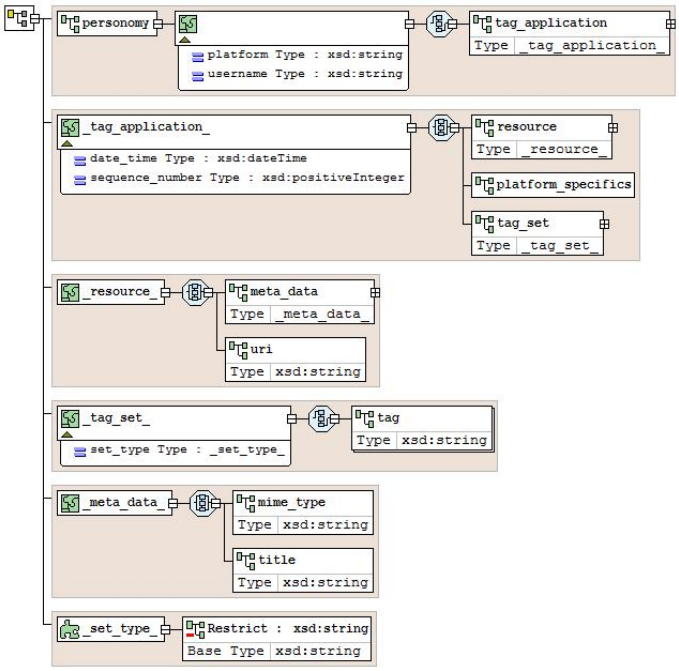


Figure A.1: XML Schema graphic for data representation

```
<?xml version="1.0" encoding="UTF-8"?>
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema">

  <xsd:element name="personomy">
    <xsd:complexType>
      <xsd:sequence>
        <xsd:element name="tag_application" type="_tag_application_"></xsd:element>
      </xsd:sequence>
      <xsd:attribute name="platform" use="required" type="xsd:string"/>
      <xsd:attribute name="username" use="required" type="xsd:string"/>
    </xsd:complexType>
  </xsd:element>

  <xsd:complexType name="_tag_application_">
    <xsd:all>
      <xsd:element name="resource" type="_resource_"></xsd:element>
      <xsd:element name="platform_specifics"></xsd:element>
      <xsd:element name="tag_set" type="_tag_set_"></xsd:element>
    </xsd:all>
    <xsd:attribute name="date_time" use="required" type="xsd:dateTime"/>
    <xsd:attribute name="sequence_number" use="required" type="xsd:positiveInteger"/>
  </xsd:complexType>

  <xsd:complexType name="_resource_">
    <xsd:all>
      <xsd:element name="meta_data" type="_meta_data_"></xsd:element>
      <!-- NOTE: We do not impose any restrictions on URI representations -->
      <xsd:element name="uri" type="xsd:string"></xsd:element>
    </xsd:all>
  </xsd:complexType>

  <xsd:complexType name="_tag_set_">
    <xsd:sequence>
      <!-- NOTE: We enforce at least one tag for every tag set of a resource -->
      <xsd:element name="tag" type="xsd:string" minOccurs="1"
        maxOccurs="unbounded"/>
    </xsd:sequence>
    <xsd:attribute name="set_type" type="_set_type_" use="required"/>
  </xsd:complexType>

```

```
41    <xsd:complexType name="_meta_data_">
42      <xsd:all>
43        <!-- NOTE: We do not restrict this to any predefined mime-types -->
44        <xsd:element name="mime_type" type="xsd:string"></xsd:element>
45        <xsd:element name="title" type="xsd:string"></xsd:element>
46      </xsd:all>
47    </xsd:complexType>
48
49    <xsd:simpleType name="_set_type_">
50      <xsd:restriction base="xsd:string">
51        <!-- NOTE:
52          * tag => general tagging purposes
53            (e.g. bookmarks in delicious)
54          * category => potential categorization
55            (e.g. sets in flickr / lists in diigo etc.)
56        -->
57        <xsd:pattern value="tag|category"/>
58      </xsd:restriction>
59    </xsd:simpleType>
60
61  </xsd:schema>
```

Listing A.1: XML-Schema code for data representation

A.1.2 Technical aspects of the crawling process

1. **User List Generation:** Whenever needed, user name collection has been done with Python based on the library Beautiful Soup[1] which is very well-suited to do any kind of HTML screen scraping in a rapid, prototypic way. The library allows to flexibly parse all kinds of HTML / XML trees, even malformed ones. By implementing individual Python-Scripts for the systems Delicious, Diigo and Flickr, it was easy to collect large lists of user names from the corresponding websites in almost no time.

2. **Tagging Data Acquisition:** The complete data acquisition process of the users' tagging activities has been done in Java and heavily relies on XML processing techniques. The main parts involved the implementation of a WebRequestHandler that performed all web-related I/O operations (i.e. direct website requests as well as REST API calls) based on URL-Templates. Additionally, specific Controller instances had to be designed in order to handle the logical flow of the data acquisition (e.g. paging / browsing through a user's bookmark history) of each different system. Finally, XMLHandlers for the XSLT transformations and XSD Schema validations had to be written, which generated and verified the resulting personomy datasets.

3. **Personomy Generation:** Wherever applicable, individual XSLT transformations had to be setup, in order to bring the heterogeneous datasets from the crawling process into the uniform XML-based data representation. The biggest advantage of the chosen approach is that further output formats (e.g. plain-text, SQL statements, SVG-Charts etc.) could be generated easily just by writing additional versions of the corresponding XSLT stylesheet transformations files. Consequently, XSLT stylesheets have been

[1] `http://www.crummy.com/software/BeautifulSoup/`

written to get SQL as well as Pajek.NET versions of each personomy, which either allows to easily import the datasets into relational databases or perform network-related calculations using Pajek and other graph libraries respectively.

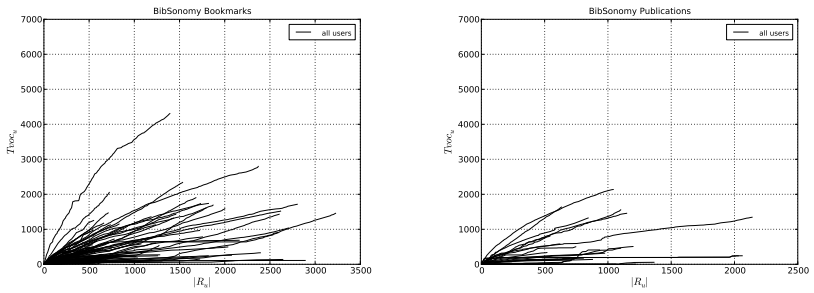
Appendix B

Further Results

This appendix chapter holds further detailed results for those datasets which could not all be presented in chapter 4 for the sake of clarity as well as due to space limitations. The same section structure has been used to illustrate the results of the statistical analysis for all investigated aspects related to tagging pragmatics.

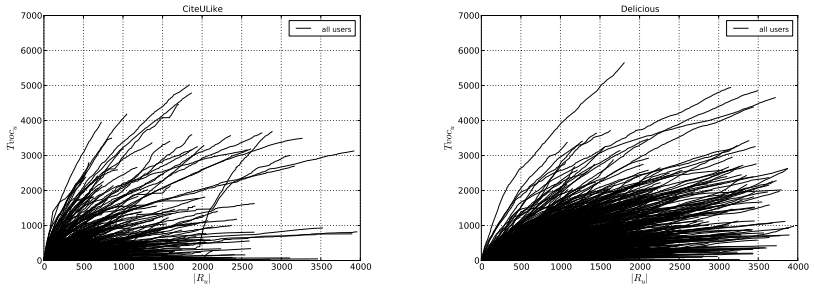
B.1 Basic tagging characteristics of the datasets

B.1.1 Growth of tagging vocabulary



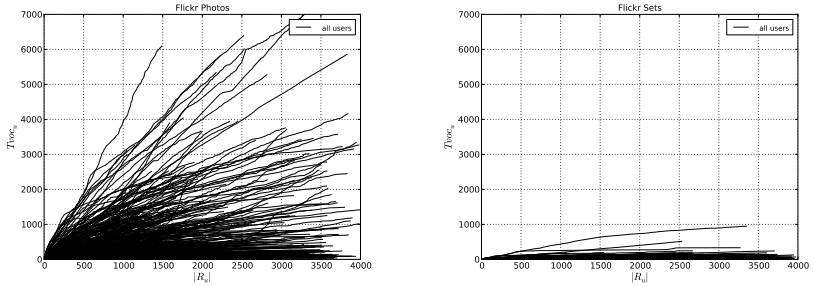
(a) BibSonomy Bookmarks (b) BibSonomy Publications

Figure B.1: Growth of tagging vocabulary for BibSonomy



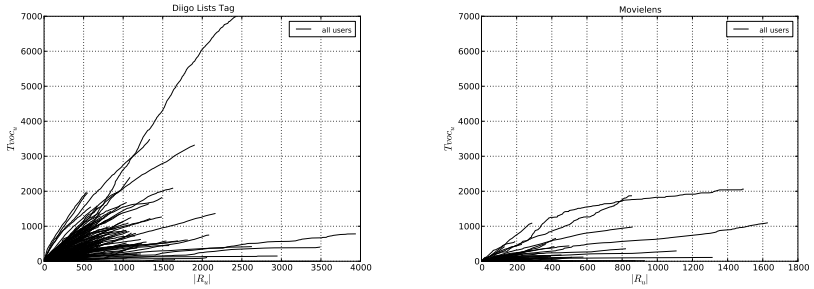
(a) CiteULike (b) Delicious

Figure B.2: Growth of tagging vocabulary for CiteULike and Delicious



(a) Flickr Photos (b) Flickr Sets

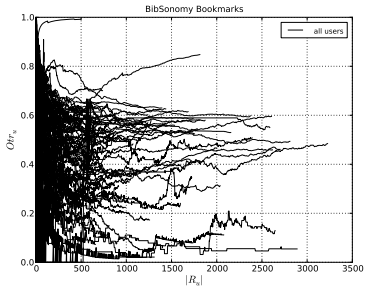
Figure B.3: Growth of tagging vocabulary for Flickr



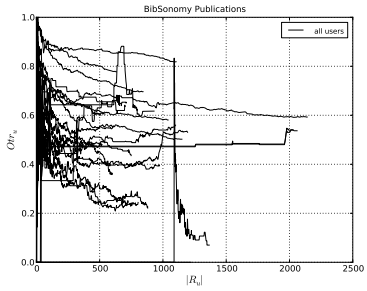
(a) Diigo Lists Tag (b) Movielens

Figure B.4: Growth of tagging vocabulary for Diigo Lists Tag and MovieLens

B.1.2 Evolution of tag orphan ratio

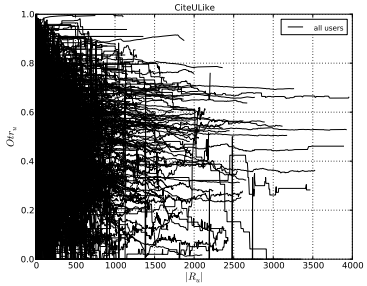


(a) BibSonomy Bookmarks

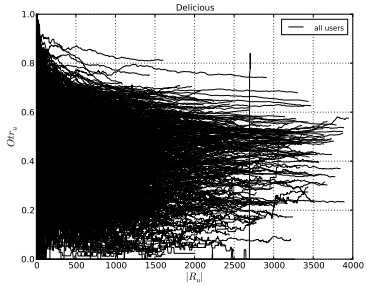


(b) BibSonomy Publications

Figure B.5: Tag orphan ratio for BibSonomy

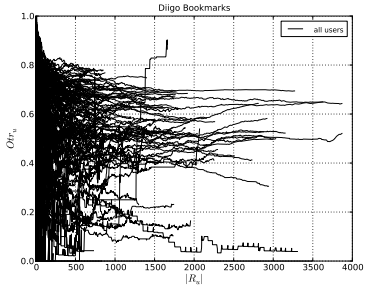


(a) CiteULike

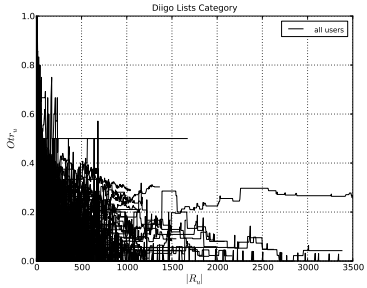


(b) Delicious

Figure B.6: Tag orphan ratio for CiteULike and Delicious

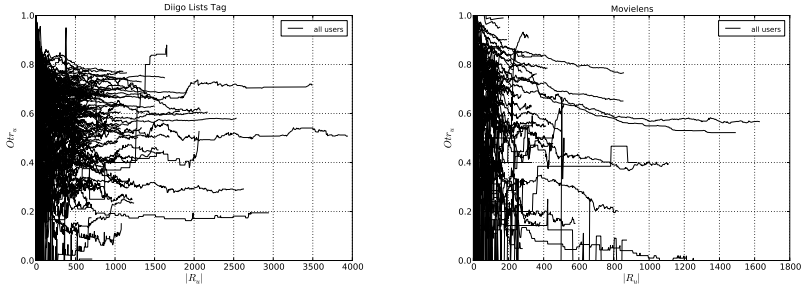


(a) Diigo Bookmarks



(b) Diigo Lists Category

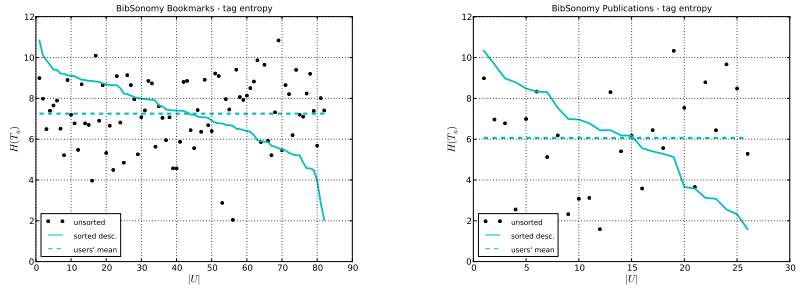
Figure B.7: Tag orphan ratio for Diigo Bookmarks and Diigo Lists Category



(a) Diigo Lists Tag (b) Movielens

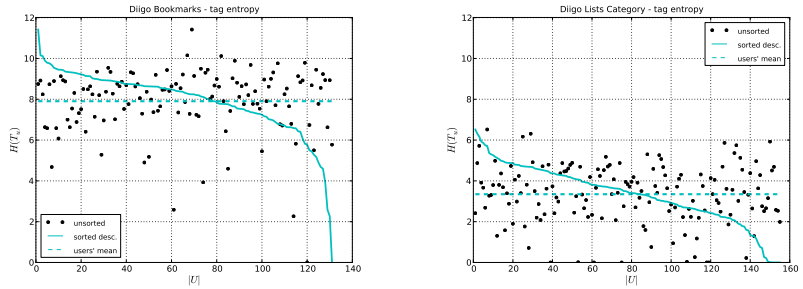
Figure B.8: Tag orphan ratio for Diigo Lists Tag and Movielens

B.1.3 Entropy and conditional tag entropy



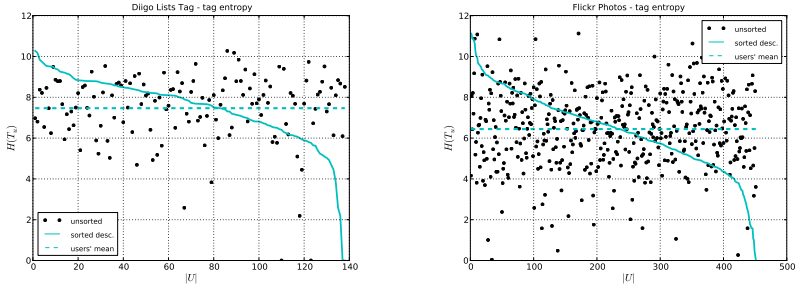
(a) BibSonomy Bookmarks (b) Bibsonomy Publications

Figure B.9: Tag entropy for BibSonomy



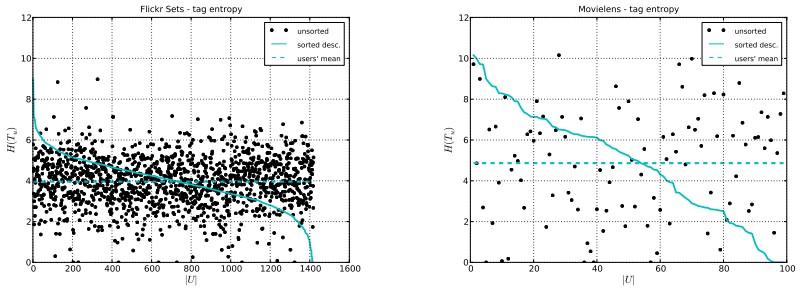
(a) Diigo Bookmarks (b) Diigo Lists Category

Figure B.10: Tag entropy for Diigo Bookmarks and Diigo Lists Category



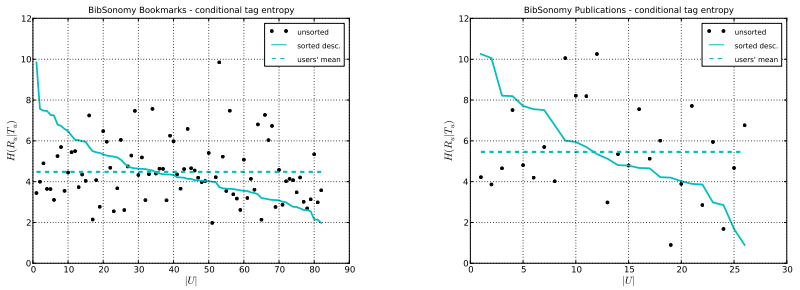
(a) Diigo Lists Tag (b) Flickr Photos

Figure B.11: Tag entropy for Diigo Lists Tag and Flickr Photos



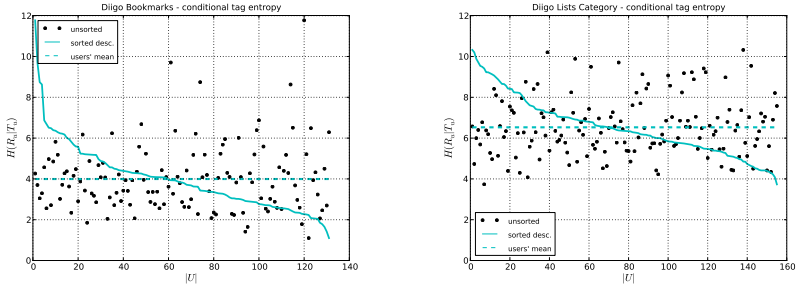
(a) Flickr Sets (b) MovieLens

Figure B.12: Tag entropy for Flickr Sets and Movielens



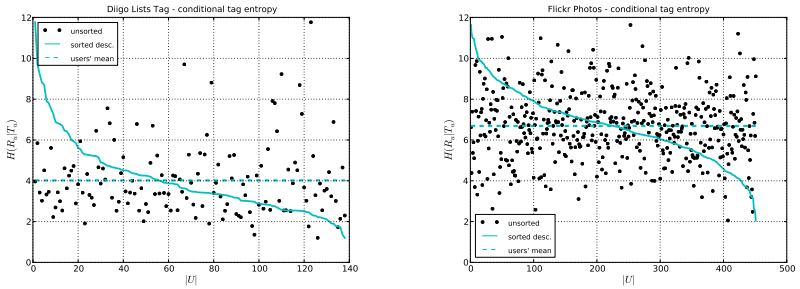
(a) BibSonomy Bookmarks (b) Bibsonomy Publications

Figure B.13: Conditional tag entropy for BibSonomy



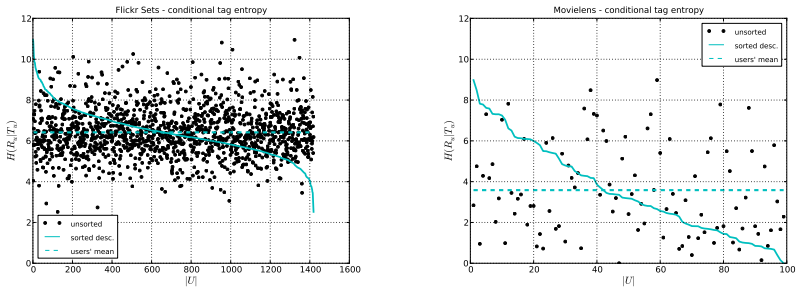
(a) Diigo Bookmarks (b) Diigo Lists Category

Figure B.14: Conditional tag entropy for Diigo Bookmarks and Diigo Lists Category



(a) Diigo Lists Tag (b) Flickr Photos

Figure B.15: Conditional tag entropy for Diigo Lists Tag and Flickr Photos



(a) Flickr Sets (b) MovieLens

Figure B.16: Conditional tag entropy for Flickr Sets and Movielens

B.2 Correlations of suggested measures

	BibSonomy Book.		BibSonomy Publ.		Diigo Book.		Diigo Lists Cat.		Diigo Lists Tag		Flickr Sets	
measures	pearson	spearman	pearson	spearman	pearson	spearman	pearson	spearman	pearson	spearman	pearson	spearman
$Tvoc_u \vert Trr_u$	**0,758**	**0,853**	**0,902**	**0,915**	0,725	0,709	**0,850**	**0,862**	0,722	**0,869**	0,685	0,690
$Tvoc_u \vert Tmean_u$	0,659	0,605	0,548	0,355	0,585	0,571	0,152	0,337	0,620	0,566	0,073	0,226
$Tvoc_u \vert Otr_u$	0,531	0,736	0,631	0,679	0,390	0,457	0,376	0,408	0,391	0,611	0,352	0,351
$Tvoc_u \vert M_{desc}$	0,530	0,630	0,461	0,509	0,498	0,625	0,248	0,300	0,420	0,530	0,251	0,305
$Tvoc_u \vert M_{cat}$	0,325	0,650	0,295	0,486	0,514	0,605	0,483	0,507	0,469	0,633	0,314	0,315
$Tvoc_u \vert M_{comb}$	0,431	0,648	0,370	0,501	0,521	0,630	0,358	0,386	0,461	0,611	0,294	0,314
$Tvoc_u \vert P_{overlap}$	0,510	0,595	0,279	0,181	0,480	0,550	0,141	0,244	0,418	0,533	0,065	0,205
$Trr_u \vert Tmean_u$	0,692	0,619	0,526	0,496	0,598	0,614	0,117	0,272	0,541	0,535	0,150	0,267
$Trr_u \vert Otr_u$	0,610	**0,803**	0,634	0,672	0,545	0,711	0,373	0,438	0,596	0,703	0,349	0,337
$Trr_u \vert M_{desc}$	0,417	0,503	0,362	0,336	0,417	0,459	0,184	0,254	0,432	0,412	0,164	0,184
$Trr_u \vert M_{cat}$	0,341	0,599	0,155	0,347	0,483	0,623	0,524	0,516	0,503	0,584	0,349	0,298
$Trr_u \vert M_{comb}$	0,393	0,565	0,242	0,350	0,467	0,575	0,333	0,364	0,486	0,534	0,255	0,256
$Trr_u \vert P_{overlap}$	0,549	0,610	0,329	0,310	0,508	0,594	0,112	0,173	0,474	0,500	0,153	0,249
$Tmean_u \vert Otr_u$	0,224	0,343	0,032	-0,068	0,000	0,100	-0,076	0,022	-0,049	0,039	-0,024	0,050
$Tmean_u \vert M_{desc}$	0,315	0,359	0,026	-0,200	0,115	0,327	-0,030	-0,034	0,110	0,296	0,149	0,147
$Tmean_u \vert M_{cat}$	0,134	0,299	-0,184	-0,235	0,162	0,313	-0,017	-0,026	0,089	0,244	0,170	0,207
$Tmean_u \vert M_{comb}$	0,219	0,339	-0,107	-0,219	0,146	0,329	-0,027	-0,032	0,101	0,276	0,168	0,200
$Tmean_u \vert P_{overlap}$	**0,848**	**1,000**	0,734	**0,928**	0,735	**0,999**	**0,968**	**0,999**	0,730	**0,999**	**0,936**	**1,000**
$Otr_u \vert M_{desc}$	**0,769**	0,700	**0,848**	**0,755**	**0,789**	0,573	0,671	**0,775**	0,747	0,561	0,596	0,641
$Otr_u \vert M_{cat}$	**0,783**	**0,838**	0,747	**0,753**	**0,798**	0,730	0,544	0,664	**0,799**	0,713	0,487	0,422
$Otr_u \vert M_{comb}$	**0,824**	**0,785**	**0,811**	**0,754**	**0,815**	0,692	0,660	**0,759**	**0,799**	0,678	0,586	0,546
$Otr_u \vert P_{overlap}$	0,230	0,332	-0,242	-0,205	0,044	0,071	-0,096	-0,047	-0,130	-0,015	-0,032	0,041
$M_{desc} \vert M_{cat}$	**0,774**	**0,939**	**0,870**	**0,938**	**0,897**	**0,875**	**0,765**	**0,794**	**0,889**	**0,853**	**0,761**	0,698
$M_{desc} \vert M_{comb}$	**0,915**	**0,980**	**0,949**	**0,968**	**0,964**	**0,944**	**0,965**	**0,962**	**0,963**	**0,940**	**0,959**	**0,878**
$M_{desc} \vert P_{overlap}$	0,317	0,349	-0,445	-0,337	0,208	0,304	-0,057	-0,113	0,089	0,255	0,125	0,138
$M_{cat} \vert M_{comb}$	**0,963**	**0,986**	**0,981**	**0,987**	**0,982**	**0,983**	**0,907**	**0,915**	**0,979**	**0,978**	**0,913**	**0,942**
$M_{cat} \vert P_{overlap}$	0,138	0,288	-0,555	-0,367	0,274	0,289	-0,057	-0,144	0,133	0,199	0,163	0,194
$M_{comb} \vert P_{overlap}$	0,222	0,329	-0,530	-0,352	0,254	0,305	-0,061	-0,145	0,118	0,233	0,149	0,186

Table B.1: Pairwise measure correlation results incorporating the remaining six datasets. Pearson's product moment correlation coefficient and spearman's rank correlation coefficient have been calculated over all users within the corresponding dataset. The heat-map style uses the following four correlation intervals: no correlation $|x| <= 0.25$ (white), low correlation $0.25 < |x| <= 0.50$ (yellow), medium correlation $0.50 < |x| <= 0.75$ (orange) and high correlation $|x| > 0.75$ (red).

B.3 Evolution of tagging vocabulary

B.3.1 Tag vocabulary change probability

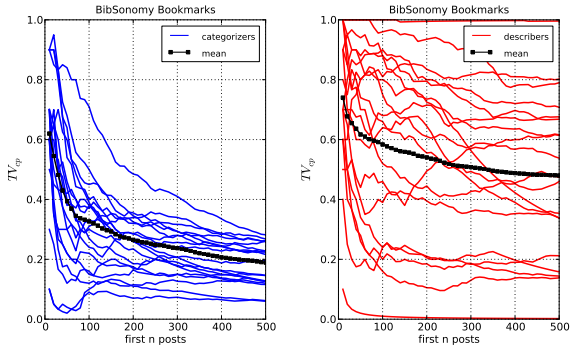


Figure B.17: Tag vocabulary change probability BibSonomy Bookmarks dataset

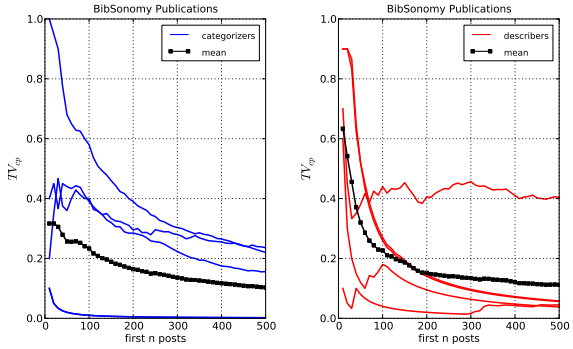


Figure B.18: Tag vocabulary change probability BibSonomy Publications dataset

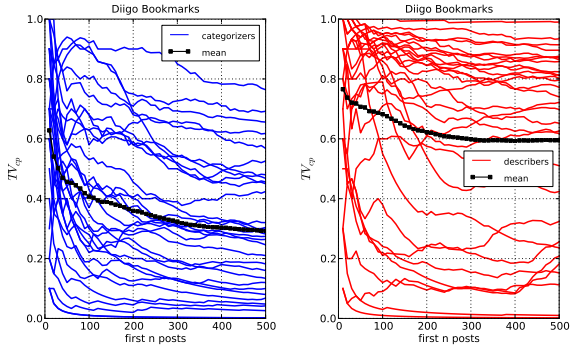


Figure B.19: Tag vocabulary change probability Diigo Bookmarks dataset

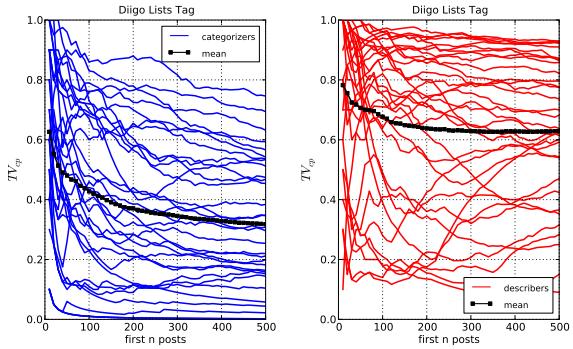


Figure B.20: Tag vocabulary change probability Diigo Lists Tag dataset

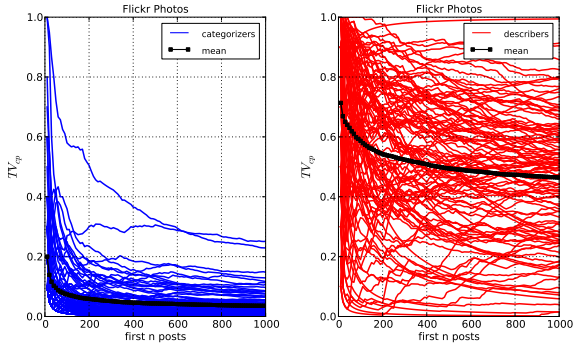


Figure B.21: Tag vocabulary change probability Flickr Photos dataset

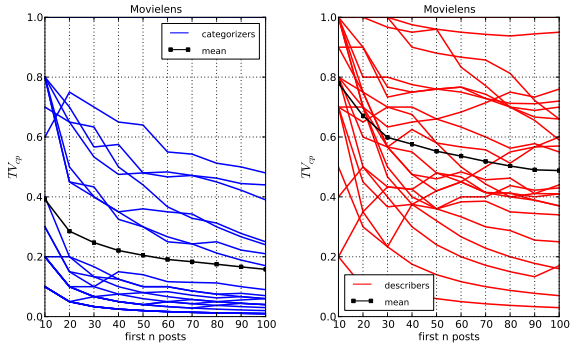


Figure B.22: Tag vocabulary change probability Movielens dataset

B.3.2 Tag vocabulary change rate

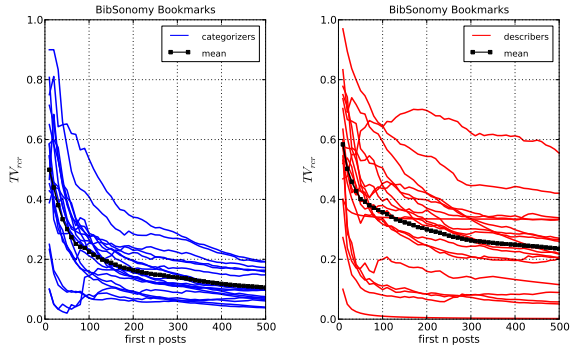


Figure B.23: Tag vocabulary change rate BibSonomy Bookmarks dataset

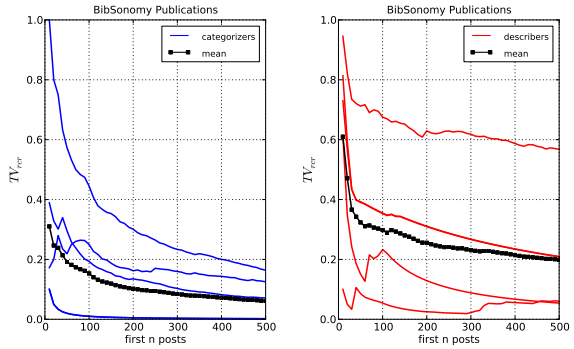


Figure B.24: Tag vocabulary change rate BibSonomy Publications dataset

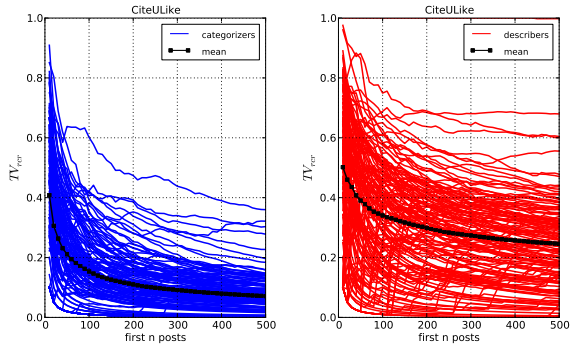


Figure B.25: Tag vocabulary change rate CiteULike dataset

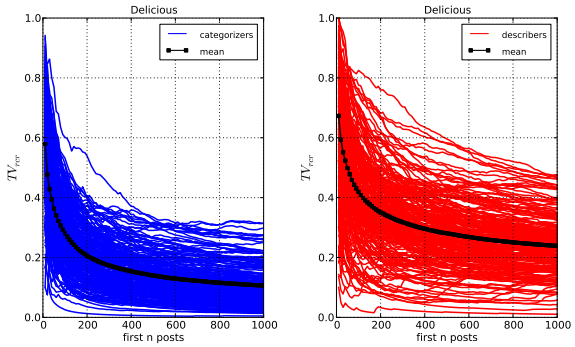


Figure B.26: Tag vocabulary change rate Delicious dataset

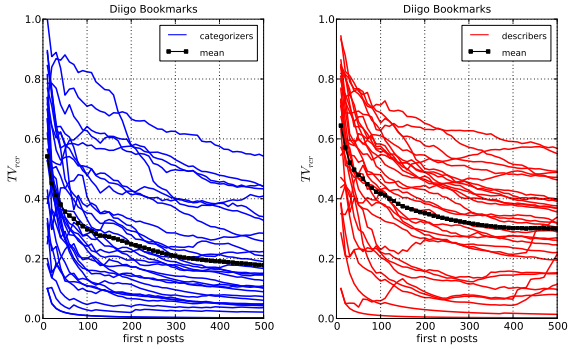


Figure B.27: Tag vocabulary change rate Diigo Bookmarks dataset

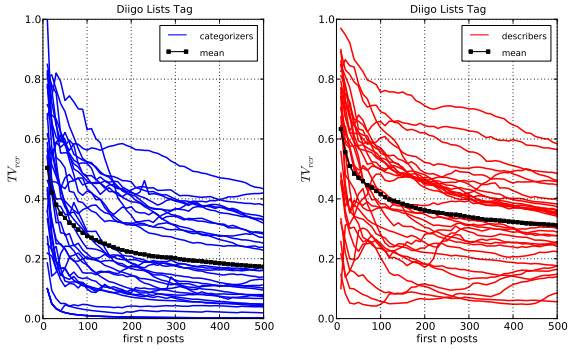


Figure B.28: Tag vocabulary change rate Diigo Lists Tag dataset

B.4 Tag cloud related properties

Figure B.29: Tag cloud of assumed categorizer from Diigo Bookmarks top-20 users (Trr_u)

Figure B.30: Tag cloud of assumed describer from Diigo Bookmarks top-20 users (Trr_u)

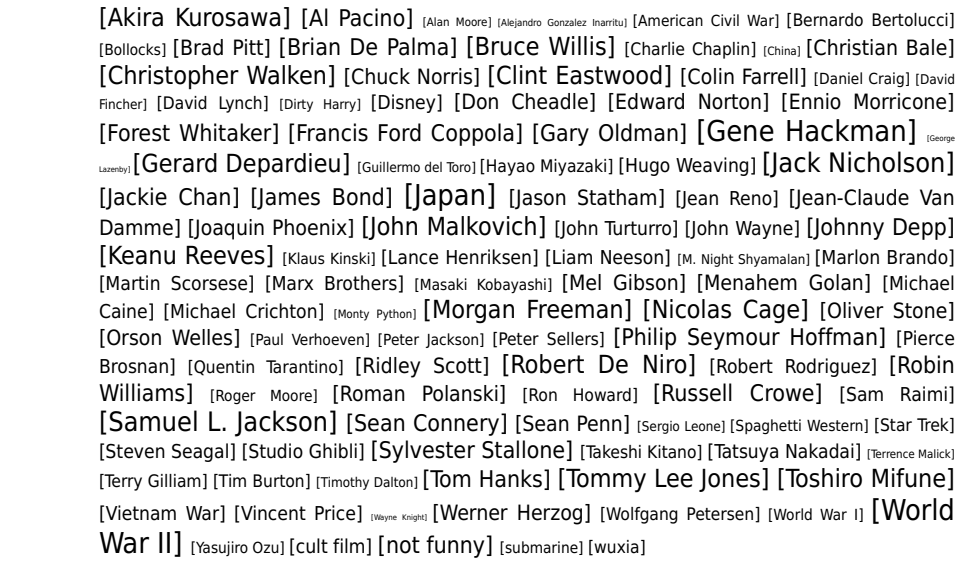


Figure B.31: Tag cloud of potential categorizer from Movielens top-20 users (M_{comb}). Note: tags are enclosed in square brackets since Movielens allows multi-word tags

Figure B.32: Tag cloud of potential describer from Movielens top-20 users (M_{comb}). Note: tags are enclosed in square brackets since Movielens allows multi-word tags

B.4.1 Resource coverage

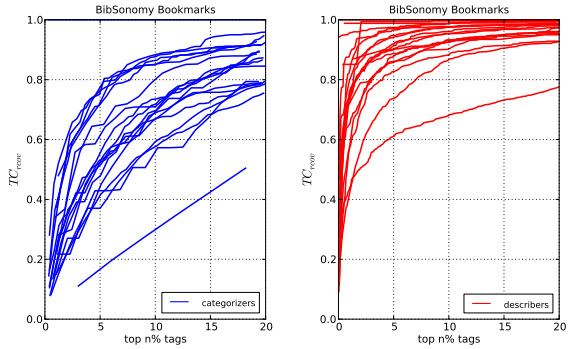


Figure B.33: BibSonomy Bookmarks resource coverage for the top-25% users and top-20% tags

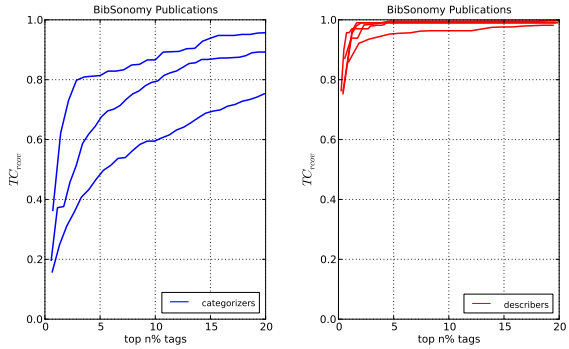


Figure B.34: BibSonomy Publications resource coverage for the top-25% users and top-20% tags

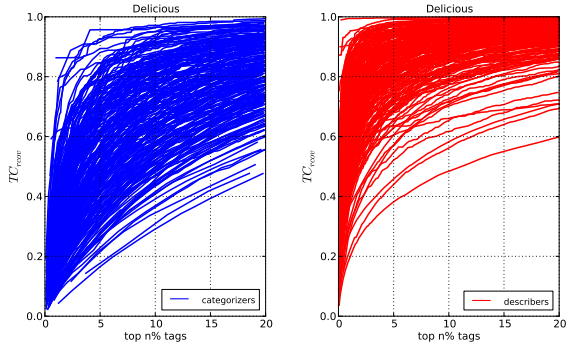


Figure B.35: Delicious resource coverage for the top-25% users and top-20% tags

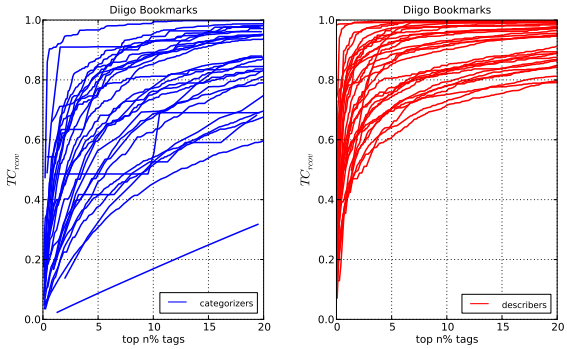


Figure B.36: Diigo Bookmarks resource coverage for the top-25% users and top-20% tags

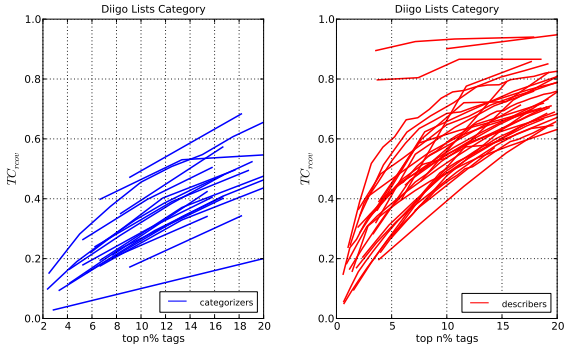


Figure B.37: Diigo Lists Category resource coverage for the top-25% users and top-20% tags

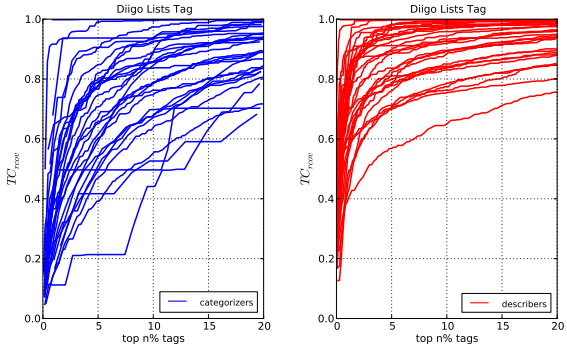


Figure B.38: Diigo Lists Tag resource coverage for the top-25% users and top-20% tags

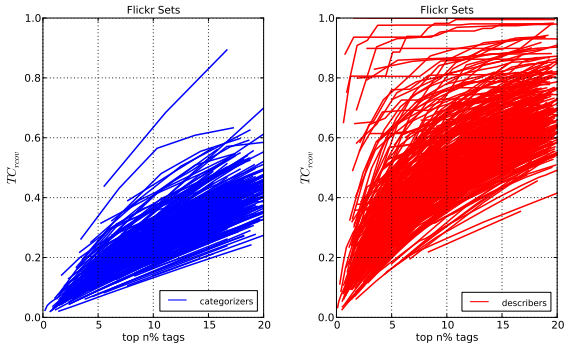


Figure B.39: Flickr Sets resource coverage for the top-25% users and top-20% tags

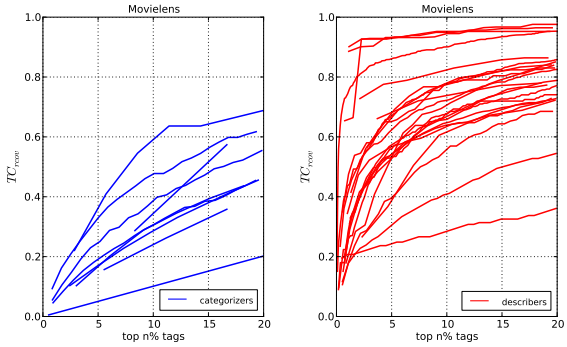
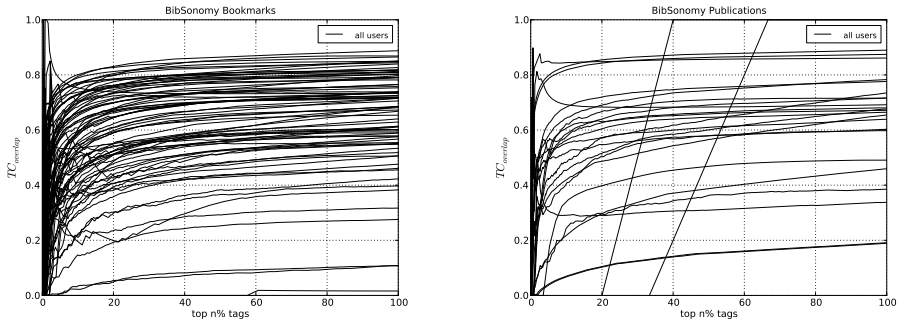


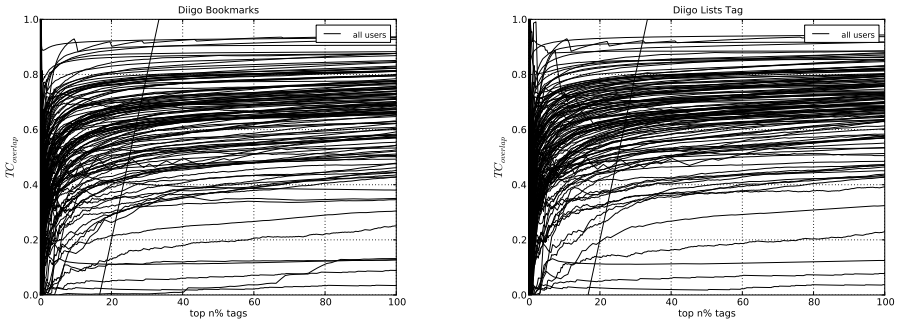
Figure B.40: Movielens resource coverage for the top-25% users and top-20% tags

B.4.2 Resource overlap



(a) BibSonomy Bookmarks (b) BibSonomy Publications

Figure B.41: BibSonomy complete resource overlap evolution



(a) Diigo Bookmarks (b) Diigo Lists Tag

Figure B.42: Diigo Bookmarks and Diigo Lists Tag complete resource overlap evolution

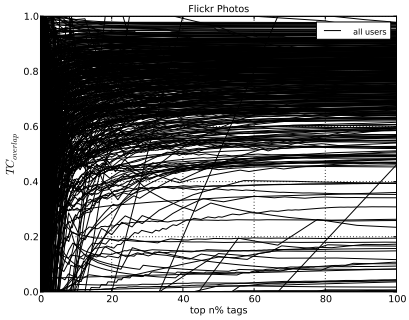


Figure B.43: Flickr Photos complete resource overlap evolution

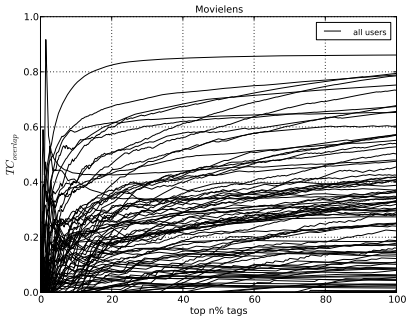
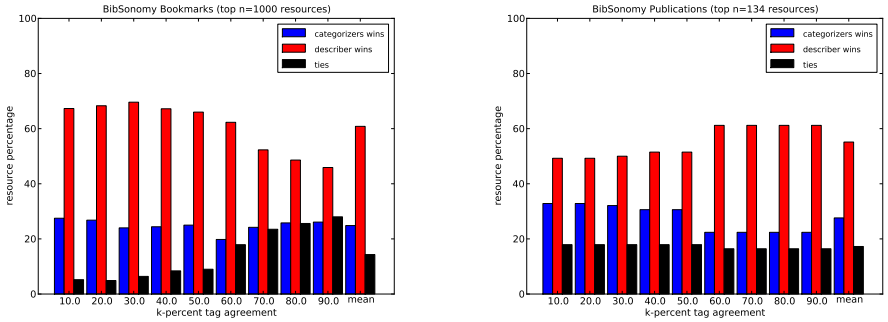


Figure B.44: Movielens complete resource overlap evolution

B.5 Differences in tag agreement



(a) BibSonomy Bookmarks (b) BibSonomy Publicationss

Figure B.45: Tag agreement results for the BibSonomy datasets

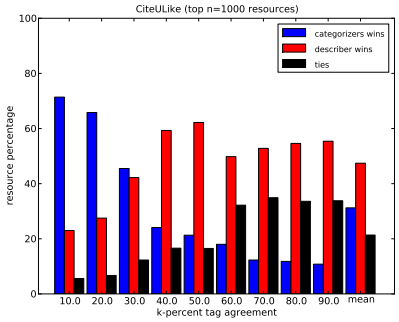


Figure B.46: Tag agreement results for the CiteULike dataset

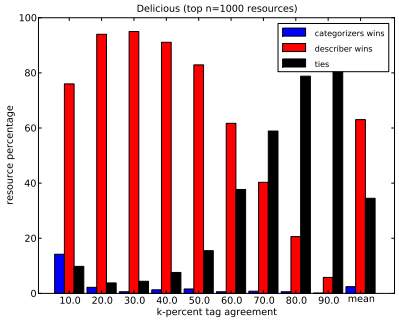
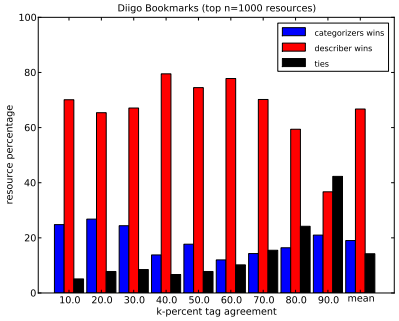
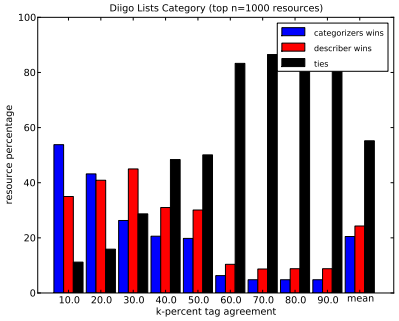


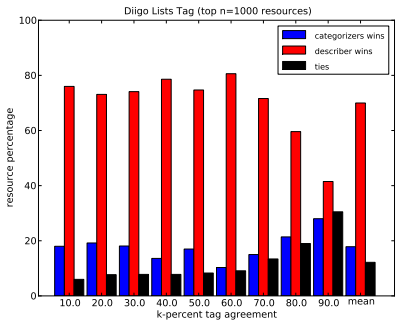
Figure B.47: Tag agreement results for the Delicious dataset



(a) Diigo Bookmarkss



(b) Diigo Lists Category



(c) Diigo Lists Tag

Figure B.48: Tag agreement results for the Diigo datasets

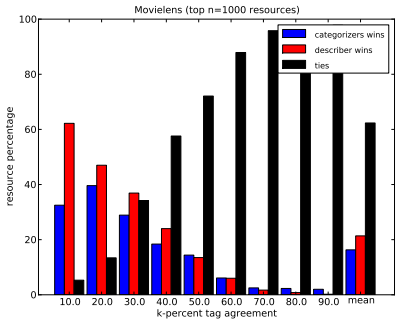


Figure B.49: Tag agreement results for the Movielens dataset

Bibliography

[Abrams et al., 1998] Abrams, D., Baecker, R., and Chignell, M. (1998). Information archiving with bookmarks: personal web space construction and organization. In *CHI '98: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 41–48, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

[Albert and Barabasi, 2002] Albert, R. and Barabasi, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47.

[Alvarado et al., 2003] Alvarado, C., Teevan, J., Ackerman, M. S., and Karger, D. (2003). Surviving the information explosion: How people find their electronic information. Technical Report AIM-2003-006, MIT AI Lab.

[Ames and Naaman, 2007] Ames, M. and Naaman, M. (2007). Why we tag: motivations for annotation in mobile and online media. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 971–980, New York, NY, USA. ACM.

[Bao et al., 2007] Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., and Su, Z. (2007). Optimizing web search using social annotations. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 501–510, New York, NY, USA. ACM.

[Barbaro and Zeller, 2006] Barbaro, M. and Zeller, T. (2006). A face is exposed for aol searcher no. 4417749. *The New York Times*.

[Boardman and Sasse, 2004] Boardman, R. and Sasse, M. A. (2004). Stuff goes into the computer and doesn't come out: a cross-tool study of personal information management. In Dykstra-Erickson, E. and Tscheligi, M., editors, *CHI*, pages 583–590. ACM.

[Brin and Page, 1998] Brin, S. and Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117.

[Broder et al., 2000] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the Web. *Computer Networks*, 33(1-6):309–320.

[Cattuto et al., 2006] Cattuto, C., Loreto, V., and Pietronero, L. (2006). Collaborative tagging and semiotic dynamics. *CoRR*, abs/cs/0605015. arXiv:cs/0605015v1.

[Civan et al., 2008] Civan, A., Jones, W., Klasnja, P., and Bruce, H. (2008). Better to organize personal information by folders or by tags?: The devil is in the details. *ASIST 2008 Annual Meeting (AM08 2008)*, Columbus, Ohio, October 24-29, 2008.

[Coates, 2005] Coates, T. (2005). Two cultures of fauxonomies collide... *Weblog*.

[Cover and Thomas, 2006] Cover, T. and Thomas, J. (2006). *Elements of information theory*. Wiley-Interscience.

[Diestel, 2005] Diestel, R. (2005). *Graph theory*. Springer, Berlin, 3rd edition.

[Easley and Kleinberg, 2010] Easley, D. and Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning about a Highly Connected World (Draft Version: February 13, 2010) To be published 2010*. Cambridge University Press.

[Golder and Huberman, 2005] Golder, S. and Huberman, B. A. (2005). The structure of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208.

[Golle, 2006] Golle, P. (2006). Revisiting the uniqueness of simple demographics in the us population. In Juels, A. and Winslett, M., editors, *WPES*, pages 77–80. ACM.

[Halpin et al., 2007] 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*, pages 211–220, New York, NY, USA. ACM.

[Hammond et al., 2005] Hammond, T., Hannay, T., Lund, B., and Scott, J. (2005). Social bookmarking tools (i): A general review. *D-Lib Magazine*, 11(4).

[Heckner et al., 2009] Heckner, M., Heilemann, M., and Wolff, C. (2009). Personal information management vs. resource sharing: Towards a model of information behaviour in social tagging systems. In *Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, San Jose, CA, USA.

[Heymann et al., 2008] Heymann, P., Koutrika, G., and Garcia-Molina, H. (2008). Can social bookmarking improve web search? In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 195–206, New York, NY, USA. ACM.

[Hotho et al., 2006] Hotho, A., Jäschke, R., Schmitz, C., and Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. In Sure, Y. and Domingue, J., editors, *The Semantic Web: Research and Applications*, volume 4011 of *LNAI*, pages 411–426, Heidelberg. Springer.

[Jacob, 2004] Jacob, E. K. (2004). Classification and categorization: a difference that makes a difference. *Library Trends*, 52(3):515–540.

[Jiang and Conrath, 1997] Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *CoRR*, cmp-lg/9709008.

[Jones, 2007] Jones, W. (2007). *Keeping Found Things Found: The Study and Practice of Personal Information Management.* Academic Press.

[Jones, 2008] Jones, W. (2008). How is information personal? *CHI 2008 Workshop, April 5-6 2008 Florence, Italy.*

[Kern et al., 2010] Kern, R., Körner, C., and Strohmaier, M. (2010). Are tags used to categorize or describe resources? In *The 14th European Conference on Research and Advanced Technology for Digital Libraries (ECDL) 2010*, Glasgow, United Kingdom.

[Kleinberg, 1999] Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632.

[Körner, 2009] Körner, C. (2009). The motivation behind tagging. In *ACM SIGWEB Hypertext'09 Graduate Student Research Challenge (Poster).*

[Körner et al., 2010a] Körner, C., Benz, D., Strohmaier, M., Hotho, A., and Stumme, G. (2010a). Stop thinking, start tagging - tag semantics emerge from collaborative verbosity. In *Proceedings of the 19th International World Wide Web Conference (WWW 2010)*, Raleigh, NC, USA. ACM.

[Körner et al., 2010b] Körner, C., Kern, R., Grahsl, H. P., and Strohmaier, M. (2010b). Of categorizers and describers: An evaluation of quantitative measures for tagging motivation. In *21st ACM SIGWEB Conference on Hypertext and Hypermedia (HT 2010)*, Toronto, Canada. ACM.

[Landis and Koch, 1977] Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

[Lansdale, 1988] Lansdale, M. (1988). The psychology of personal information management. *Applied Ergonomics*, 19(1):55–66.

[Lipczak and Milios, 2010] Lipczak, M. and Milios, E. E. (2010). The impact of resource title on tags in collaborative tagging systems. In Chignell, M. H. and Toms, E., editors, *HT*, pages 179–188. ACM.

[Malone, 1983] Malone, T. W. (1983). How do people organize their desks? implications for the design of office information systems. *ACM Transactions on Office Information Systems*, 1:99–112.

[Marlow et al., 2006] Marlow, C., Naaman, M., Boyd, D., and Davis, M. (2006). Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA. ACM.

[Mathes, 2004] Mathes, A. (2004). Folksonomies - cooperative classification and communication through shared metadata. `http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html` [last accessed 2010-08-23].

[Miller, 1995] Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, 38(1):39–41.

[Morrison, 2008] Morrison, P. J. (2008). Tagging and searching: Search retrieval effectiveness of folksonomies on the world wide web. *Information Processing Management*, 44(4):1562 – 1579.

[Narayanan and Shmatikov, 2008] Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. *Proc. of. 29th IEEE Symposium on Security and Privacy*, 0:111–125.

[Newman, 2003] Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

[Nov et al., 2009] Nov, O., Naaman, M., and Ye, C. (2009). Motivational, Structural and Tenure Factors that Impact Online Community Photo Sharing. In *ICWSM '09: Proceedings of AAAI International Conference on Weblogs and Social Media*.

[Ohm, 2009] Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. working draft.

[Page et al., 1998] Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia.

[Sen et al., 2006] Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F. M., and Riedl, J. (2006). tagging, communities, vocabulary, evolution. In *CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 181–190, New York, NY, USA. ACM.

[Shirky, 2005] Shirky, C. (2005). Ontology is overrated: Categories, links and tags. `http://www.shirky.com/writings/ontology_overrated.html` [last accessed 2010-08-23].

[Simon, 1955] Simon, H. A. (1955). On a class of skew distribution functions. 42(3/4):425–440.

[Sinha, 2005] Sinha, R. (2005). A cognitive analysis of tagging. `http://rashmisinha.com/2005/09/27/a-cognitive-analysis-of-tagging` [last accessed 2010-08-23].

[Smith and Medin, 1981] Smith, E. E. and Medin, D. (1981). *Categories and Concepts*. Harvard university press, Cambridge, MA.

[Strohmaier, 2008] Strohmaier, M. (2008). Purpose tagging - capturing user intent to assist goal-oriented social search. In *Workshop on Search in Social Media SSM'08, in conjunction with CIKM'08, Napa Valley, USA*.

[Strohmaier et al., 2010a] Strohmaier, M., Körner, C., and Kern, R. (2010a). Why do users tag? detecting users' motivation for tagging in social tagging systems. In *International AAAI Conference on Weblogs and Social Media (ICWSM2010), Washington, DC, USA, May 23-26*.

[Strohmaier et al., 2010b] Strohmaier, M., Trattner, C., Helic, D., and Andrews, K. (2010b). Network-theoretic potentials and limitations of tag clouds as a tool for social navigation. (to be published 2010).

[van Rijsbergen, 1979] van Rijsbergen, C. (1979). *Information Retrieval second edition*. Dept. of Computing Science, University of Glasgow, London: Butterworths.

[Vander Wal, 2004] Vander Wal, T. (2004). Folksonomy :: vanderwal.net. `http://vanderwal.net/folksonomy.html` [last accessed 2010-08-23].

[Wash and Rader, 2007] Wash, R. and Rader, E. (2007). Public bookmarks and private benefits: An analysis of incentives in social computing. *Proceedings of ASIST Annual Meeting '07*.

[Wasserman and Faust, 1994] Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press.

[Weller and Peters, 2008] Weller, K. and Peters, I. (2008). Seeding, weeding, fertilizing. different tag gardening activities for folksonomy maintenance and enrichment. In Auer, S., Schaffert, S., and Pellegrini, T., editors, *Proceedings of I-Semantics '08, International Conference on Semantic Systems*, pages 100–117.

[Whittaker and Sidner, 1996] Whittaker, S. and Sidner, C., editors (1996). *Email overload: exploring personal information management of email: CHI '96: Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press.

[Xu et al., 2006] Xu, Z., Fu, Y., Mao, J., and Su, D. (2006). Towards the semantic web: Collaborative tag suggestions. In *Proceedings of the Collaborative Web Tagging Workshop at the WWW 2006*, Edinburgh, Scotland.

[Yanbe et al., 2007] Yanbe, Y., Jatowt, A., Nakamura, S., and Tanaka, K. (2007). Can social bookmarking enhance search in the web? In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 107–116, New York, NY, USA. ACM.

[Yule, 1925] Yule, U. G. (1925). A mathematical theory of evolution based on the conclusions of dr. j. c. willis, f.r.s. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character:*, 213:21–87.