Quantifying the Value of Peer-Produced Information in Social Tagging Systems

by

Elizeu Santos-Neto

B. Computer Science, Universidade Federal de Alagoas, 2002

M. Computer Science, Universidade Federal de Campina Grande, 2004

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Electrical and Computer Engineering)

The University Of British Columbia
(Vancouver)

August 2014

Abstract

Commons-based peer production systems are marked by three main characteristics, they are: radically decentralized, non-proprietary, and collaborative. Peer production is in stark contrast to maket-based production and/or on a centralized organization (e.g., carpooling vs. car rental; couchsurfing vs. hotels; Wikipedia [1] vs. Encyclopaedia Britannica).

Social tagging systems represent a class of web systems, where peer production is central in their design. In these systems, *decentralized* users collect, share, and annotate (or tag) content *collaboratively* to produce a *public* pool of annotated content. This uncoordinated effort helps filling the demand for labeling an ever increasing amount of user-generated content on the web with textual information. Moreover, these labels (or simply *tags*) can be valuable as input to mechanisms such as personalized search or content promotion.

Assessing the value of individuals' contributions to peer production systems is key to design user incentives to bring high quality contributions. However, quantifying the value of peer-produced information such as tags is intrinsically challenging, as the value of information is inherently contextual and multidimensional. This research aims to address these two issues in the context of social tagging systems.

To this end, this study sets forth the following hypothesis: *assessing the value of peer-produced information in social tagging systems can be achieved by harnessing context and user behavior characteristics.* The following questions guide

[1] http://www.wikipedia.com

the investigations:

Characterization: (*Q1*). What are the characteristics of individual user activity? (*Q2*). What are the characteristics of social user activity? (*Q3*). What are the aspects that influence users' perception of tag value?

Design: (*Q4*). How to assess the value of tags for exploratory search? (*Q5*). What is the value of peer-produced information for content promotion?

This study applies a mixed methods approach. The findings show that patterns of user activity can inform the design of supporting mechanisms for tagging systems. Moreover, the results suggest that the proposed method to assess value of tags is able to differentiate between valuable tags from less valuable tags, as perceived by users. Moreover, the analysis of the value of peer-produced information for content promotion shows that peer-produced sources can oftentimes outperform expert-produced sources.

Preface

Although I am the main author of the studies presented in this dissertation, the results presented in the following chapters are the product of collaborative efforts. I had the pleasure to work with researchers from the University of British Columbia (my advisor, Matei Ripeanu), University of South Florida (David Condon and Adriana Iamnitchi), Universidade Federal de Campina Grande (Nigini Oliveira and Nazareno Andrade), Universidade Federal de Minas Gerais (Flavio Figueiredo, Tatiana Pontes, and Jussara Almeida), and HP Labs - Bristol (Miranda Mowbray).

It is worth noting that this dissertation consists of research studies that have been published (or are under review) in peer-reviewed international conferences, workshops, and journals. First, the characterization study presented in Chapter 2, Chapter 3, and Chapter 4 led to the four publications and submissions below:

- Elizeu Santos-Neto, Flavio Figueiredo, Nigini Oliveira, Nazarendo Andrade, Jussara Almeida, Matei Ripeanu. *Assessing Tag Value for Exploratory Search*. Under review.

- Elizeu Santos-Neto, David Condon, Nazareno Andrade, Adriana Iamnitchi, Matei Ripeanu. *Reuse, Temporal Dynamics, Interest Sharing, and Collaboration in Social Tagging Systems*. First Monday, Vol. 19 (7), August, 2014.

- Elizeu Santos-Neto, David Condon, Nazareno Andrade, Adriana Iamnitchi, Matei Ripeanu. *Individual and Social Behavior in Tagging Systems*. In the

20th ACM Conference on Hypertext and Hypermedia, Torino, Italy, June 2009 (acceptance rate: 32%)

- Elizeu Santos-Neto, Matei Ripeanu, Adriana Iamnitchi. *Content Reuse and Interest Sharing in Tagging Communities*. The *AAAI 2008 Spring Symposia on Social Information Processing*, Stanford, CA, USA, March 2008.

- Elizeu Santos-Neto, Matei Ripeanu, Adriana Iamnitchi. *Tracking User Attention in Collaborative Tagging Communities*. In Proceedings of the *International ACM/IEEE Workshop on Contextualized Attention Metadata: Personalized Access to Digital Resources*, Vancouver, BC, Canada, June 2007.

The results related to the system design part, which are reported in Chapter 5 and Chapter 6, are presented in the following refereed publications:

- Elizeu Santos-Neto, Flavio Figueiredo, Nigini Oliveira, Nazarendo Andrade, Jussara Almeida, Matei Ripeanu. *Assessing Tag Value for Exploratory Search*. Under review.

- Tatiana Pontes, Elizeu Santos-Neto, Jussara Almeida, Matei Ripeanu. *Where Are the 'Key' Words? On the Optimization of Multimedia Content Textual Attributes to Improve Viewership*. Under review.

- Elizeu Santos Neto, Tatiana Pontes, Jussara Almeida, Matei Ripeanu. *On the Choice of Data Sources to Improve Content Discoverability via Textual Feature Optimization*. In the Proceedings of ACM Hypertext Conference (HT'2014), Santiago de Chile, Chile, September, 2014.

- Elizeu Santos Neto, Tatiana Pontes, Jussara Almeida, Matei Ripeanu. *Towards Boosting Video Popularity via Tag Selection*. In the Proceedings of the ACM Workshop on Social Multimedia and Storytelling (SoMuS), April 2014, Glasgow, UK.

- Elizeu Santos-Neto. *Characterizing and Harnessing Peer-Production of Information in Social Tagging Systems*. In the Proceedings of the *5th ACM International Conference on Web Search and Data Mining* Doctoral Consortium (WSDM 2012 Doctoral Consortium), Seattle, WA, USA, February 2012.

- Elizeu Santos-Neto, Flavio Figueiredo, Jussara Almeida, Miranda Mowbray, Marcos Gonalves, Matei Ripeanu. *Assessing the Value of Contributions in Tagging Systems*. In the Proceedings of the *2nd IEEE International Conference on Social Computing (SocialCom'2010)*, Minneapolis, MN, August 2010. (acceptance rate 15%)

Finally, it is worth noting that part of this research has been approved by the UBC Behavioral Research Ethics Board under the approval number H11-02039.

Table of Contents

List of Tables

List of Figures

Acknowledgments

First and foremost, I must thank Matei Ripeanu – an academic advisor turned into a friend – for challenging me all the way during this journey; and, for being a risk-taker by investing his time on investigations I decided to pursue.

Second, this work would have been no fun without my fabulous collaborators and *external advisors*. My special thanks to you all (in order of appearance): Walfredo Cirne (UFCG, Google), Franscisco Brasileiro (UFCG), Nazareno Andrade (UFCG), Ian Foster (U. of Chicago), Adriana Iamnitchi (USF), David Condon (USF), Abdullah Gharaibeh (UBC), Lauro Beltão Costa (UBC), Flavio Figueiredo (UFMG), Jussara Almeida (UFMG), Miranda Mowbray (HP Labs, Bristol), Dinan Gunawardena (Microsoft Research, UK/MSRC), Thomas Karagiannis (MSRC), Milan Vojnovic (MSRC), Alexandre Poutiere (MSRC), Tatiana Pontes (UFMG), and Nigini Oliveira (UFCG).

Third, thanks to all members of NetSysLab for their feedback and discussions; the volunteers who participated in parts of this study; and the support from: Univertisy of British Columbia (UBC), British Columbia Innovation Council (BCIC), Natural Sciences and Engineering Research Council of Canada (NSERC), and Association of Universities and Colleges of Canada (AUCC).

Fourth, thanks to my Google colleagues for enriching internship experiences and great SF2G rides: Thomas Kotzmann, Foad Dabiri, Viresh Ratnakar, Daniel Fireman, and Guilherme Germoglio.

Finally, thanks to Cypress Mt. and Mt. Seymour for being there when I needed them the most.

Dedication

To my family, friends, Luciana, Anaïs, and Amália,
who constantly help me to be a better human being.

Chapter 1

Introduction

Commons-based peer production systems (or simply, peer production systems) are marked by three main characteristics, they are: radically decentralized, non-proprietary, and collaborative [8]. Peer production is in stark contrast to modes of production based on markets and/or on a centralized organization (e.g., Wikipedia [1] vs. Encyclopaedia Brittanica; ontology created by experts vs. folksonomies).

As peer production abounds in today's World Wide Web, with hundreds of millions of distributed users collaborating towards the production of information goods (e.g., Wikipedia, *Flickr* [2], *YouTube* [3], and *Delicious* [4]), studying these systems to understand users' motivation to contribute to their communities and their perception of value of peer-produced information is important. In fact, understanding peer production of information the World Wide Web will enable us to improve user experience both in existing and to design the next generation of such systems.

Social tagging systems (often referred to as *collaborative tagging systems*, or simply tagging systems) are web systems, where peer production is central in their design. In social tagging systems, users collect, share and annotate (or tag) content

[1] http://www.wikipedia.com

[2] http://www.flickr.com

[3] http://www.youtube.com

[4] http://www.delicious.com

collaboratively. For example, in Delicious, users bookmark URLs and annotate these URLs with free form words (i.e., *tags*). As multiple users may annotate the same URL, the user community collaboratively produces a large-scale catalog of annotated content that may enable the design of improved mechanisms such as personalized search and recommendation [56, 93, 100].

As tagging becomes an effective way to collect useful metadata about content and users' interests, user behavior characteristics at the individual and social level, and their relationship to peer production are important to the design of mechanisms that improve users' experience [40, 99]. Therefore, unveiling usage characteristics is important to inform the design of current and future tagging systems, in particular, and peer production systems, in general.

More importantly, social tagging systems are inherently collaborative, and, as such, studying social tagging through the lenses of commons-based peer production has the potential to enable the design of key supporting mechanisms such as incentives to boost high quality participation. More specifically, designing and evaluating methods that assess the value one participant produces to other individuals in the community is a fundamental building block to build other mechanisms. For example, measuring how one user's tags helps other users search more efficiently could help designing incentive mechanisms that improve the quality of tags.

To fill these gaps, this thesis present efforts on two main directions:

- First, a characterization of social tagging systems based on both a qualitative investigation of user perceptions of value and a quantitative analysis of records of user activity. In particular, the characterization aims to understand how and *why* users contribute to tagging systems and *what* are their perceptions about others' contributions;

- Second, the design and evaluation of methods to assess the value of user contributions in social tagging systems. In particular, this thesis investigates the value of tags in two contexts: exploratory search and content promotion.

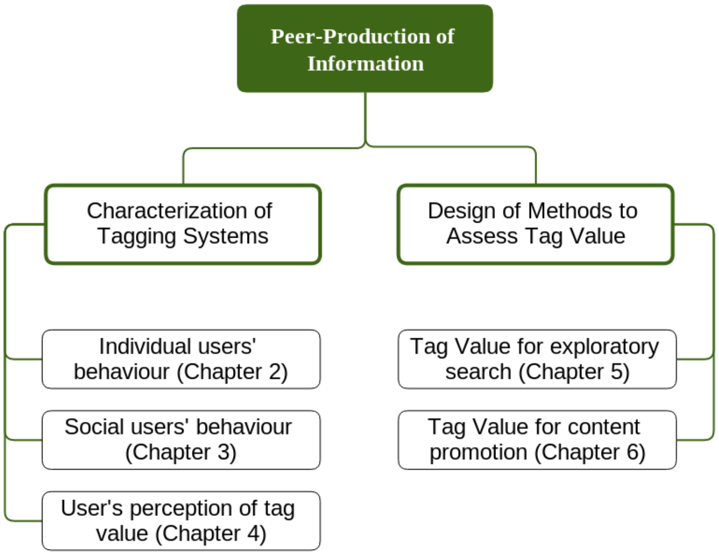

Peer-Production of Information
Characterization of Tagging Systems
Design of Methods to Assess Tag Value
Individual users' behaviour (Chapter 2)
Social users' behaviour (Chapter 3)
User's perception of tag value (Chapter 4)
Tag Value for exploratory search (Chapter 5)
Tag Value for content promotion (Chapter 6)

Figure 1.1: The logical structure of this research.

Figure 1.1 illustrates the logical structure of this research, while the rest of this chapter presents the background of this research, the research questions, and the contributions. First, it presents the definition of *commons-based peer production systems* [8] (Section 1.1). Next, it discusses the peer production of information in *tagging systems*, which is the specific area this research contributes to (Section 1.2). The discussion follows with the introduction of research questions that guide this thesis (Section 1.3), the methodology adopted to address these questions (Section 1.4), and a summary of contributions (Section 1.5). Finally, this chapter concludes with the presentation of the structure of this dissertation.

1.1 Online Peer Production Systems

Commons-based peer production systems are "*systems where production is radically decentralized, collaborative and non-proprietary*" [8]. Benkler uses the term *commons* to highlight that the production mode in the these systems resembles that of a common property regime, where participants share a common pool of resources [67]. In this sense, commons-based peer production systems, such as carpooling [5], represent the opposite of production systems based on private property, such as car rental.

Commons-based peer production spans a wide range of scenarios, both in the physical world (e.g., carpooling, community choirs) and over the Internet (e.g., open source software, Q&A portals, social media websites). This research focuses on the latter class of scenarios. In particular, it concentrates on a subclass of peer production systems that are concerned with the *online peer production of information* (see Figure 1.2). Systems in this subclass use the Internet to mobilize *decentralized* participants who *collaboratively* produce and share information. As it happens in most peer production systems, individuals contribute to a common pool of resources without any enforced hierarchy.

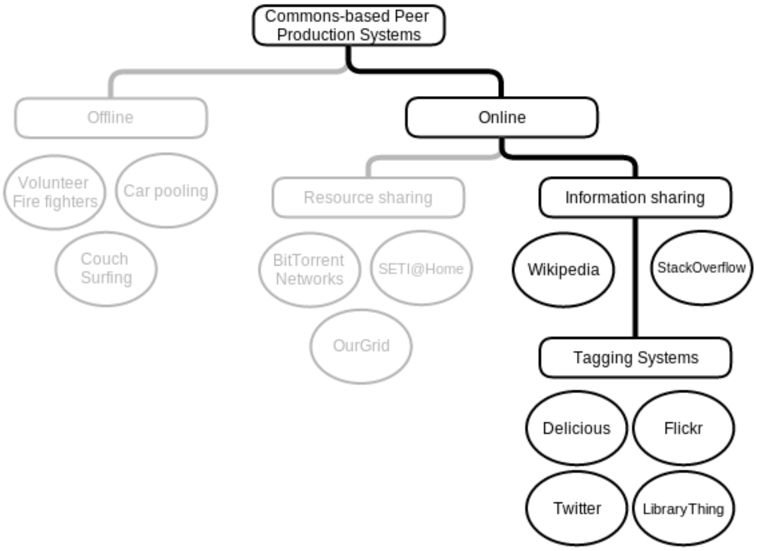


Figure 1.2: Illustration of classes (squares) and instances (ellipsis) of commons-based peer production systems.

[5] http://en.wikipedia.org/wiki/Carpool

Encyclopedia Britannica and *Wikipedia*, for instance, illustrate the contrasts between the proprietary mode of information production and its commons-based counterpart. In the former, the production of articles follows a traditional model, where authors are coordinated by a centralized entity, to produce a proprietary good (i.e., the encyclopedia articles). In the latter, however, authors are inherently decentralized, contribute to a public pool of articles, and hierarchy emerges by consensus. It is worth noting that *Encyclopedia Britannica* has recently incorporated a few elements of peer production to its services (e.g., users can submit articles), which serves as evidence of the advantages of peer production in some contexts.

1.2 Social Tagging Systems

Along the same lines as *Wikipedia* in terms of collaboration, yet focusing on the production of a different type of information, many social systems target the demand for social content sharing and personal content management [37]. Systems like *CiteULike* [6], *Delicious*, *YouTube* [7], and *Flickr* [8] are commonly referred to as *social tagging systems*. These systems provide users with the capability to annotate content with free-form words (or tags), as well as social networking features. In fact, tagging features are commonplace even in major online social systems like *Twitter* [9], Facebook [10], and *Google+* [11], where social networking is a more prominent fature.

Although each of these systems targets different types of content, users, and provide unique features, their tagging capabilities are conceptually the same, in terms of the abstract entities. In social tagging systems each user maintains a

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

[7] http://www.youtube.com

[8] http://www.flickr.com

[9] http://www.twitter.com

[10] http://www.facebook.com

[11] http://plus.google.com

library: a collection of annotated items (e.g., photos, videos, URLs, or textual posts). For example, in *CiteULike*, users collect citation records linked to online articles, while in *Delicious*, users bookmark URLs to generic web pages, and in *Google*+, *Twitter*, and *Facebook*, users post textual or multimedia content. A user may assign tags to items in her library (e.g., tags in *Delicious*, or hashtags in *Twitter*). Additionally, a user may also tag items in other user's public library. Tags may serve to group items, as a form of categorization, or to help find items in the future [29, 66]. The tagging activity can be private (i.e., only the user who generated the tags and items can access these annotations) or public. A user can see what (public) tags other users assigned to an item when she is tagging it, thus the user is able to reinforce the choice of tags as appropriate by repeating the tags previously assigned to that item.

1.3 Research Questions

This section describes the specific goals of this research towards both the *characterization* of tagging systems and the *design* of methods to assess the value of tags.

Characterization of Individual User Activity (Chapter 2)

The rationale behind characterizing tagging systems is that usage patterns can inform the design of mechanisms and supporting infrastructure of these systems. In this mindset, this research focuses on the following aspects of tagging systems:

- **RQ1**. *What are the characteristics of individual user activity?* In tagging systems, activity can be described in terms of production (i.e., content publication and annotation) and consumption of information (i.e., search, navigation). This research question is divided into two specific aspects of users' activity, as follows:
  - **RQ1.1**. *Are there any patterns of information production in social tagging systems?* Besides the activity distributions, the rate of infor-

mation production overtime provides valuable insights on the growth of the system. More specifically, the goal is to determine whether users are annotating more frequently the content that is available in the system, or they are more likely to publish new content. Understanding this aspect is important to inform the design of techniques that aim at predicting user behavior based on past activity such as recommendation systems, as a high rate of new items poses an extra challenge to recommendation algorithms [2], while a high rate of tag reuse among users may help alleviating this problem [78, 100] (Section 2.3).

  - **RQ1.2**. *What are the temporal dynamics of tag vocabularies?*. Studying the evolution of tag vocabularies for individual users, in terms of vocabulary size and tag usage frequency, complements the characterization of information production in a tagging system. Patterns in vocabulary evolution may help designing mechanisms that rely on tags as indicators of user preferences. For example, assuming that tags are used to represent user interests, personalization mechanisms [18, 56] could use the rate of vocabulary change (growth or tag frequency usage) to determine the shifting window of interests users might have) (Section 2.4).

Characterization of Social User Activity (Chapter 3)

Social tagging systems provide features to users to engage in online social behaviour (e.g., collaborative content creation, curation, and sharing). This part of the thesis investigation focures on the following questions:

- **RQ2**. *What are the characteristics of social user activity?* Understanding the social aspects of user activity is an important step to complement the analysis of individual user activity. More importantly, it can unveil characteristics that can inform system design. This characterization focuses on two main aspects of user social activity, as guided by the following particular questions:

- **RQ2.1** *How is the strength of implicit user ties based on activity similarity distributed across the system?* As a first step on understanding social user behaviour in tagging systems, this work characterizes whether the strength of implicit ties between users are concentrated on a small subset of user pairs, or evenly spread across the population. Understanding this aspect can be harnessed by mechanisms that aim to detect communities of users with shared interest over specific topics, for instance (Section 3.2).
- **RQ2.2** *Are there relationships between implicit and explicit user ties?* There are different types of ties between users: *implicit ties,* as inferred from the similarity on their tagging activity; and, *explicit ties* such as declared co-membership in discussion groups or friendship links. The goal is to characterize the intensity of implicit ties and determine whether one type of ties contains information about its counterpart. Understanding this relationship between the type of ties enables a finer interpretation of explicit ties, as the implicit ties provide a measure of the strength of existing explicit links between users (Section 3.3).

Characterizing Users' Perception of Tag Value (Chapter 4)

Tagging systems are a subclass of online peer production systems. However, quantifying users' contribution in tagging systems poses new challenges compared to other peer production systems, where the contribution value is often linked to the amount of resource shared. In particular, users *contribute information* to tagging systems, which is fundamentally different from other peer production systems, where users share units of physical resources (e.g., bandwidth, CPU and storage). This contrast demands a study on the users' perception of value of peer-produced information in the context of social tagging systems.

To this end, this research moves to a qualitative characterization of the perception of value and their records of activity. The goal is to inform the design of methods that quantify the value of tags. In particular, this work focuses on the following two aspect of tagging systems design:

RQ3. *What are the aspects that influence users' perception of value of tags?* All tags are not created equal. Due to many factors such as context in which the the tag is being used and personal interests, a user will naturally consider some tags more important than others. To inform the design of methods that assess the value of tags, it is important to understand what are the aspects users take into account when choosing to use tags in some contexts such as exploratory search (also known as navigation). The results of this investigation directly inform the design of methods that assess the value of tags (Chapter 4).

Assessing the Value of Tags (Chapter 5, Chapter 6)

Inspired by both the qualitative and the quantitative characterizations of users' perception of tag value and users' tagging activity, this thesis moves towards investigating methods that automatically assess the value of peer-produced information in two relevant contexts for both information seekers and information producers in social tagging systems. In particular, this part of the thesis focuses on assessing the value of peer-produced information in two contexts: *exploratory search* and *content promotion*, as described in the following:

RQ4. *How to assess the value of tags for exploratory search?* In the light of the aspects that influence users' perception of tag value, one can design methods to automatically quantify the value of tags while aiming to capture the identified aspects. The goal is to formalize the aspects, design, and evaluate methods that quantify the value of tags from the perspective information seekers (i.e., users who use tags to discover new content) (Chapter 5).

RQ5. *How to assess the value of tags for content promotion?* The characterization of how users perceive tag value also provides insights on the aspects that influence tag production. In particular, some users have clear goals that drive the choice of tags such as promoting online content. Therefore, given the availability of tags and peer-produced information, it is paramount to assess the value of such sources as inputs to content producers and whether these sources can improve their success metrics (e.g., popularity of videos) (Chapter 6).

1.4 Summary of Methodology

This research was conducted using *mixed methods* [39] (i.e., a combination of both quantitative and qualitative research methods). In particular, quantitative methods are used to study the overall characteristics of tagging systems by analysing traces of user tagging activity. To this end, I have applied statistical methods and simulations. Part of the characterization of user behavior resorts to qualitative methods. More specifically, I used grounded theory [39] and in-depth interviews [39] to study what are the aspects information consumers take into account when choosing tags in an exploratory search (Chapter 4).

The combination of these methods provides primarily two important angles on user behavior in social tagging systems. Additionally, it enables the design of mechanisms while starting from the users' perspective of what is important, instead of using unvalidated assumptions.

1.5 Summary of Contributions

This section summarizes the contributions of this thesis. I note that the contributions are part of a series of refereed articles and technical reports [65, 75–81]. Each contribution briefly described below maps to each specific research question, as stated in the previous section, in order.

Characterization of Individual User Activity

- **RQ1.1.** *Users tend to reuse tags already present in the system more often than they repeatedly tag existing items* [77, 78]. This finding supports the intuition that tags are primarily a content categorization instrument. Additionally, the results show that the difference between the levels of tag reuse and repeated item tagging vary across different systems. This observation suggests that features such as tag recommendation and the type of content play a role in the patterns of peer production of information in tagging systems.

- **RQ1.2.** *The tag vocabulary of a user can be approximated by a small portion of her activity* [79]. The experiments on the evolution of user tag vocabularies show that only to accurately approximate the characteristics of a tag vocabulary, only a small percentage of the initial tag assignments performed by a user is necessary. These observed results can applied in the context of applications that rely on activity similarity scores between users, for example, as it provides a way to reason about the trade offs between accuracy of a user activity profile and the computational cost of updating the similarity scores.

Characterization of Social User Activity

- **RQ2.1.** *The strength of implicit social ties is concentrated over small portion of user pairs*. Moreover, the observed strength of activity similarity between pairs of users are the result of shared interest as opposed to generated by chance. The distributions of activity similarity strength deviate significantly from those produced by a Random Null Model (RNM) [71]. This suggests that the implicit ties between users, as defined by their activity similarity levels, capture latent information about user relationships that may offer support for optimizing system mechanisms.

- **RQ2.2.** *The average strength of implicit ties is stronger for user pairs with explicit ties* [78]. This investigation analyzes the similarity between users according to their tagging activity and its relation to explicit indicators of collaboration. The results show that the users' activity similarity is concentrated on a small fraction of user pairs. Also, the observed distributions of users' activity similarity deviate significantly from those produced by a *Random Null Model* [71]. Finally, an analysis of the relationships between implicit relationships based on activity similarity and other more explicit relationships, such as co-membership in discussion groups, shows that user pairs that tag items in common have in average higher similarity in terms of co-membership in discussion groups.

Characterization of Users' Perception of Tag Value

To complement the quantitative characterization and to inform the design of methods that assess the value of tags, this research conducts a qualitative characterization of user' perception of tag value. A summary of the major findings in this investigation is presented below:

- **RQ3.** *Users perception of tag value in exploratory search is multidimensional and the key aspects that influence users' perception are: relevance of items retrieved and reduction of search space [81].* Based on a qualitative characterization of users' perception of tag value in the context of exploratory search, this study finds that the two most salient aspects that influence users' perception of tag value are: *ability to retrieve relevant content items* and *ability to reduce the search space*. These findings inform the design of a method that quantifies the value of tags automatically by taking into account the important aspects, which are identified by the qualitative analysis.

Methods to Assess Value of Peer-Produced Information

Finally, this research proposes new techniques that exploit the usage characteristics of tagging systems to improve their design. The next paragraphs briefly describe the contributions related to studying social tagging as commons-based peer production systems and the design of methods to assess the value of user contribution in these collaborative contexts. Chapter 5 and Chapter 6 distills the proposed approaches and results in details.

Important to note that there are two perspectives to the problem of assessing the value of peer-produced information in tagging systems: the consumer and the producer. The goal is to design methods that cater for each of these perspectives. For consumers, assessing the value of tags are considered in the context of exploratory search, while for producers, the method takes into account the ability of a tag to improve the viewership of content (e.g., a YouTube video).

- **RQ4.** *An information-theoretical approach to assess the value of tags for exploratory search provides accurate estimates of value as perceived by users.* This study first provides a framework that help specifying components of methods that assess the value of user contributios in social tagging systems. In particular, this part of the research provides a method that automatically quantifies the value of tags that caters for the two desirable properties in the context of exploratory search, as identified by the qualitative user study. A proof shows that the proposed method has desirable theoretical properties while quantifying these two aspects. Additionally, an experiment using real tagging data that shows that the proposed method accurately quantifies the value of tags according to users' perception.

- **RQ5.** *Peer-produced information, though lacking formal curation, has comparable value to that of expert-produced information sources when used for content promotion.* An analysis of online videos provides evidence that the tags associated with a sample of popular movie trailers can be optimized further by an automated process: either by incorporating human computing engines (e.g., Amazon Mechanical Turk) at a much lower cost than using dedicated *channel managers* (the current industry practice); or, at an even lower cost, by using recommender algorithms to harness textual produced by a multitude of data sources that are related to the video content. To this end, I perform a comparison of the effectiveness of using peer- and expert-produced sources of information as input for tag recommender that aim to boost content popularity.

1.6 Dissertation Structure

This dissertation is naturally divided into three parts: *i*) quantitative characterization of user activity; *ii*) qualitative characterization of users' perception of tag value; and, *iii*) system design. The first part consists of Chapter 2 (RQ1) and Chapter 3 (RQ2) that presents a characterization of both individual and social user behaviour while using quantitative research methods. The second part, as presented in Chapter 4 (RQ3), focuses on a qualitative analysis of users' perception of tag value to inform system design. The third part, which is presented as Chapter 5 (RQ4), Chapter 6 (RQ5), focus on the design of methods to assess the value of tags from the perspective of both information seekers and content promotion. Finally, Chapter 7 presents the final remarks and directions for future work. Note that each chapter contains its respective related work section to position the contributions among the related literature.

Chapter 2

Characterizing Users' Individual Behavior

Tagging systems [63] are a ubiquitous manifestation of online peer-production of information [8], a production mode commonplace in today's World Wide Web [70]. The annotation feature, often referred to as simply *tagging*, has been originally designed to support personal content management. However, as this feature exposes user preferences and their temporal dynamics, similarities between users, and the aggregated characteristics of the user population, annotations have been recognized for their potential to support a wider range of mechanisms such as social search [24], recommendation [32, 56], and spam detection [54]. Therefore, a better understanding through characterization and modeling of usage patterns is necessary to fully realize the full potential of this feature.

This chapter presents quantitative characterization results that complements previous characterization studies (presented in Section 2.1) [1]. In particular, this chapter focuses on two major aspects of the tagging activity that have attracted relatively little attention in the past: *i*) the dynamics of peer production of tags and items; and, *ii*) the temporal dynamics of users' tag vocabularies [64, 75, 76, 78]. More specifically, the following questions guide the quantitative characterization

[1] The results presented in this chapter appeared at the following references: [64, 76, 78, 81]

of individual user behavior:

- **RQ1.1**. *Are there any patterns of information production in social tagging systems?* (Section 2.3).
- **RQ1.2**. *What are the temporal dynamics of tag vocabularies?* (Section 2.4).

To study the patterns of information production in social tagging systems, Section 2.3 concentrates on two metrics: *i*) item re-tagging, a measure of the degree to which items are repeatedly tagged; and *ii*) tag reuse, a measure of the degree to which users reuse a tag to perform new annotations.

The analysis of the evolution of the users tag vocabularies (i.e., the set of tags a user assigns to her items) in Section 2.4 focuses on the evolution of the user vocabularies over time.

This study uses activity traces from three distinct tagging systems: *CiteULike*, *Connotea* and *Delicious* (Section 2.2). This selection of systems samples the diversity of the tagging ecosystem, as they are three emblematic tagging systems for the type of content they target, with *CiteULike* and *Connotea* concentrating in bookmarking of academic citations, and *Delicious* focusing on general URLs. The in-depth analysis of these three systems reveals regularities and relevant variations in tagging behavior.

The main findings of this characterization study are:

- The characteristics of peer production of information are qualitatively similar across systems but differ quantitatively, as suggested by the observed rates of item re-tagging and tag reuse. In all three systems investigated, users produce new items at higher rate than they produce new tags. However, the observed rates in *CiteULike* and *Connotea* are different from *Delicious*. As the three systems provide essentially similar annotation features, these findings suggest that the target audience and the type of annotated content play an important role in the users' tagging behavior (Section 2.3).

- User tag vocabularies are constantly growing, but at different rates depending on the age of the user. However, despite the constant increase in size, the relative usage frequency of tags in a vocabulary converges to a stable ranking at early stages of a user's lifetime in the system. These observations have implications for applications that rely on tag-vocabulary similarity (e.g., recommender systems): these applications can use only a sub sample of the entire user activity to estimate vocabulary similarity between users. Moreover, applications can aim to strike a balance between the accuracy of similarity estimates, the data volume used for estimation, and the freshness of the data. (Section 2.4)

These characteristics have practical implications for the design of mechanisms that rely on implicit user interactions such as collaborative search [15], spam detection [54] and recommendation [19, 35].

2.1 Related Work

This section positions this work among the related literature of two main topics: i) general studies about the characteristics of social tagging systems; and, ii) characterization of the evolution of tag vocabularies.

2.1.1 General Characterization Studies

Previous characterization studies focusing on tagging systems vary along three main aspects: *i*) the system analyzed; *ii*) the focus of the characterization (i.e., system-, tag-, item- or user-centric analysis); and, *iii*) the method of investigation - qualitative or quantitative research methods. Nevertheless, these works share the same intent: to characterize the usage patterns observed and gaining insight into the underlying processes that generate them. These works propose models that can be used to explain the observed characteristics of tagging activity such as the incentives behind tagging, the relative frequency of tags over time for a given item, the interval between tag assignments performed by users and the distributions of activity volume.

Hammond et al. [34] present, perhaps, the first study and discussion about the characteristics of social tagging, its potential, and the incentives behind tagging itself. The study comments on the features provided by different social tagging systems and discusses preliminary reasons that incentivize users to annotate and share content online. Following on the question of incentives, Ames et al. [3] study tagging in online social media websites by interviewing 13 users on the fundamental question of *why do people tag*? Based on user answers, the authors suggest that tagging serves to support content organization or to communicate aspects about the content. These actions can be either socially- or personally-driven. More recent studies have followed the analysis of incentives at a larger scale [90]. Our study supports and, more importantly, extends these result by performing a large-scale user behavior analysis (covering more than 700,000 users) in three tagging systems. Although, we do not focus on the question of incentives particularly, the quantitative analysis we present highlight and provide stronger evidence of existing incentives hypothesized by previous works.

One of the first works on the quantitative characterization of tagging systems is an item-centric characterization of *del.icio.us* that proposes the Eggenberger-Polya's urn model [23] as an explanation to the observed relative frequencies of tags applied to an item [29]. Addtionally, Cattuto et al. [12] show in a tag-centric characterization that the observed tag co-occurrence patterns in *del.icio.us* is well modeled by the Yule-Simon's stochastic process [85]. Similarly, Capocci et al. [11] models the tag interarrival time distribution to show that it follows a power-law. Using a different approach to characterize tagging activity, Chi and Mytkowicz [16] study the impact of user population growth in the efficiency of tags to retrieve items in *Delicious*. More recent works, however, focus on a characterization of social tagging systems that analyzes the impact of using tagging on external applications such as information retrieval and expert-generated content [31, 58, 61, 82].

Another stream of characterization studies focuses on user-centric analysis. Nov et al. [66] present a user-centric qualitative study on the motivations behind content tagging in *Flickr*, where they suggest that users tag content due to a mixture of individual like personal content organization, and social motivation such as to help others in finding photos from a particular place. In a previous study, we characterize the user-centric properties of tagging activity from two social bookmarking systems designed for academic citation management: *CiteULike* and *Bibsonomy*. The observations suggest that user activity across the system follows the Hoerl model [76].

The investigations presente in this chapter complement and extend these previous studies, as I study the characteristics of a combination of user-, item- and tag-centric tagging activity. Moreover, it explores different aspects of tagging activity, such as the levels of item re-tagging and tag reuse over time and the relationship between implicit and explicit user ties in tagging systems. By applying a quantitative approach on a broad population of users and multiple tagging systems, this study also offers new insights on user behavior that complement previous qualitative work by Ames and Naaman [3].

2.1.2 Evolution of Users' Tag Vocabularies

Tags represent to a certain extent the user perception or intended use of an item. It is natural, therefore, to assume that the set of tags (i.e, tag vocabulary) of a given user provides information about her topics of interest, which is useful to design other mechanisms that support efficient content usage such as recommender systems. Naturally, if inclusion of new tags or shifts in the tag usage frequency observed in a vocabulary are rare (i.e., if tag vocabularies are stable over time), a mechanism that relies on vocabulary snapshots can focus less on shifts of users preferences overtime when computing personalized predictions. Indeed, the results show that this is the case (Section 2.4).

Previous studies on the characterization of the evolution of tag vocabulary can be divided in two categories: first, studies that aim to quantify and model the growth of tag vocabularies at both the system- and user-level [12, 13]; and, second, studies that estimate shifts in the tag vocabularies over time such as evolution of the tag popularity distribution of item-level tag vocabularies [33], and the variation of tag usage frequency across manually predefined tag classes [29] (i.e., factual tags, subjective tags and personal tags) [83].

In summary, these previous studies show that: *i*) the system-level and user-level tag vocabulary growth is sublinear; *ii*) item-level tag popularity distribution converges to a power-law; and, *iii*) the usage frequency of tag categories shifts over time.

This study extends previous works by evaluating different facets of the vocabulary evolution. First, this work goes beyond the estimation of vocabulary growth, focusing on the evolution of tag usage frequency, as opposed to the frequency of tag categories. Second, it concentrates on individual, user-level tag vocabularies, as opposed to the item-level vocabularies; more precisely, the approach used makes no assumptions about the categories of tags that appear in the user tag vocabularies. Finally, it uses a different methodology to estimate the difference between tag vocabularies from different points in time.

2.2 Data Collection and Notation

This section describes the activity traces collected and analyzed in this study. Additionally, it also introduces the basic notation used in the rest of this chapter.

Table 2.1 presents a summary of the data sets used in this investigation. The *CiteULike*and *Connotea*data sets consist of all tag assignments since the creation of each system in late 2004 until January 2009. The *CiteULike*dataset is available directly from its website. For *Connotea*, I built a crawler that leverages *Connotea*'s API to collect tagging activity since December 2004 (no earlier activity was available). Finally, the *Delicious*dataset is available at the website of a previous study by Görlitz et al. [30][2].

[2] http://www.tagora-project.eu/

Table 2.1: Summary of data sets used in this study

	CiteULike	*Connotea*	*Delicious*
Activity Period	11/2004 – 01/2009	12/2004 – 01/2009	01/2003 – 12/2006
# Users	40,327	34,742	659,470
# Items	1,325,565	509,311	18,778,597
# Tags (distinct)	274,982	209,759	2,370,234
# Tag Assignments	4,835,488	1,671,194	140,126,555

Note that we do not have access to browsing or click traces. The traces analyzed in this work contain records that indicate when items are annotated with a given tag and who was the user, but the traces do not inform whether a tag is subsequently used by a user to navigate through the system, for example. The data sets are 'cleaned' to reduce sources of noise, such as the default tag 'no-tag' in *CiteULike*, tags composed only of symbols and other tags like the automatically generated 'bibtex-import', which are clear outliers in the popularity distribution.

Notation. The rest of this chapter uses the following notation. A tagging system is composed of a set of users, items and tags, respectively denoted by U, I, T. The tagging activity in the system is a set of tuples (u,i,w,t), where $u \in U$ is a user who tagged item $i \in I$ with tag $w \in T$ at time t. The activity of a user $u \in U$ can be characterized by A_u, I_u and T_u, which are respectively the set of tag assignments performed by u, the set of items annotated, and the vocabulary or set of tags used by u. The user's activity from the beginning of the trace up to a particular point in time is denoted by $A_u(t_0,t)$, $I_u(t_0,t)$ and $T_u(t_0,t)$, where t_0 and t are timestamps, t_0 represents the begin of the trace, and $t_0 \leq t$.

2.3 Tag Reuse and Item Re-Tagging

Let a new item (or tag) be an item (or tag) that has never been used in an annotation in the tagging system. If users introduce new items and tags frequently, efficiently harnessing information based on collective action is difficult, if not impossible.

This is so because in this case information about future user actions towards the annotation of an item or use of a tag is then hard to predict: prediction relies on the historical use of items and tags; new items or tags have no history in the system. Understanding the degree to which items are repeatedly tagged and tags reused can therefore help estimating the potential efficiency of techniques that rely on similarity of past user activity (e.g., recommender systems). To this end, this section addresses the following specific questions:

- **RQ 2.3.1** *What is the rate of repeated item annotation and tag reuse*? (Section 2.3.1)
- **RQ 2.3.2** *Is the flow of new incoming users a major factor in the observed rates of repeated item annotation*? (Section 2.3.2)
- **RQ 2.3.3** *Are the observed reuse patterns the result of a group of high-volume* power *users*? (Section 2.3.3)

The rest of this section first formalizes the metrics item re-tagging and tag reuse used to address these questions. Second, it characterizes the levels of item re-tagging and tag reuse as well as the level of activity generated by returning users. Finally, it discusses the implications of the usage characteristics discovered.

2.3.1 Levels of Item Re-tagging and Tag Reuse

An item is re-tagged (repeatedly tagged) if one or more users annotate it more than once (with the same or different tags). Similarly, a tag is reused if it appears in the trace more than once (for the same or different items) with different timestamps. The goal is to determine which portion of the activity falls in these categories.

Definition 1 *The level of item re-tagging during a time interval $[t_{f-1}, t_f)$ is the ratio between the number of items tagged during that interval that have also been tagged in the past $[t_0, t_f)$ to the total number of items tagged during the interval $[t_{f-1}, t_f)$, as expressed by Equation 2.1. Tag reuse, denoted by $tr(t_{f-1}, t_f)$, is similarly defined.*

$$ir(t_{f-1}, t_f) = \frac{|I(t_0, t_{f-1}) \cap I(t_{f-1}, t_f)|}{|I(t_{f-1}, t_f)|} \tag{2.1}$$

This definition is used to determine the aggregate level of item re-tagging and tag reuse in *CiteULike*, *Connotea* and *Delicious*. Table 2.2 presents the median daily item re-tagging and tag reuse over the entire traces (i.e., the time interval $[t_{f-1}, t_f)$ encompasses a day). The results show that *CiteULike* and *Connotea* have relatively low levels of item re-tagging while *del.icio.us* has a higher level of item re-tagging, yet all three systems present similarly high levels of tag reuse. One hypothesis is that the observed difference in item re-tagging between *Delicious* and their counterparts in *CiteULike* and *Connotea* is due to the type of content users bookmark in each system (with URLs of any type in the former, and academic literature in the latter).

Table 2.2: A summary of daily item re-tagging and tag reuse. The higher the score the more an item/tag is re-tagged/reused.

	Re-Tagged Items		Reuse Tags	
	Median	Std. Dev.	Median	Std. Dev.
CiteULike	0.15	0.07	0.84	0.12
Connotea	0.07	0.06	0.77	0.21
del.icio.us	0.45	0.17	0.86	0.07

To test whether these aggregate levels are a result of stable behavior over time, Figure 2.1 presents the moving average (with a window size of 30 days) of daily item re-tagging and tag reuse. Overall, these results show that all three systems go through a bootstrapping period, after which they stabilize, with the levels of item re-tagging and tag reuse stabilizing much sooner for *CiteULike* and *Connotea* than that for *del.icio.us*. However, the tag reuse levels have a similar evolution pattern in all three systems.

On the one hand, from the perspective of personal content management, the observed levels of item re-tagging and tag reuse, together with the much larger number of items than that of tags in these systems, suggest that users indeed exploit tags as an instrument to categorize items according to, for example, topics of

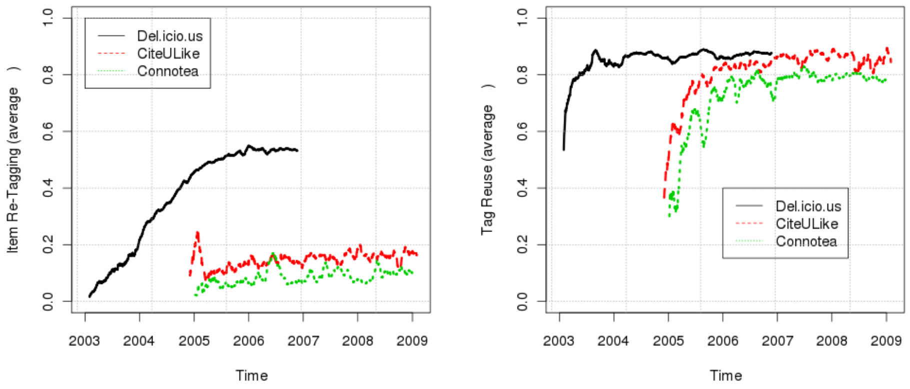


Figure 2.1: Daily item re-tagging (left) and tag reuse (right). The curves are smoothed by a moving average with window size $n = 30$

interest or intent of usage ('toread', 'towatch'). On the other hand, from the social (or collaborative) perspective, the relatively high level of tag reuse taken together with the low level of item reuse suggests that users may have common interest over some topics, but not necessarily over specific items. These quantitative results suggest that tags are used in the way previous exploratory qualitative study Ames and Naaman discusses [3].

A question that arises from the above observations is whether the levels of item re-tagging and tag reuse are generated by the same user or by different users. An inspection of the activity trace shows that virtually none of the item re-tagging events are produced by the user who originally introduced the item to the system: generally, users (in our trace) do not add new tags to describe the items they collected and annotated once.

As illustrated by Figure 2.2 (left), about 50% of tag reuse is self-reuse (i.e., the reuse of a tag by a user who already used it first). This level of tag self-reuse indicates that users will often tag multiple items with the same tag, a behavior consistent with the use of tagging for item categorization and personal content management, as discussed above. Additionally, the fact that half of the tag reuse

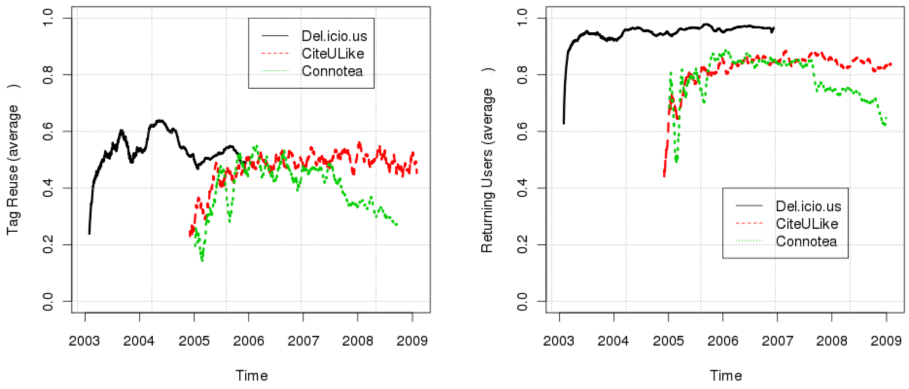


Figure 2.2: Self-tag reuse (left) and daily activity generated by returning users (right). The curves are smoothed by a moving average with window size $n = 30$

is not self-reuse reinforces the notion that users do share tags, which indicates potentially similar interests. Chapter 3 further investigates this social aspect of tag reuse by defining and evaluating *interest sharing* among users, as implied by the similarity between users' activity (i.e., tags and items).

2.3.2 New Incoming Users

To understand whether the observed low level of item re-tagging is due to a high rate of new users joining the community, it is necessary to estimate the levels of activity generated by returning users (as opposed to new users that join the community). Figure 2.2 (right) shows that, after a short bootstrap period, the level of tagging activity generated by returning users remains stable at about 80% over the rest of the trace for both *CiteULike* and *Connotea*. In *del.icio.us*, the percentage of activity represented by returning users is even higher, with above 95% of daily activity performed by returning users.

Thus, the low levels of item re-tagging are the outcome of new items being added by returning users, instead of a constant stream of new users joining the community.

2.3.3 The Influence of Power Users

Finally, this study looks into the influence of highly active users in the observed item re tagging and tag reuse levels. To this end, I conduct an experiment that consists of comparing the observed item re-tagging and tag reuse with and without the activity produced by such power users. This experiment assumes the *power users* as the top-1% most active users according to the number of annotations produced, and calculates item re-tagging and tag reuse as before.

The experiments test the hypothesis that the levels of item re-tagging and tag reuse are the same with and without the activity produced by these *power users*. To this end, I apply the Kolmogorov-Smirnov test (KS-test) on the two samples of activity (i.e., with and without the power users) with the null hypothesis that the item re-tagging and tag reuse observed in the two samples come from the same distribution (i.e, $H_0 =$ *the item re-tagging and tag reuse levels are equally distributed with and without the power users*).

At a confidence level of $99\%(\alpha = 0.01, p = 1 - \alpha)$, the null hypothesis can be rejected for all the systems, except the item re-tagging levels for *Delicious* (see the *p-values* in Table 2.3). This means that removing the activity produced by the power users leads to statistically different levels of item re-tagging and tag reuse as indicated by the D-statistic in Table 2.3 (i.e., the maximum difference between the two distributions) [87].

An explanation for the observations above is that *Delicious* is a system that focuses on social bookmarking of URLs of any type (as opposed to be restricted to scientific articles in *CiteULike* and *Connotea*), removing the top 1% most active users do not affect the observed levels of item re-tagging because some items will attract the attention of many other less active users. These users contribute, therefore, in large part for the observed levels of item re-tagging in *del.icio.us*.

Table 2.3: The statistical test results reject the hypothesis that the item re-tagging and tag reuse observations with and without the power users are equal. However, in most cases the difference has a small magnitude.

	Re-Tagged Items	
	D-Statistic	p-value <
CiteULike	0.03516	2.2×10^{-16}
Connotea	0.1889	2.2×10^{-16}
Delicious	0.0475	0.0768
	Reuse Tags	
	D-Statistic	p-value <
CiteULike	0.2858	2.2×10^{-16}
Connotea	0.2132	2.2×10^{-16}
Delicious	0.1371	3.23×10^{-16}

2.3.4 Summary and Implications

The observed user behavior impacts the efficiency of systems that rely on the inferred similarity among items, such as recommender systems. On the one hand, the relatively low level of item re-tagging suggests a highly sparse data set (i.e., attempting to connect users based on similar items will connect only few user pairs). A sparse data set poses challenges when designing recommender systems as they typically rely on the similarity of users based on their past activity to make recommendations.

On the other hand, the higher level of tag reuse confirms that analyzing tags has the potential to circumvent, or at least alleviate, the sparsity problem described above. The tags and users that relate to each item could not only serve to link items and build an item-to-item structure, but could also potentially provide semantic information about items. This information may help, for instance, to design better bibliography and citation management tools for the research community.

The results on analyzing the impact of power users in the observed levels of item re-tagging and tag reuse support two ideas: first, the notion that some users are instrumental on reducing the sparsity on tagging data sets (i.e., without power users, tags and items would be reused less, therefore potentially lesser items would be connected through tags and users). In fact, recommender systems benefit directly from the activity produced by such power users, as they can connect more items via repeated tag usage. Second, the role of power users differs from system to system, potentially due to effects of population size and diversity of interests. In the largest and most diverse system studied here, reuse is a result of the activity of less active users rather than only power users.

2.4 Temporal Dynamics of Users' Tag Vocabularies

The item re-tagging and tag reuse analysis presented in the previous section shows that users constantly produce new information in the system, by adding both new items to their libraries and tags to their vocabularies, though at different rates.

Although user tag vocabularies are constantly growing, it is unclear whether the growth rate is uniform over time. More importantly, vocabulary growth may or may not imply changes in the relative tag usage frequency by a given user. Changes in these frequency can indicate shifts in user interests over time.

To better understand these aspects of tagging activity, this section characterizes the temporal dynamics of user tag vocabularies. In particular, we study the rate of change of user vocabularies over time, as it quantifies the growth rate and changes in tag usage frequency for each user vocabulary. The following question guides this investigation:

- **RQ1.2** *What are the temporal dynamics of tag vocabularies?*?

To address this question, this section quantifies the evolution of user tag vocabularies by considering both their vocabulary growth and the tag usage frequency at different points in time. More specifically, the experiments first characterize the growth of user vocabularies, and, second, estimate the distance between tag vocabularies as expressed by the distance between snapshots of a user's vocabulary at various points in time and her final vocabulary. To take into account tag usage frequency the tags are ordered according to their frequency (i.e., the number of times the user annotated an item with the tag).

2.4.1 Methodology

Time is introduced in the definition of a user vocabulary by defining the tag vocabulary of a user $T_u(s, f)$ as the set of tags used within the tag assignment interval $[s, f]$. A particular case is $T_u(1, n)$ when 1 and n indicate the timestamps of the first and the last observed tagging assignment by user u, respectively. Thus $T_u(1, n) = T_u$ and represents the user's entire vocabulary.

Vocabulary growth. To analyze the vocabulary growth, it is necessary to track the distribution of growth rates across the user population for the duration of the traces. The goal is to understand whether the growth rate changes according to the user age. Therefore, the growth is measured by following ratio:

$$\frac{|T_u(1, k+1)| - |T_u(1, k)|}{|T_u(1, k+1)|} \tag{2.2}$$

where $k \in [1, n]$ for all users in the system (i.e., 1 and n represent the timestamp of the first and last tag assignments of a particular users, respectively).

Vocabulary change. To measure the rate of change in the content of the vocabularies, this investigation considers vocabularies as sets of tags ordered in decreasing order of usage frequency (i.e., number of times the tag was used to annotate any item), and apply a distance metric as follows.

In this context, the final tag vocabulary, $T_u(1, n)$ is taken as a reference point to study the evolution of tag vocabularies in terms of the usage frequency of individual tags. The rationale behind the choice of this reference is that according to the tag reuse results in Section 4, user tag vocabularies are constantly growing. Therefore, it is unlikely that splitting the activity trace into disjoint windows could help identifying meaningful evolution patterns. Instead, we trace the evolution of a user's tag vocabulary by comparing the distance of incremental snapshots to her final vocabulary. This way, it is possible to understand the rate of convergence of user vocabularies over time. The experiment consists of calculating the distance from the tag vocabularies $T_u(1, k)$ ($k \in [2, n]$), to the reference tag vocabulary $T_u(1, n)$.

A traditional metric to calculate the distance between two lists of ordered elements is the Kendall's τ distance [51], which considers the number of pairwise swaps of adjacent elements necessary to make the lists similarly ordered. However, Kendall's τ distance assumes that both lists are composed of the same elements. Since we are interested in the evolution of tag vocabularies over time, this assumption is not valid in our case: tag vocabularies are likely to contain different tags at different times due to the constant inclusion of new tags.

Therefore, we apply the *generalized Kendall's τ distance*, as defined by Fagin et al. [25], which relaxes the restriction mentioned above and accounts for elements that are present in one permutation, but are missing in the other. Similar to the original Kendall's τ distance, the generalized version of the metric counts the number of pairwise swaps of items necessary to make the lists similarly ordered. Additionally, the generalized version counts the absence of items via a parameter p. This parameter can be set between 0 and 1, which allows various levels of certainty about the order of absent items. For example, in the case that two items are missing from one list, but present on the other, setting $p = 0$ indicates that there are not enough information to decide whether the two items are in the same other or not. Conversely, setting $p = 1$ indicates that there is full information available to consider the absence as an increase in the distance between the lists. In the experiments that follow we use $p = 1$.

2.4.2 Results and Implications

Our analysis filters out users that had negligible activity considering only users with at least 10 annotations. This sample is responsible for approximately 93%, 61%, and 90% of the total system activity in terms of tag assignments in *CiteULike*, *Connotea*, and *Delicious*, respectively.

Vocabulary growth rate. Figure 2.3, Figure 2.4, and Figure 2.5 illustrates vocabulary growth rate across the user population in the three systems studied. The x-axis indicates categories of users according to their age (i.e., number of days since their first recorded tag assignment), while the y-axis indicates the growth rate relative to each user vocabulary. For each of the systems studied we present two plots: labeled 'median' and '90th percentile'. A point in the median plot indicates that 50

The results show that, for the duration of the traces analyzed, the median growth rate (e.g., Figure 2.3 – left) is relatively larger for older users. On the other hand, if we take the 90th percentile growth rate (e.g., Figure 2.3 – right), except the very young users, we observe that the rate is relatively the same for all age groups with a slightly smaller rate for users in the middle of the age spectrum. An important observation is that except for the growth rate of young vocabularies, the 90^{th} percentile reaches a maximum rate of 0.1. This means that for 90% of users, their vocabularies growth rate upper bound is 10%.

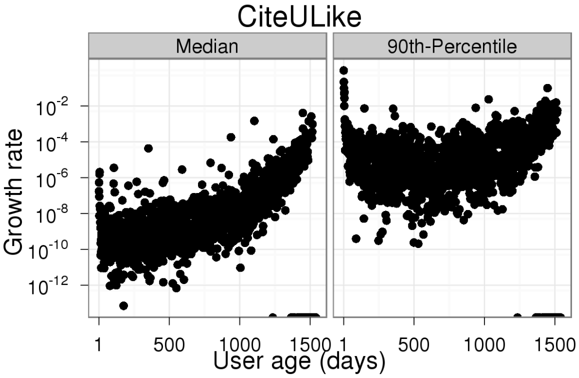


Figure 2.3: The vocabulary growth pattern in *CiteULike*

Vocabulary change. Figure 2.6, Figure 2.7, and Figure 2.8 changes the focus from growth rate to the rate of change in users' vocabularies. The figures present the rate of change in the contents of user vocabularies by taking into account the frequency of tags and calculating the distance between vocabulary snapshots. The results show that the distance from the vocabulary at earlier ages to its final state (i.e., Kendall-tau distance $t(T_u(1,k),T_u(1,n))$, where $k \in [2,n]$) decreases rapidly in the first 100 days for 50% of users.

These findings have direct consequences for the design of similarity-aware

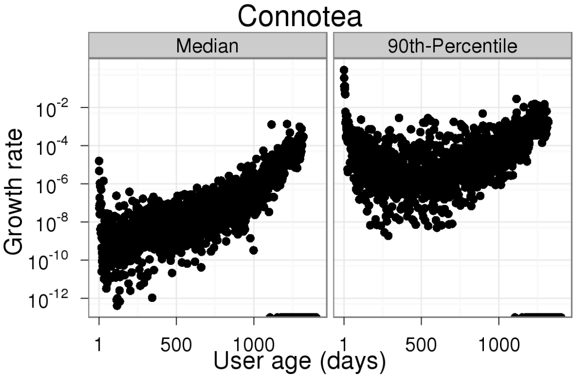


Figure 2.4: The vocabulary growth pattern in *Connotea*

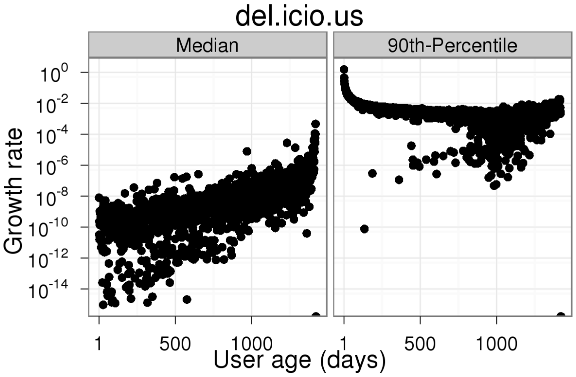


Figure 2.5: The vocabulary growth pattern in delicious

applications. The rapid convergence to the final vocabulary shown in Figure 2.6, Figure 2.7 and Figure 2.8 suggest that it is possible to obtain a relatively accurate approximation of a user's vocabulary based on her initial and limited tag assignment history, which leads to potential reduction in computation costs when attempting to estimate user similarity. In particular, users' vocabulary are used as the input for methods that quantify the value of tags, as presented in Chapter 5. Therefore, using part of the user vocabulary can be beneficial as one reduce the cost of computing tag values without compromising accuracy.

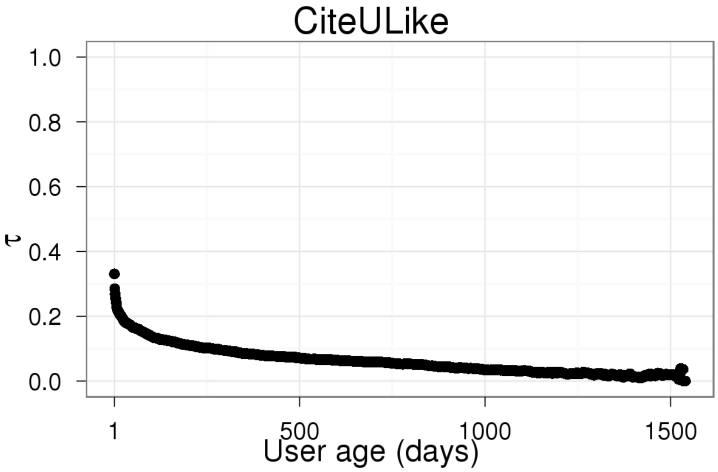


Figure 2.6: Rate of change in the tag usage frequency in the user vocabularies of *CiteULike*

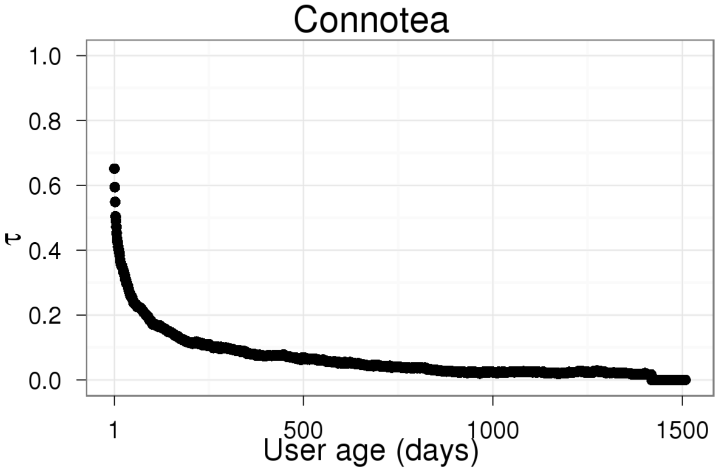


Figure 2.7: Rate of change in the tag usage frequency in the user vocabularies of *Connotea*

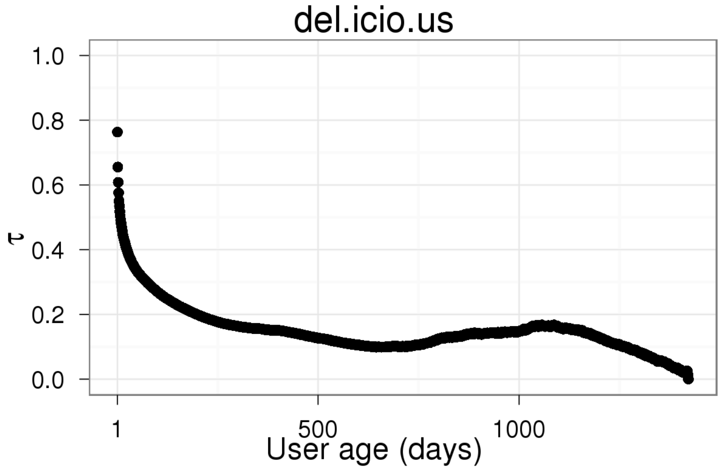


Figure 2.8: Rate of change in the tag usage frequency in the user vocabularies of *Delicious*

Chapter 3

Characterizing Social Aspects in Tagging Systems

Tagging systems are inherently social, as users can oftentimes annotate content shared by others or use others' annotations to discover new content of interest. Therefore, besides understanding the individual characteristics of user behaviour, it is also important to study this social dimension of social tagging. This chapter present results on the characterization of social user behaviour [1]. The focus lies on the characteristics of the social ties between users in these systems [64, 76, 78, 81].

The investigation of social ties between pairs of users focuses first on unveiling the characteristics of the implicit ties between users based on the similarity between their tagging activities. Additionally, this work explores the relationship between the strength of such implicit ties and those of more explicit social ties such as co-membership in discussion groups and semantic similarity of tag vocabularies. Studying the relationship among the implicit and explicit ties is relevant, as we test whether the implicit ties based on usage similarity provide information about the potential creation of explicit social ties and ultimately for collaboration. This characterization focuses on two main aspects of user social activity, as guided by the following particular questions:

[1] The results presented in this chapter appeared at the following references: [78, 81]

- **RQ2.1** *How is the strength of implicit user ties based on activity similarity distributed across the system?*
- **RQ2.2** *Are there relationships between implicit and explicit user ties?*

To address these questions, I applied a quantitative approach to characterize traces of activity collected from real social tagging systems. The main findings of this characterization study are:

- **RQ2.1**. *The observed levels of activity similarity between pairs of users are the result of shared interest as opposed to generated by chance.* The distributions of activity similarity strength deviate significantly from those produced by a Random Null Model (RNM) [71]. This suggests that the implicit ties between users, as defined by their activity similarity levels, capture latent information about user relationships that may offer support for optimizing system mechanisms (Section 3.2).
- **RQ2.2.** *The implicit social ties are related to explicit indicators of collaboration.* We show that user pairs that share interests over items (i.e., annotate the same items) have higher similarity regarding the groups they participate together and higher semantic similarity of their tag vocabularies (even after eliminating the portions of tagging activity that is related to the items they tag in common) (Section 3.3).

These characteristics have practical implications for the design of mechanisms that rely on implicit user interactions such as collaborative search [15], spam detection [54] and recommendation [56] as outlined in Section 3.2.4 and Section 3.3.3.

3.1 Related Work

This section contextualizes this work along the topic of graph-based approaches to study activity similarity among users.

An alternative way to characterize tagging systems is a graph-centric approach. Two users are connected by a weighted edge with strength proportional to the similarity between the tagging activities of these two users. In this study, this similarity is referred to as an implicit social tie between users. Note that other types of connections between users are possible. In particular, we refer to explicit social ties as explicit indicators of user collaboration, such as co-membership in discussion groups.

This approach has been used by Iamnitchi et al. [45, 46] to characterize scientific collaborations, the web, and peer-to-peer networks. The same model has been used by Li et al. [58] to target the problem of finding users with similar interests in online social networking sites. The authors use a *Delicious* data set and define links between users based on the similarity of their tags. Their conclusions support the intuition that tags accurately represent the content by showing that tags assigned to a URL match to a great extent the keywords that summarize that URL. Additionally, they design and evaluate a system that clusters users based on similar interests and identifies topics of interests in a tagging community.

Another focus of graph-centric characterizations is to determine structural features in the graph formed by connecting users, items and tags based on similarity. Hotho et al. [43] models a collaborative tagging system as a tripartite network (the network connects users, items and tags in a hypergraph) and design a ranking algorithm to enable search in social tagging systems. Using the same tripartite network model, Cattuto et al. [12] study *Bibsonomy* and show the existence of small-world patterns in such networks representing social tagging systems. Krause et al. [55] also explore the topology of a tagging system, but the one formed by item similarity, to compare the folksonomy inferred from search logs and tagging systems. Their results suggest that search keywords can be considered as tags to URLs. More recently, Kashoob et al. [50] characterizes and model the temporal evolution of sub-communities in social tagging systems by looking into the similarity between users vocabularies.

Our study differs from these previous investigations in three aspects: first, the characterization of tagging activity similarity between users focuses on the system-wide concentration and intensity of pairwise similarities, as opposed to the topological characteristics. Second, our methodology provide a principled way to test whether the user similarity observed in social tagging systems is the product of interest sharing among users or chance. Finally, we investigate possible correlations between the observed levels of activity similarity between users (i.e., the implicit social ties) and the external indicators of explicit collaboration (i.e., the explicit social ties) as co-membership to discussion groups and semantic similarity of tag vocabularies (Sections 3.2 and 3.3). We note that our methodology is inspired by a previous work by Reichardt and Bornholdt that studies the patterns of similarity of product preferences among buyers and sellers on eBay [71].

3.2 Interest Sharing

The analysis of item re-tagging and tag reuse in Section 2.3 suggests that the observed level of re-tagging is the result of different users interested in the same item and annotating it. We dub this similarity in item related activity *item-based interest sharing*. Similarly, we dub the similarity in tag related activity *tag-based interest sharing*. This section defines and characterizes pairwise interest sharing between users as implied by their annotation activity in *CiteULike*, *Connotea* and *Delicious*.

Analyzing interest sharing is relevant for information retrieval mechanisms such as search engines tailored for tagging systems [98, 101], which can exploit pairwise user similarity to estimate the relevance of query results. However, this section goes one step further and studies the system-wide characteristics of interest sharing and the implicit social structure that can be inferred from it. Moreover, the next section investigates the relationship between interest sharing (as inferred from activity similarity) and explicit indicators of collaboration such as co-membership in discussion groups and semantic similarity between tag vocabularies (Section 3.3).

In particular, this section focuses in particular on characterizing interest sharing distributions across the user-pairs in the system and addresses the following question:

- **RQ2.1.** *How is interest sharing distributed across the pairs of users in the system*?

3.2.1 Quantifying Activity Similarity

This study uses the Asymmetric Jaccard Similarity Index [47] to quantity similarity between the item (or tag-) sets of two users. Note that previous work (including ours) has used the Jaccard Index to quantify interest sharing: Stoyanovich et al. [89] used this index to model shared user interest in *Delicious* and to evaluate its efficiency in predicting future user behavior. Chi et al. [14] applied the symmetric index to determine the diversity of users and its impact in a social search setting.

More formally, the item-based interest-sharing metric is defined as follows (the tag-based version is defined similarly and denoted by w_T):

Definition 2 *The level of item-based interest sharing between two users, k and j, as perceived by k, is the ratio between the size of the intersection of the two item sets and the size of the item set of that user, where I_k is the set of items annotated by user k.*

$$w_I(k,j) = \frac{|I_k \cap I_j|}{|I_k|} \tag{3.1}$$

Equation 3.1 captures how much the interests of a user u_k match those of another user u_j, from the perspective of u_k. We opt for the asymmetric similarity index rather than the symmetric version (which uses the size of the union of the two sets as the denominator in Equation 3.1) to account for the observation that the distribution of item set sizes in our data is heavily skewed. As a result, the situation where a user has a small item set contained in another user's much larger item set happens often. In such cases, the symmetric index would define that there is little similarity between interests, while the asymmetric index accurately reflects that, from the standpoint of the user with smaller item set, there is a large overlap of interests. From the perspective of the user with a large item set, however, only a small part of his interests intersect with those of the other user.

3.2.2 How is Interest Sharing Distributed across the System?

This section presents the distribution of pairwise interest sharing in *CiteULike*, *Connotea* and *Delicious*. The first observation is that approximately 99.9% of user pairs in *CiteULike* and *Delicious* share no interest over items (i.e., $w_I(k, j) = 0$). In *Connotea*, the percentage is virtually the same: 99.8%. For the tag-based interest sharing, the percentage of user pairs with no tag-based shared interest (i.e., $w_T(k, j) = 0$) is slightly lower: 83.8%, 95.8% and 99.7% for *CiteULike*, *Connotea* and *Delicious*, respectively. Such sparsity in the user similarity supports the conjecture that users are drawn to tagging systems primarily by their personal content management needs, as opposed to the desire of collaborating with others (Section 4.4.1 discusses further the qualitative aspects of tag production).

The rest of this section focuses on the remaining user pairs, that is, those user pairs that have shared interest either over items or tags. To characterize these user pairs, we determine the cumulative distribution function (CDF) of item- and tag-based interest sharing for these sets of user pairs in all three systems.

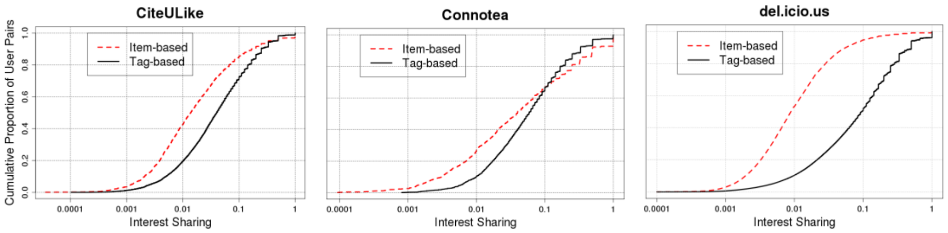


Figure 3.1: Distributions for item- and tag-based interest sharing (for pairs of users with non-zero sharing) in the studied systems

Figure 3.1 shows that, in all three systems, the typical intensity of tag-based interest sharing is higher than its item-based counterpart. This is not surprising: after all, all three systems include two to three times more items than tags. However, there is qualitative difference across systems with respect the concentration of item-based and tag-based interest sharing levels, with *Delicious* showing a much wider gap between the distributions.

The difference between the levels of item- and tag-based interest sharing suggests the existence of latent organization among users as reflected by their fields of interest. We hypothesize that this observation is due to a large number of user pairs that have similar tag vocabularies regarding high-level topics (e.g., computer networks), but have diverging interests in specific sub-topics (e.g., internet routing versus firewall traversal techniques), which could explain the relatively lower item-based interest sharing compared to the observed tag-based interest sharing.

Finally, to provide a better perspective in the tag-based interest sharing levels, we compare the observed values to that of controlled studies on the vocabulary of users describing computer commands [28]. The tag-based interest sharing level, as observed in Figure 3.1 is approximately 0.2 (or less) for 80% of the user pairs that have some interest sharing, while Furnas et al. [28] show that in an experiment where participants are instructed to provide a word to name a command based on its description such that it is an intuitive name and more likely to be understood by other people, the ratio of agreement between two participants is in the interval $[0.1, 0.2]$ (i.e., number of times two participants use the same word divided by the total number of participant pairs).

These observations suggest that the tag-based interest sharing is due to conscious choice of terms from vocabularies that are shared among users, rather than by chance. The next section looks more closely into this aspect by constructing a baseline to compare the observed interest sharing levels to that of a random null model.

3.2.3 Comparing to a Baseline

The goal of this section is to better understand the interest sharing levels we observe. In particular, we focus on the following high-level question:

- **RQ.2.2**. *Do the interest sharing distributions we observe differ significantly from those produced by random tagging behavior?*

For this investigation, we compare the observed interest sharing distribution to that obtained in a *system with users that have an identical volume of activity and the same user-level popularity distributions for items or tags, but do not act according to their personal interests*. Instead, in the random null model (RNM) [71], the chance that a user is interested in an item or tag is simply that item or tag's popularity in the user's vocabulary.

The reason to perform this experiment is the following: we aim to validate the intuition that the interest sharing metric distils useful user behavior information. If the interest-sharing levels we observe in the three real systems at hand are more concentrated than those generated by the RNM, then interest sharing metric captures relevant information about similarity of user preferences, rather than simply coincidence in the tagging activity.

To reiterate, the random null model (RNM) is produced by emulating a tagging system activity that preserves the main macro-characteristics of the real systems we explore (such as the number of items, tags, and users, as well as item and tag popularity, and user activity distributions), but where users make random tag assignments. As such, random assignments are used here as the opposite of interest-driven assignments.

To test this hypothesis, the experiment compares the two sets of data (real and RNM-generated) in terms of the numbers of user pairs with non-zero interest sharing and the interest-sharing intensity distribution. Because of its probabilistic nature, we use the RNM to generate five synthetic traces corresponding to each of the real systems we analyze. For the rest of this section, the RNM results represent averages over the five RNM traces for each system. We confirmed that the five synthetic traces represent a large enough sample to guarantee a narrow 95% confidence interval for the average interest sharing observed from the RNM simulations.

The data analysis presented in this section confirms that interest sharing deviates significantly from that generated by random behavior in two important respects.

First, interest sharing (and, consequently, the similarity between users) is more concentrated in the real systems than in the corresponding simulated RNM. More specifically, the number of user pairs that share some item-based interest (i.e., $w_I(k, j) > 0$) is approximately three times smaller in the real systems than in the RNM-generated ones. Tag-based interest sharing follows a similar trend.

Second, interest sharing distribution deviates significantly from that produced by a RNM. We compare the cumulative distribution function (CDF) for the interest sharing intensity for the user-pairs that have some shared interest (i.e., $w(k, j) > 0$). Figure 3.2 presents the Q-Q plots that directly compare the quantiles of the distributions of interest-sharing levels derived from the actual trace and those derived from the simulated RNM. A deviation from the diagonal indicates a difference between these distributions: The higher the points are above the diagonal, the larger the difference between the observed interest-sharing levels and those generated by the RNM.

Note that the only interest-sharing distribution that is close to the one produced by the RNM is for *Connotea*'s tag-based interest sharing (Figure 3.2). However, there is still a significant deviation from randomness: the real activity trace leads to three times fewer user-pairs that share interest than the corresponding RNM.

3.2.4 Summary and Implications

This section provides a metric to estimate pairwise interest sharing between users, offers a characterization of interest-sharing levels in *CiteULike*and *Connotea*; and investigates whether the observed interest sharing in these systems deviates from that produced by chance, given the amount of activity users had. Such reference

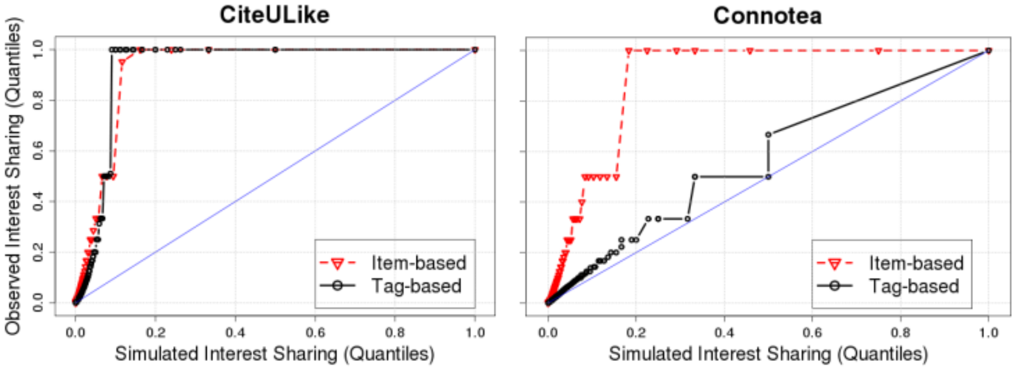


Figure 3.2: Q-Q plots that compare the interest sharing distributions for the observed vs. simulated (i.e., the RNM model) for *CiteULike* (left) and *Connotea* (right)

is given by a random null model (RNM) that preserves the macro characteristics of the systems we investigate, but uses random tag assignments.

The comparison highlights two main characteristics of the interest sharing: first, interest sharing is significantly more concentrated in the real traces than in the RNM-generated activity: in quantitative terms, three times fewer user pairs share interests in the real traces. Second, most of the time, for the user pairs that have non-zero interest sharing the observed interest-sharing intensity is significantly higher in each real system than in its RNM equivalent.

A conjecture to explain these observations is as follows. Let us consider that the set of tags that can be assigned to an item is largely limited by the set of topics that item is related to. In this case, intuitively, the probability of choosing a tag is conditional to the set of topics the item is related to. At one extreme, the maximum diversity of topics occurs when there is a one-to-one mapping between topics and tags, that is, when each tag introduces a different topic. The RMN simulates the other extreme, a single topic that encompasses all tags in the system.

However, in real systems, the interests for each individual user are limited to a finite set of topics, which is likely to determine their tag vocabulary. This leads to a concentration of interest sharing, as implied by the tag similarity, on few user pairs, yet at higher intensity than that produced by the RNM.

Finally, and most importantly, the divergence between the observed and the RNM-generated interest sharing distributions shows that activity similarity, our metric to quantify interest sharing intensity, embeds information about user self-organization according to their preferences. This information, in turn, could be exploited by mechanisms that rely on implicit relationships between users. The next section seeks evidence about the existence of such information by analyzing the relationship implicit user ties, as inferred from the similarity between users' activity, and their explicit social ties, as represented by co membership in discussion groups or semantic similarity between tag vocabularies.

3.3 Shared Interest and Indicators of Collaboration

The previous section characterizes interest sharing across all user pairs in each system and suggests that it encodes information about user behavior, as its distribution deviates significantly from that produced by a random null model.

This section complements this characterization and evaluates whether the implicit user relationships that can be derived from high levels of interest sharing correlate with explicit online social behavior. More specifically, this section addresses the following question:

- **RQ2.2** *Are there correlations between interest sharing and explicit indicators of social behavior*?

Before starting the analysis, it is important to mention that the number of externally observable elements of user behavior to which we have access is limited by the design of the tagging systems themselves (e.g., the tagging systems collect limited information on user attributes) and by our limited access to data (e.g., we do not have access to browsing traces or search logs).

One *CiteULike* feature, however, is useful for this analysis: *CiteULike* allows users to explicitly declare membership to groups and to share items among a selected subset of co-members – an explicit indicator of user collaboration in the system. Thus, this feature enables an investigation about the relationship between interest sharing and group co-membership (which we assume to indicate collaboration). Note that a similar experiment could be performed using the explicit friendship links in *Delicious*, for example. However, this data is not available to our study.

Along the same lines, we use a second external signal: semantic similarity between tag vocabularies. More specifically, we test the hypothesis that item-based interest sharing relates to semantic similarity between user vocabularies. The underlying assumption here is that users who (have the potential to) collaborate employ semantically similar vocabularies.

This section presents the methodology and the results of these two experiments that mine the relationship between interest sharing and indicators of collaboration. In brief, our conclusions are:

- User pairs with positive item-based interest sharing have a much higher similarity in terms of group co-membership and semantic tag vocabulary, than users who have no interest sharing.
- On the other side, we find no correlation between the intensity of the interest sharing and the collaboration levels as implied by group co-membership or vocabulary similarity.

3.3.1 Group Membership

In *CiteULike*, approximately 11% of users declare membership to one or more groups. While the percentage may seem small, they are the most active users: these users generate 65% of tag assignments, and introduce 51% of items and 50% of tags. For this section we limit our analysis to the user pairs for which both users are members of at least one group. Also, the analysis focuses on groups that have two or more users (about 50% of all groups) as groups with only one user are obviously not representative of potential collaboration.

The goal is to explore the possible relationship between item-based interest sharing and co-membership in one or more groups. Let H_u be the set of groups in which the user u participates. We determine the group-based similarity $w_H(u,v)$ between two users u and v using the asymmetric Jaccard index, similar to the item-based definition in Eq. 3.1, but considering the sets of groups users participate in. Based on this similarity definition, we study whether the intensity of item-based interest sharing between two users with non-zero interest sharing (i.e., $w_I(u,v) > 0$) correlates with group membership similarity.

The experiments show no correlation between $w_I(u,v)$ – the item-based interest sharing – and $w_H(u,v)$ – the group-based similarity. More precisely, Pearson's correlation coefficient is approximately 0.12, and Kendall's τ is about 0.05. This is surprising as one would expect that being part of the same discussion groups is a good predictor to the intensity in which users share interest over items. Therefore, we look into these correlations in more detail.

To put these correlation results in perspective, we look at group similarity for two distinct groups of user pairs: those with no item-based interest sharing ($w_I(u,v) = 0$) and those with some interest sharing ($w_I(u,v) > 0$). We observe that, although the group information is relatively sparse, pairs of users with positive interest sharing are more likely to be members of the same group than the user pairs where $w_I(u,v) = 0$. In particular, 4% of the user pairs with $w_I(u,v) > 0$ have $w_H(u,v) > 0.2$, while twenty times fewer user pairs with $w_I(u,v) = 0$ have $w_H(u,v) > 0.2$.

These observations suggest that activity similarity (defined according to Eq. 3.1) is a necessary, but not sufficient condition for higher-level collaboration, such as participation in the same discussion groups. Although users share interest over items, and may implicitly benefit from each other tagging activity (e.g., using one another's tags to navigate the system), this may not directly lead to users actively engaging in explicit collaborative behavior. Conversely, the lack of interest sharing strongly suggests a lack of collaborative behavior.

3.3.2 Semantic Similarity of Tag Vocabularies

This section complements the previous analysis on the relationship between item-based interest sharing and collaboration indicators via group co-membership. It investigates the potential relation between item-based interest sharing of a pair of users and the semantic similarity between their tag vocabularies, that is, the set of tags each has applied to items in its library. Since, through this experiment we aim to understand the potential for user collaboration through similar vocabularies, when comparing vocabularies for a user pair, we exclude the tags applied to the items the two users have tagged in common – a these tags have a likely high similarity.

The rest of this section is organized as follows: it presents the metric used to estimate the semantic similarity of two tag vocabularies; discusses methodological issues; and, finally, presents the evaluation results.

Estimating semantic similarity: This experiment uses the lexical database *WordNet* to estimate the semantic similarity between individual tags. *WordNet* consists of a set of hierarchical trees representing semantic relations between word senses such as synonymy (the same or similar meaning) and hypernymy/hyponymy (one term is a more general sense of the other). Different methods have been implemented to quantify semantic similarity using *WordNet*. In particular, WordNet::Similarity – a Perl module – provides a set of semantic similarity measures [68].

The experiments use the Leacock-Chodorow similarity metric [10], as previous experiments, based on human judgments, suggest that it best captures the human perception of semantic similarity. The metric is derived from the negative log of the path length between two word senses in the *WordNet* ”is-a” hierarchy, and is only usable between word pairs where each has at least one noun sense.

Additionally, we explore a method to extend coverage to a larger subset of users’ tag vocabularies, with an approach that builds on the YAGO ontology, developed and described by Suchanek et al. [91, 92]. YAGO (”Yet Another Great

Table 3.1: The share of tagging activity captured by the tag vocabularies in *CiteULike* and *Connotea* that is found in *WordNet* and WorldNet combined with YAGO lexical databases. As we use an anonymous version of the *del.icio.us* dataset, with all the users, items, and tags identified by numbers, this precluded us to perform the same analysis using *WordNet* and YAGO for *del.icio.us*.

	WordNet only	*WordNet* + YAGO
CiteULike	62.1%	79.5%
Connotea	51.3%	65.3%
Combined	57.4%	73.4%

Ontology") is built from the entries in Wikipedia [2], a collaborative online encyclopedia. The standardized formatting of Wikipedia makes it possible for information to be automatically extracted from the work of thousands of individual contributors and used as the raw material of a generalized ontology. The primary content of the YAGO ontology is a set of fact tables consisting of bilateral relations between entities, such as "bornIn", a table of relations between persons and their birthplaces. Five of the relations are of particular interest to us because they contain links between entities mentioned in Wikipedia and terms found in *WordNet*. In this way, we are able to identify some tags as probable personal, collective, or place names, and use the *WordNet* links from YAGO to map these on to a set of corresponding *WordNet* terms.

A merged tag vocabulary that combines tags from *CiteULike* and *Connotea* datasets show that a little over 13% of the tags had direct matches in *WordNet*. By adding the tags matched through comparison with YAGO's *WordNet* links, this was increased to 28.6% of unique tags applied by users of both systems. Note, however, that these tags cover up to 75% of the tagging activity in the two systems, as shown in Table 3.1.

In order to match tags gathered from the two systems with corresponding entities in YAGO, all non-ASCII characters, such as accented letters, are replaced

[2] http://wikipedia.org

by their nearest ASCII equivalents; also, the experiment removes all characters other than letters and numerals, and reduced all the YAGO entities to lower case (tags from both systems being already reduced to lower case). Finally, partial matches are allowed, but to consider the partial match it is required that the end of a tag correspond to a word boundary in the YAGO entity or vice-versa. This procedure enable the construction of a mapping between about 58,600 tags from the merged vocabulary and 57,900 distinct *WordNet* senses, with most tags matching multiple *WordNet* senses. Given that the addition of *WordNet* terms identified by mapping through YAGO effectively increases the total depth of the tree being considered, the Leacock-Chodorow algorithm required that we adjust all tag pair similarity scores accordingly in order to fairly compare the *WordNet*-only and *WordNet*+YAGO scores. The maximum possible similarity with *WordNet* alone is $\log(1/40)$ or 3.689; whereas with *WordNet* + YAGO it is $\log(1/42)$ or 3.738.

The similarity $sim(t_1,t_2)$ between two tags (t_1,t_2) is defined as the maximum Leacock-Chodorow similarity between every available noun sense of t_1 and t_2. Thus, the semantic similarity between the tag vocabularies T_u and T_v of two users, u and v, as perceived by u, is denoted by $s(u,v)$, and determined by the ratio between the sum over the pairwise tag similarities and the size of u's vocabulary, as expressed by Eq. 3.2 below.

$$sim(u,v) = \frac{\sum_{t_1 \in T_u, t_2 \in T_v} sim(t_1,t_2)}{|T_u|} \tag{3.2}$$

We then calculate the corresponding value of $s(v,u)$ by reversing the u and v terms in Eq. 3.2 and record the smaller of the two – i.e. $min(s(u,v),s(v,u))$ – as the undirected tag vocabulary similarity between the two users u and v. We note that this metric is based on the Modified Hausdorff Distance (MHD) [22].

Methodological issues. There are three practical issues regarding our experimental design that deserve a note. First, to avoid bias, if two users assigned the same tags to the same item, we omit these tags from their vocabularies, before determining the aggregate similarity. By eliminating from vocabularies the tags that have been used on exactly the same items, we eliminate the tags on which

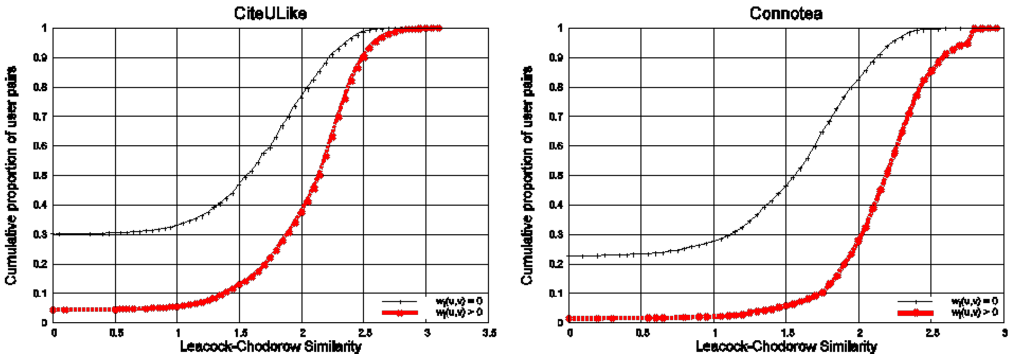


Figure 3.3: CDFs of tag vocabulary similarity for user pairs with positive (bottom curve) and zero (top curve) activity similarity (as measured by the item-based interest sharing defined in Eq. 3.1). *CiteULike* (left); *Connotea* (right)

the two users have most likely already converged. We look only at the remaining parts of the vocabularies where convergence is not apparent. Second, the Leacock-Chodorow similarity metric only considers words that have noun senses in *WordNet*, because it is calculated from paths through the ”is-a” hierarchy, only defined for nouns. Tags in both systems considered may include words or phrases from any language, abbreviations, or even arbitrary strings invented by the user, while *WordNet* consists mainly of common English words. A third methodological issue was that matching tags to YAGO entries, in some cases, returned an unmanageably large set of distinct *WordNet* senses. We accordingly eliminated those tags that were above the 99th percentile in distinct *WordNet* senses matched, which were those returning more than 167 distinct senses.

Results. We use sampling to test, in both *CiteULike* and *Connotea*, whether there is a significant difference in tag vocabulary similarity between two sets of user pairs: one where all users have no item-based interest sharing and one with positive item-based interest sharing (we sample each group with $n = 4000$ pairs). This analysis shows that the vocabularies of user pairs with interest sharing are significantly more similar than those of user pairs with no interest sharing (Figure 3.3). The median vocabulary similarity for user pairs with positive interest sharing $\mu_c = 2.112$ (±0.02, 99% c.i.) is about 1.6 times that of user pairs with no interest sharing $\mu_u = 1.308$ (±0.04, 99% c.i.). This salient difference in the vocabulary similarity suggests that the item-based interest sharing embeds information about the ”language” shared by the users to describe the items they are interested in.

3.3.3 Summary and Implications

This section takes a first step towards understanding the relationship between the implicit user ties, as inferred from pairwise interest sharing, and their explicit social ties. First, we look at correlations between the item-based interest sharing and the group-based similarity. The observations indicate that although the intensity of item-based activity interest sharing does not correlate with explicit collaborative behavior, as implied by group co-membership, user pairs with some interest sharing are more than one order of magnitude more likely to participate in similar groups.

Second, we evaluate the relationship between item-based interest similarity and the semantic similarity of tag vocabularies. We discover that, although the two do not yield a Pearson’s correlation, item-based interest similarity does embed information about the expected semantic similarity between user vocabularies.

These results have implications on the design of mechanisms that aim to predict collaborative behavior, as these mechanisms could exploit item-based similarity to set expectations about group-based and vocabulary-based similarity. Moreover, assuming that the tagging activity characteristics of spammers differ from legitimate users, one could use deviations from observed relationship between item-based similarity and the two indicators of collaborative behavior presented here to detect malicious user behavior.

Chapter 4

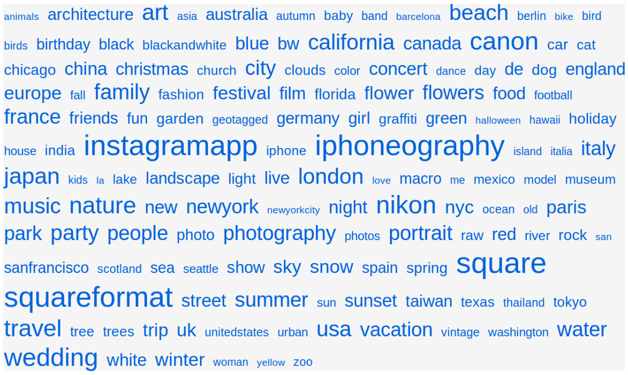
Understanding Users' Perception of Tag Value

The first part of this thesis (Chapter 2 and Chapter 3) presents a quantitative characterization of tag production with the goal to understand how users' individually produce tags and socially interact (share interest and collaborate). This chapter moves towards a characterization of users' perception of tag value [1]. In particular, it considers users' perception of value when both producing and using tags for particular tasks. The goal of this chapter is to understand the qualitative aspects users take into account when both producing tags (i.e., annotating content) and using tags in *exploratory search tasks*.

In exploratory search tasks, information seekers (i.e., users who are looking to satisfy an information need) navigate the set of items by using tag clouds, as opposed to traditional keyword search. Users tend to prefer tag-based navigation when they are exploring a topic and want to retrieve a set of related items, as opposed to the single most relevant item [86]. *Tag clouds* (or similar user interface artifacts) are the default interaction mode provided by systems like *Delicious*, StackOverflow, or MrTaggy [49]. Figure 4.1 illustrates what a tag cloud typically look like. Tag clouds are generally initialized with the set of most popular tags.

[1] The results presented in this chapter appeared at the following references: [80]

Information seekers start the navigation by entering a tag-query (typing or clicking). The system, in turn, retrieves items that are annotated with that tag-query and related tags (e.g., in the form of a tag cloud). The navigation continues further if the user selects one of the available tags presented by the system. The search result at each navigation step is generally composed of items annotated by all the tags selected by the user. In this sense, we assume that the tagging system provides AND-semantics [6].



animals architecture art asia australia autumn baby band barcelona beach berlin bike bird
birds birthday black blackandwhite blue bw california canada canon car cat
chicago china christmas church city clouds color concert dance day de dog england
europe fall family fashion festival film florida flower flowers food football
france friends fun garden geotagged germany girl graffiti green halloween hawaii holiday
house india instagramapp iphone iphoneography island italia italy
japan kids la lake landscape light live london love macro me mexico model museum
music nature new newyork newyorkcity night nikon nyc ocean old paris
park party people photo photography photos portrait raw red river rock san
sanfrancisco scotland sea seattle show sky snow spain spring square
squareformat street summer sun sunset taiwan texas thailand tokyo
travel tree trees trip uk unitedstates urban usa vacation vintage washington water
wedding white winter woman yellow zoo

Figure 4.1: Example of a tag cloud extracted from the most popular tags in *Flickr*. Information seekers interact with the tag clouds by clicking on each term to retrieve items that are annotated with that specific tag. The tag cloud is reconstructed at each step.

The investigation presented in this chapter addresses the following research question:

- **RQ3.** *What are the aspects that influence users' perception of tag value for exploratory search?*

To address this question, this study uses qualitative research methods. In particular, in-depth contextual interviews help collecting the data, while the analysis cycle resorts to grounded theory methods [39]. This leads to a characterization of aspects that influence users' perception of tag value for exploratory search and when producing tags to annotate content. The rest of this chapter focuses on the former issue, while one of the latter aspects is further explored in Chapter 6.

In summary, this chapter presents two major contributions:

- Present findings about aspects of users' production of tags that contribute to solidify the existing body of research on the motivation behind tagging. Moreover, it reveals that sometimes there is a disconnect between the motivations behind producing tags and the aspects that makes a tag valuable to users when solving exploratory search tasks (Section 4.4.1).
- A qualitative characterization of users' perception of tag value in the context of exploratory search; based on the qualitative analysis of 9 contextual interviews of social tagging users, we find that the two most salient aspects that influence users' perception of tag value are: ability to retrieve relevant content items and ability to reduce the search space (Section 4.4).

The next section provides background and positions the investigation in this chapter among the related literature.

4.1 Related Work

In a nutshell, this work differs from previous efforts in two main aspects. First, it is motivated by the view that social tagging systems are inherently online peer production systems [8]. Thus to improve the quality of user contributions, it is necessary to first quantify their value, so that one can then think of designing incentives for the production of high quality content. Second, this research focuses both on characterizing users' perception of tag value, and on the design and analysis of a method to assess tag value in practice (as opposed to only studying the impact of tags in other information retrieval tasks such as recommendation [3, 9, 26, 90]).

This section starts by positioning the work in this chapter among the related studies on characterizing users' motivation behind tagging (Section 4.1.1). Next, it discusses previous studies on the economics of information that provides a background to understand the perceived value of peer-produced information (Section 4.1.2).

4.1.1 *Why Do We Tag?*

Hammond et al. [34] provide, perhaps, the rst study that discusses the characteristics of social tagging, its potential, and the motivations users have to produce tags. The study comments on the features provided by different social tagging systems, and discusses preliminary reasons that incentivize users to annotate and share content online. Marlow et al. [63] discuss the properties of several tagging systems while pointing out their similarities and differences. Additionally, the authors conjecture the motivations that can potentially drive the production of tags in these systems. Ames and Naaman [3] go deeper on the study of motivations behind tagging and investigate why people tag in mobile (i.e., ZoneTag) and web applications (i.e., *Flickr*). They interviewed 13 users to address the question: *Why do people tag?* Their findings indicate that there are both personal and social motivations behind tagging. Moreover, the study builds a taxonomy for the motivations behind tagging in these systems along two dimensions: sociality and function. More recent studies have extended the analysis of motivations at a larger scale [90].

This work differs from these previous studies as it concentrates on understanding the use of tags to engage in exploratory search tasks (e.g., exploring the set of items available in a social bookmarking tool), as opposed to focusing on the motivations behind tagging (i.e., the production of tags).

4.1.2 Economics of Information

The value of information in market settings is contextual [88], as it requires one to make use of it to assess its expected value. Hischleifer [42] adds to the study of characteristics of information goods by enumerating and discussing a set of economically significant information attributes that can influence its perceived value, namely: Certainty, Diffusion, Applicability, Content (environmental vs. behavioral), and Decision-relevance; as described below.

Certainty – the value of information goods depends on the amount of certainty it provides about the outcome of a particular process. For example, an annotation that increases the probability of finding an item of interest to a user is more valuable than an annotation that retrieves items of marginal interest to the user.

Diffusion – the availability of information goods across the user population may affect their value, as few users may have the privilege to possess that information. In tagging systems, one may think of particular items or annotations that are kept private.

Applicability – an information good can be of general or particular applicability or interest. Indeed, a tag or item may serve a general audience or only a small fraction of the user population. For instance, tags can be general enough (e.g., networks) to be of interest to several sub-communities of users that use it to retrieve relevant content. Conversely, other tags (e.g., Agneta[2]) are only applicable in a more restricted subset of the user population.

Content – naturally, the value of information may be affected by the characteristics of its contents. Hirschleifer points out to common subclasses of this aspect in market settings, where it distinguishes information about the environment from information about the behavior of other individuals in the market. In tagging systems, the content aspect of a peer-produced information can map to the semantic of tags, for example. For example, a tag may reveal information about how the user intends to use the annotated content (e.g. 'to-read'). Additionally, a tag may contain information about topics of interest for a user.

Decision-relevance – this dimension captures how important is the information in the context of a decision problem.

As noted by Bates [74], the five aspects provide an idea on *what* factors influence the value of tags and items, but it shed little light on *how* to determine their value. Arrow[**?**] and Stigler [88] seems to provide a way out by adding to the five aspects described above the observation that the value of information is determined from its use.

One may expect that these attributes also influence the perceived value of peer-produced information such as tags. However, while attributes that influence infor-

[2] A character in Pedro Juan Gutierrez's *Tropical Animal* novel.

mation value have been investigated and discussed in market contexts, it is unclear what role these attributes play, if any, in the context of peer-production systems, in general, and in social tagging systems, in particular.

Stigler states the value of information goods can only be assessed by its use [88]. Repo [73] goes further with this statement and discusses two major approaches on assessing value of information: value-in-use and exchange value.

Our study uses the same notion of a multidimensional value concept as discussed by Hischleifer [42] and Repo [73]. However, we inquire further on the human perception by performing interviews with users to understand their perception of value of peer-produced information (which departs from the type of information focused on in previous work). More precisely, we investigate what aspects users take into account when choosing tags in the context of exploratory search tasks. Understanding the value of information in online peer production systems, as perceived by information seekers, extends the existing body of knowledge on the value of information in markets' context and in the design of information systems discussed in the next section.

4.1.3 Perceptions of Information Value in System Design

Another context where the perception of information value plays a role is on the design of social information-sharing systems. In a recent study, Lampe et al. [57] investigate the users' perception of Facebook's value as an information source by conducting a survey with non-faculty staff at Michigan State University. Their analysis shows that Facebook users are not likely to engage in information seeking with their Facebook network. However, users who do engage in information seeking show common characteristics: they tend to be female, younger, and have more total and actual friends on Facebook than those who do not engage in information seeking. Similarly, André et al. [4] investigate the perception of value of tweet's content. The authors show that while 35% of tweets are rated by users as worth reading, 25% are marked otherwise (not worth reading). Their analysis also shows that the tweets that are considered valuable are those that provide information about a topic of interest or have some humor taste.

Our study differs from these previous work as it focus on the instrumental value of tags (peer-produced information) for a particular application exploratory search as opposed to the conceptual value of information of expert-produced information or the value assessment of a platform (e.g., online social network or particular search engine) as a source for information.

4.2 Methodology

This section presents a brief description of the methodology adopted for the qualitative study (i.e., recruiting, data collection, and analysis methods). Figure 4.2 illustrates the qualitative research cycle used in this work. It consists of an iterative process that starts by defining a set of research questions, recruiting participants, performing contextual interviews, and iteratively refining the data collection procedures based on the ongoing analysis of data.

Participants. The target population for this experiment was any Internet user who is familiar with search and navigation tasks in social tagging systems. The recruiting method used a combination of advertising via email and snowball recruiting techniques (i.e., where a participant suggests others who may qualify for the study) [39]. New participants were continuously recruited until a level of diversity among participants and saturation in the data collection was observed. Participants were asked to complete a background and demographics questionnaire [3]. We recruited 12 participants. The first two interviews were used as pilots and were not used for the final results. We discarded one participant because she failed to demonstrate basic knowledge about social tagging.

The 9 interviewed participants are mostly young males (all are 19 year old or older): only one is female and only two reported to be over 30 years old (two preferred not to report their age). Brazilian nationals are the majority (5), followed by Iranians (2) and USA nationals (2). The group is highly educated: all of them have at least a graduate degree. The majority of the participants has an

[3] Available at: http://goo.gl/uLVkRl

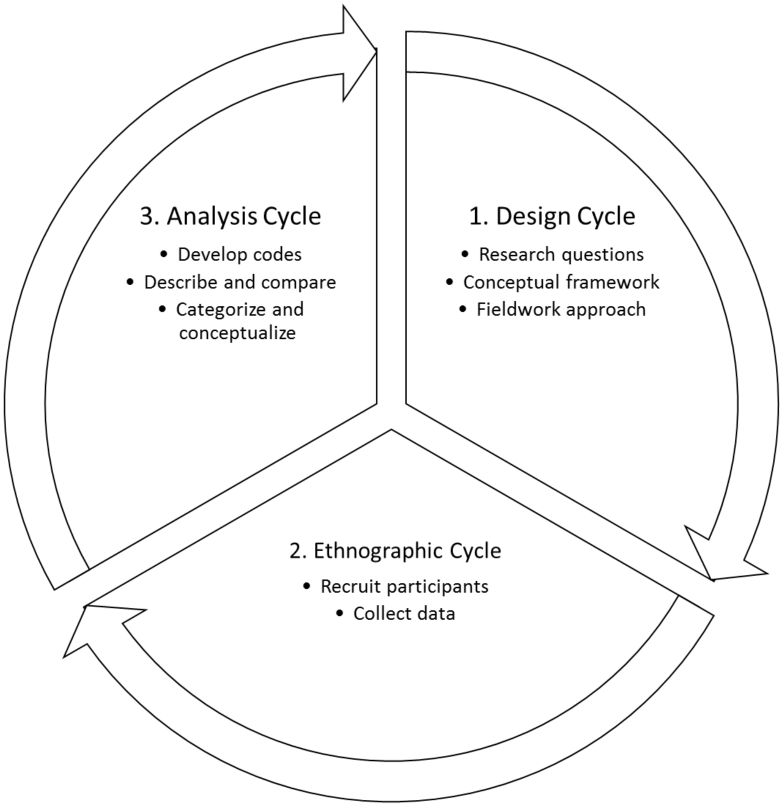


Figure 4.2: Illustration of the qualitative research cycle inspired by the methodology described in Hennik et al. [39] and applied in this work.

engineering/computer science background, while two others have background in linguistics and arts. All of them reported to be fully capable of performing exploratory search tasks, and 8 reported to be able to develop software.

Data collection. The data is collected using semi-structured contextual interviews, as this technique provides flexibility in approaching participants about their tag-based search habits. The interview protocol consisted of open-ended questions that explore the users' application of tagging features in different systems(Appendix A).

The interviews were performed either face-to-face or using a video chat tool at the participants convenience. The duration of each interview was roughly one hour, and consisted of two parts. Both parts consist of contextual inquiries where participants are encouraged to use a social tagging system to illustrate usage and explain their choices of tags while searching.

In the first part of the interview, participants were free to use any system they are familiar with and used to produce tags or use tags to search. The goal is to gain an insight of the users' habits as they explain their understanding of tags and their personal usage choices. In the second part, the participant used a *Delicious*-clone system [4] that is populated with a snapshot of bookmarks and tags collected from *Delicious* (i.e., more than six hundred thousand entries collected in September 2009 [5]). The goal of this task-driven interview is to inquire deeper on the users' decision making process during exploratory searches. By motivating the user with a real search task that is similar to the user' common tasks, we can explore specific aspects that influence the choice for one tag versus another.

All sessions were recorded as a video of the participant's screen and the audio of the conversation. The data collected via the interviews were transcribed, coded, and, finally, analyzed using Grounded Theory methodology [39]. It is worth noting that data collection was conducted iteratively with data analysis. This approach allowed the research to stop recruiting and interviewing new participants when *saturation* in the data is observed (i.e., new issues are not found in the analysis of the interviews transcripts) together a *diverse* demographics in the participant sample.

Analysis. In summary, the analysis cycle consists of *transcription* of interview recordings, coding, description and comparison of codes, codebook consolidation, analysis plan design, categorization, production of a thick description for each identified issue, and, finally, conceptualization. It is important to highlight that that this process is iterative and each step is constantly refined by the output of the previous.

- *Coding*. The codebook is seeded with deductive codes (i.e., codes origi-

[4] Code available at: https://github.com/nigini/GetBoo

[5] http://arvindn.livejournal.com/116137.htm

nated from the related literature). During the process of coding a transcribed interview, new inductive codes may surface from the data. This implies a refinement of the codebook by the addition of new codes. In collaboration with other researcher, the codebook is refined by applying *triangulation*. In this process, each researcher reviews the coding performed by each other independently to consolidate the codebook by means of a discussion and revision of each others codes and coded interviews. The final codebook that results from this step is available at Appendix B.

- *Analysis plan*. The next step consists of defining an analysis roadmap that guides the data searches on the coded interviews with the purpose of addressing the initial research questions. The analysis plan with the result of the data searches per issue is available at: http://goo.gl/YL38nA.

- *Categorization*. It is necessary to group codes with similar attributes to build a set of categories such that an understanding of higher level concepts can be extracted from the data. The categorization is connected to the step of producing *thick descriptions* (see Appendix C of issues and the related codes as resulted from the searches on the data that is guided by the analysis plan.

- *Conceptualization*. Finally, it is possible to move towards the extraction of concepts that connect the issues and help addressing the research questions. The main result of this step is the production of a concept map that highlights aspects that influence users perception of value. This result ultimately can inform the design of methods that assess the value of tags in the context of exploratory search. The concept map is discussed in details in Section 4.5.

The next sections provide a description of findings from the analysis of contextual inquiries based on in-depth interviews, as guided by the research questions mentioned above.

4.3 Which Systems Do Participants Use?

This section discusses the systems in which participants have reported to use tags either to annotate content or to seek information. More importantly, we also inquire about the users' motivation behind using the tagging features of these systems with the intent to confirm previous qualitative studies on the motivations behind tagging.

Systems. Participants reported to use a variety of systems with different motivations for each of them. Twitter was mentioned by eight participants, while Flickr, Delicious, and Facebook were discussed by three participants each. Other systems mentioned were *CiteULike*, *Dribble, Diigo, Evernote, Instagram, Pinterest, StackOverflow, Vimeo*, and *YouTube*.

We note that this set of systems provides an opportunity for a more comprehensive understanding of tag production and consumption compared to previous studies, as they represent almost all categories of Marlow's tagging systems design taxonomy [63], from self-tagging usage (e.g., Evernote) to systems where free-for-all tagging is allowed (e.g., Twitter, Delicious); different types of objects as web pages (*Delicious*), images (Instagram, *Flickr*), and micro-blog posts (*Twitter*); or even considering different types of resource connectivity (e.g., Flickr photos can be grouped, scientific works in Citeulike cite each other) and social connectivity (from following users in *Twitter*to a private usage in Evernote).

Motivation behind tagging. Participants declared a variety of reasons to use tagging systems. Participants often provided many reasons for using a single system or using tagging in general. These motivations can be driven by aspects unrelated to the tagging feature such as the perception that *the process of making sense is faster* in *Twitter*, as declared by participant P1; or, by aspects closely related to the tagging features provided by the system, such as the ability to *bookmark items to read them later*, as declared by participant P3. Other participants declared the need for collaboratively maintaining a list of bookmarks that help them to organize and share a reading list with others was a driving motivation to use systems like *CiteULike*.

In summary, although there are multiple reasons that motivate a participant to use tagging systems, the personal and social information management provided by tagging systems are the main reasons that drive users to contribute to social tagging systems.

4.4 Users Perception of Tag Value

This section presents a qualitative investigation on the aspects that influence users' perception of tag value in social tagging systems. In particular, the qualitative analysis that follows mainly focuses on the value of tags in the context of exploratory search tasks; however, the analysis also contribute to the existing body of research on the motivation behind tagging by presenting the findings about aspects of tag production (Section 4.4.1). Next, Section 4.4.2 presents the analysis of users' perception of tag value for exploratory search.

4.4.1 Aspects of Tag Production

While our main goal is to characterize the perception of tag value in exploratory search tasks to inform the design of methods that quantify tag value, we start by probing about the aspects that influence users' perception of tag value when producing annotations (as opposed to using them to search).

A prevalent theme, as observed from the data collected during the interviews, is that users perceive tags as valuable when they help describing items they are annotating. Such tag assignments are deemed useful as they improve sense-making about a set of items and by making individual items searchable. In particular, interviewees comment on the need of tags to describe images and videos with these two purposes. In this context, tags that describe features of the object such as location, people , and aesthetics characteristics are considered useful. For textual items (e.g., tweets), which themselves are searchable, tags are reported as useful to augment their meaning by making explicit a feeling about the text or providing context for the textual item.

While creating annotations to improve the ability to find the item later, some participants report that there is a tension between using general and specific terms. On the one hand, general tags are likely memorable, and will come to mind as search terms when looking for an item. On the other hand, such tags provide little discriminative power; as they are likely used in many items. Another aspect repeatedly raised by our subjects is the potential of tags to attract attention to items they create or post. Several subjects were concerned with annotating items so that they would likely become more popular or at least with more chances to be finded. As participant P11 describes a strategy to promote content by the use of tags:

P11 – *Instead of writing 'got first place in the fencing championship (state league)', I write 'got first place in the #fencing champion (state league)' as it makes easier to other people find my tweet when searching for that tag*.

Finally, interviewees commented how annotations may attach content to a trend or the contributor to a group. Annotating an item with a tag that is currently used in a trending topic or which is specific to groups is seen as connecting the user with others. Different subjects were motivated to participate in the collective use of a trending tag or were concerned with not using a tag that are normally used by users of a different opinion group.

In summary, the aspects that influence one user's perception of value during the production of tags may not be in tandem with the expectation of another user when searching for items. This is based on the observation that some of the driving forces behind tagging and perception of value during production of tags is highly personal (e.g., feelings), and thus other users may not consider the same tag valuable when trying to locate the same item.

In the next section, we address the question of whether there is indeed a mismatch between the perceptions of tag value while annotating items and searching for them.

4.4.2 Tag Value in Exploratory Search

Exploratory search is the process of acquiring a set of information resources to an information need (e.g., a particular domain) with a certain level of uncertainty. This section presents the qualitative analysis of aspects that influence users' decision-making along the steps of exploratory search such as '*What tag to use?*' or '*When to stop the search?*'.

The contextual interviews, and in particular the search tasks users performed in the second part of the session, enabled us to observe and identify patterns of user behaviour while they engaged in information seeking tasks (see Appendix A for details on the contextual inquiry guidelines used in this study). Users provided data about their decision-making either voluntarily or by answering specific questions about their actions while trying to locate items that fulfill their information needs. Based on our observations that synthesize the behaviour across all participants, the exploratory search process can be illustrated by the high level model shown in Figure 4.3.

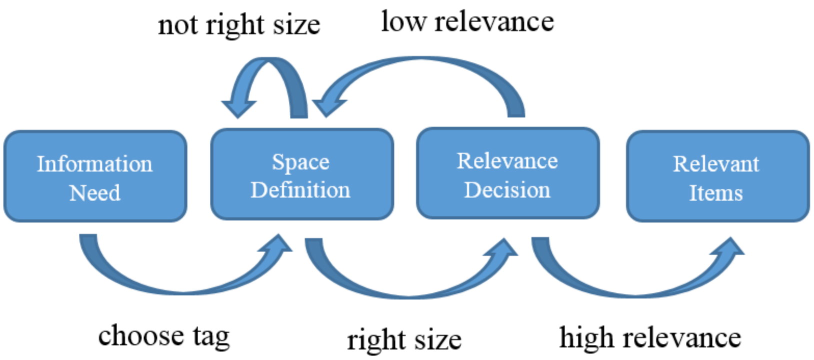


Figure 4.3: Illustration of observed users' transitions and decision making during exploratory search tasks.

The illustrative model in Figure 4.3 points out that users stay in a loop (i) deciding which tags to use to define a search space at each step that better reflects her information need; and (ii) judging the relevance of returned items. More than just reflecting a general intuition about exploratory search processes (and previously proposed models for information foraging [69]), this model is useful when discussing our more specific results about the aspects that influence the participants' decision-making, which we do in the next paragraphs.

Search space definition. A search space is basically a set of items indexed and retrieved by tags. Participants normally define a search space to describe how their information need is translated into tags. such that it provides a search space definition which Search space definition is an essential part of the exploratory search process and involves different perspectives of the set of items retrieved by each tag. Participant P11, for example, during the execution of a Search Task 1 (see Appendix A), clicked on '*web2.0*' and reported that this tag *is more representative of web social networks* (which was the main topic of that participant's information need). The same idea was expressed by participant P6 when choosing the tag '*tutorial*', as the participant explains that '*tutorial*' is a better representation of her particular information need (the participant was looking for material to learn more about programming).

Known vocabulary. As users try to translate their information needs into tags, these tags tend to come from users' known vocabulary. Participant P8 is clear about that when saying that she chooses *hashtags that are alike terms that I hear*, when performing exploratory search on Twitter. The same user goes further and, right after the previous statement, comments on the '*cryptic*' aspect of the tag *#DAADC13* saying: *this one here I would probably not click on because I do not know what it means*. However, this is not simply a matter of a tag to be 'known or unknown' to a given user. Participant P3 justifies choosing '*computer_science*' to search instead of '*computing*' saying that *basically it's because I use it more often*. These observations suggest that the more a tag is already used by a user, the higher its perceived value is.

Search space size. A characteristic of a search space (as defined by the use of a tag) is the search space size (i.e., the number of items it contains). Users tend to refer to the 'right size' of a search space in exploratory search when talking about the decision to continue searching for items. Participant P11, for instance, mentions that *A lot of results is confusing and you'll not be able to find what you want a number of results that doesn't even fill system's first page is kind of frustrating*. Additionally, P4 expressed a *lack of confidence* in the results when *too many items were retrieved*. In contrast, participant P1 took the action of removing an added tag because *it might have filtered too much*. Interestingly, many participants mentioned the number of retrieved items (reported by the system in the search results page) as a way to gauge whether the tags are helping on controlling the search space size. As mentioned by the users, search space size affects their perception of the value of a tag.

Relevance. Besides finding the 'right size' of a search space, no search is complete without locating relevant items. The relevance aspect in our data becomes salient when users are deciding whether a space (composed of retrieved items) fulfills their information needs. In fact, this aspect has been raised and described across all participants, which strongly suggests that this aspect is a major influence on users' perception of tag value.

Participant P7 points out to this aspect by stating: *I am going to take a look at the first five or ten entries to have an idea about my results*. After a brief inspection, the participant decides that *they* (the results) *still have a lot of noise, so I am going to add one more tag*. Similarly, participant P8 reports an analysis of the relevance of the space defined by a tag as saying that *this* (set of items) *is still not sufficient I gave a quick look but the first* (entries) *were not interesting*. Participant P7 is more direct in suggesting that relevance influences the perceived value of a tag when reasoning about a particular choice of tags. The participant selected '*software*' instead of '*programming*' based on the perception that she *will find more things related* (to my information needs) using the former instead of the latter.

Combination of space size and relevance of items. Participants use words like '*focused*', '*specific*', '*restrict*', and '*refined*' to describe a desired search space that balanced well size and relevance. Participant P7 supports this observation by explaining a click decision: *as 'opensource' is already a subset of* (software) *development/programming then I'll start clicking at 'opensource'*. Similarly, participant P1 reasons that adding an additional tag to the navigation is beneficial because *it might give more focused results*. Another strong example related to the influence of space characteristics combination is raised by Participant *P3* when deciding to redefine the space at a particular point of the navigation: *It looks like this* (result) *is really related to 'storage' but there is nothing to do with research. I need to refine it more*. This combination of characteristics of a space (as defined by a tag) influences positively the value of a tag, as a tag can define both a smaller space that contains highly relevant items.

Diversity and neighbouring spaces. Finally, two other identified aspects are connected to tags related to a currently defined search space: diversity and neighboring spaces. To some degree, these two aspects are opposite concepts if one considers that related tags to a given space (presented as a tag cloud) can be perceived as increasing the diversity of items in that space or simply retrieving similar neighboring spaces.

Participant P4 considers confusing to have '*artists*' as part of the tag cloud when the current space is already defined by the tag '*artist*', which suggests that more diversity in the tag cloud improves the perception of value for the tags in the tag cloud relative to the currently defined space. Similarly, participant P1 is even more emphatic about this aspect while performing Search Task 1 (see Appendix A) by stating that: *type and typography both of them point to the same thing, web and website, icon and icons, it's a bit of useless to have these two similar, very similar tags together, this is something that impacts the value, icons have zero value here because you have icon here*. When inquired about whether replacing these highly similar tags by more diverse set of tags would improve the perceived value, the participant replied: *Yes, meaningful diversity within the tags*.

On the other hand, participant P2 selects the tag '*user experience*' after using '*ux*', while reporting that these two terms are considered synonyms. Participant P2 explains that she perceives that the tag '*user experience*' can retrieve results similar to those retrieved (but not annotated with) by the tag '*ux*' (i.e., a neighboring space to the currently defined one). We were unable, however, to identify whether each of these two aspects (i.e., neighboring search space and diversity) is more important than the other regarding the characteristics of a tag cloud to users.

4.5 Concept Map

The analysis reported in the previous section leads to several insights into the aspects that influence the users' perception of tag value. These insights are summarized by the concept map in Figure 4.4.

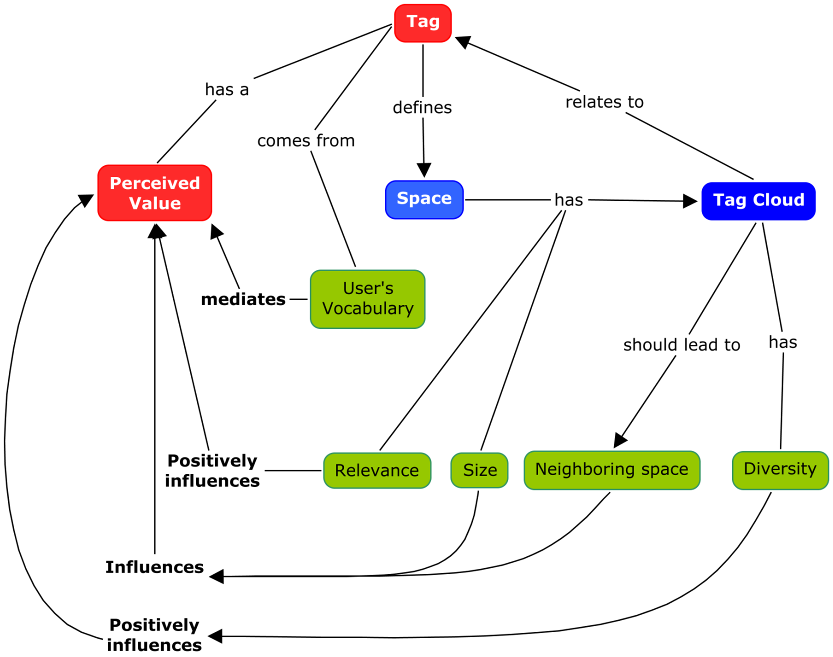


Figure 4.4: Concept map that illustrates the influence of several aspects on the perceived value of tags for exploratory search.

The central entity of the map is a *tag* that has a *perceived value* by users. Tags may come from a *user's vocabulary*, which in turn mediates the perception of value a user has about a tag. Moreover, users express their information need via a tag that defines a *search space*. In turn, a search space has many aspects: *relevance* (of items it contains), *size* (number of items), and a related *tag cloud* (set of tags). Users report, two characteristics of the tag cloud (relative to the tag that already defines the space) influence their perception of value of a tag: first, the ability to explore similar *neighboring spaces* and *diversity* among tags.

These findings are key to design methods that quantify the value of tags to information seekers in the context of exploratory search. The two aspects of search space, as defined by a tag, can therefore guide the design of functions that measure the value of tag. This definition and evaluation is presented in details in Chapter 5.

It is important highlighting that the concepts and their interactions, as presented by the concept map in Figure 4.4, expose both limitations of current technology and fundamental characteristics of user behaviour in exploratory search. For instance, users express that diversity among tags in a tag cloud is important. This fact is generally expressed by showing discontent about the selection of tags by current systems. This is an indication that current systems can be improved. On the other hand, the observation that users define a search space by extracting tags their known vocabulary is a fundamental characteristic of user behavior when performing exploratory search regardless of the system in which this occurs.

4.6 Summary

In summary, at least four aspects, as discussed by the users during the contextual inquiry and revealed by the analysis, influence their perceived value of a tag. In particular, two of them are more salient: *search space size* and *relevance*. Therefore, the findings suggest that the perceived value of a tag is largely influenced by its ability to retrieve items that are relevant to a user while reducing the search space size. The tag reduces the search space by filtering out items, and maximizes relevance by retaining the items that address the user's information needs during exploratory search.

It is also worth highlighting that this study provides an important characterization that can help designing, apart from methods to quantify value of tags, other new social tagging features (e.g., tag cloud algorithms, user interface design, and ranking mechanisms) in many systems, as it improves our understanding of what users consider valuable when searching with tags.

Finally, it is also worth discussing potential threats to validity. As in any qualitative study, this study is subject to some design decisions that may impact its validity. In particular, the most important aspects is the external validity, which I discuss below.

External validity. Although the qualitative study is performed at small scale, which limits our ability to make general claims about the findings, the list of systems described in Section 4.3 shows that the qualitative analysis covers usage scenarios of both tag production and tag-based search in a variety of systems. We believe that this variety of systems provides a broad set of real usage scenarios and reduces the threat to external validity. It is also important to recognize that the diversity of participant demographics is limited.

Chapter 5

Assessing Tag Value for Exploratory Search

The previous chapter provides a characterization of users' perception of tag value. In particular, the analysis focus on aspects users consider when producing tags (i.e., annotating content) and using tags in exploratory search tasks.

This chapter takes the lessons from the qualitative analysis to inform the design of methods that quantify the value of tags [1]. In particular, the investigation presented in this chapter addresses the following question:

- **RQ4**. *How to quantify the value of tags as perceived by information seekers in exploratory search*?

To this end, this chapter presents a formalization of the aspects that, according to the participants of a qualitative study, influence their perception of tag value when they are performing exploratory search tasks. In summary, this chapter contains the following major contributions:

- A framework that helps reasoning about the problem of quantifying the value of user contributions in tagging systems (Section 5.3).

[1] The results presented in this chapter appeared at the following references: [75, 80]

- A method that quantifies the value of tags that caters for the two desirable properties in the context of exploratory search, as identified by the qualitative user study. We prove that this method has desirable theoretical properties while quantifying these two aspects (Section 5.5).
- An experiment using real tagging data that shows that the proposed method accurately quantifies the value of tags according to users' perception (Section 5.6).

The rest of this chapter starts by positioning this study among the related literature.

5.1 Related Work

In a nutshell, this work differs from previous efforts in two main aspects: first, it is motivated by the view that social tagging systems are inherently online peer production systems. Thus to improve the quality of user contributions, it is necessary to first quantify their value, so that one can then think of designing incentives for the production of high quality content. Second, this research focuses on the design and analysis of a method to assess tag value in practice (as opposed to only studying the impact of tags in other information retrieval tasks such as recommendation [3, 9, 27, 90]).

Section 5.1.1 starts by discussing what makes the design of methods to quantify the value of peer-produced information challenging; and, Section 5.1.2 reviews previous works that study the quality of tags in different contexts.

5.1.1 Contributions in Peer Production Systems

Online peer production systems can be categorized into systems where users produce/share resources or information. In the former category, as we have already mentioned, quantifying the value of user contribution is based largely on counting the resource units one user produces and donates to other users (and implicitly to the system). For example, in P2P content sharing systems (e.g., BitTorrent) the value of contributions is estimated by the volume of content a peer donates to others citeAndrade2009. Similar methods have been applied for volunteer computing (e.g., BOINC), where contributions are quantified in terms of CPU hours.

Valuing contributions in these resource-sharing peer production systems relies on: first, the fact that the amount of resources donated are easily quantifiable; second, the assumption that contribution value can be directly linked to the resources consumed to deliver a service; and, third, on the simplifying assumption that a unit of contributed resources has a uniformly perceived value across all users of the system.

In contrast, none of these assumptions holds for systems that support production/sharing of information. First, it is impossible to directly quantify the 'effort' that has led to the production of a specific piece of information; and, second, the value of information (e.g., tags or items in tagging systems) is subjective to users' opinions, interests, and task at hand (an aspect shared with other information goods).

To address this latter issue of contextual value, some peer production systems, such as StackOverflow.com, use intangible rewards (e.g., points) by allowing users to rate content items, as a way to express how much they like a particular item. Ratings can, therefore, be interpreted as an estimate of the value of one user's contribution from the perspective of another. Although this approach generates rich feedback about what users like (or sometimes dislike), it has two limitations. First, rating information is generally sparse (i.e., the majority of users do not express their preferences via ratings); and second, in tagging systems, item rating does little to address the problem of valuing tags. Thus, while this information can support a solution, a direct estimation of value that covers the entire spectrum of peer-produced information is necessary. In this study, we cope with the contextual nature of tag value by: first, using a qualitative analysis to identify the aspects that influence information seekers' perception of tag value (in the context of exploratory search); and, second, using the result of this analysis to inform the design of a method that quantifies the value of tags.

5.1.2 Characterizing the Quality of Tags

Several studies focus on characterizing the quality of tags or tagging (as a feature of an information system) in general. These studies instantiate the notion of 'quality' in various ways, which we comment in turn.

Search and Recommendation. Focusing on the quality of tags for information retrieval tasks such as content classification and search, Figueiredo et al. [27] and Bischoff et al. [9] evaluate the quality of information provided by tags (in comparison to other textual features) to improve the efficiency of recommendation or search mechanisms. Similar studies aim to harness tags to improve web search [41, 99].

Tagging efficiency in decentralized search. Helic et al. [38] study tagging systems from a network theoretical perspective by analyzing whether tagging networks (i.e., formed by the associations of tags to items) have properties that enable efficient decentralized search [1]. In particular, the authors study the impact of different algorithms that build such tag networks (i.e., tag hierarchies, known as folksonomies) on a decentralized search process. The rationale is that a folksonomy is better than another if a decentralized search that uses that folksonomy as background knowledge [1] performs more efficiently.

Tagging as a categorization mechanism. Moving the focus to a different application, Heymann and Garcia-Molina [41] investigate whether tags help users to categorize content by analogy with widely deployed classification tools for library management systems. They use a qualitative analysis to evaluate the power of tags to build classification systems rather than a user-centric quantitative approach to assess value. Lu et al. [61] perform a similar study, by comparing peer-produced tags and expert-assigned terms to classify books aiming at showing that tags can improve accessibility of items in a library catalog.

Quality of textual content. Other studies focus on the content of peer-produced information. Suchanek et al. [92] study the quality of tags by determining the descriptive power of a tag (i.e., its efficiency in describing an item). Similarly, Gu et al. [31] propose a way to measure the confidence in which a tag describes a particular item. In their context, confidence equates to the relevance of the tag to the topic in which the item is part of. More recently, Baeza-Yates et al. [5] characterize the lexical quality of several web sources, including a social tagging system (*Flickr*), finding that the lexical quality of texts in Flickr is better than that in the general web. Other work has focused on methods to detect and mitigate the impact of tag spam [54].

Building tag clouds. Helic et al. [38] and Venetis et al. [94] analyze algorithms to build tag clouds. Their approach to evaluate the quality of a tag cloud is directly related to the present study. However, their method resorts to metrics that aim to capture intuitive aspects of users' information needs – such as novelty, diversity, and coverage of content items in the system – our method concentrates on the relevance and reduction of search space. A comparison of all these methods is a direction for future work.

Our approach differs from previous efforts as we start by characterizing users' perception of tag value to inform the design of a method that quantifies the value of tags for exploratory search. To the best of our knowledge there are no previous attempts to neither characterize the perception of tag value for exploratory search nor design methods to quantify tag value in such context.

5.2 System Model

Before presenting the design of a method that quantifies the value of tags in the context of exploratory search, it is necessary to introduce the system model and some notation.

Let $\mathbb{S} = (U, I, A)$ be a social tagging system, where U represents the set of users in the system, I denotes the set of items, and A represents the set of annotations. An annotation is a tuple that specifies its author, the annotated item, a tag assigned to the item, and the time the annotation happened. Formally, $A = \{(s,i,t,e) | s \in U, i \in I\}$, where t is a *tag*, a word selected by the user from an uncontrolled vocabulary to annotate the item at timestamp e).

The set of annotations A_s characterizes a particular user s, where individual annotations can be distinguished by their timestamps. More formally, $A_s = \{(q,i,t,e) \in A | q = s\}$. From the set of annotations A_s, it is possible to derive the set of items (or user library), and the set of tags (or user vocabulary), respectively annotated and used by particular user s. The library and vocabulary of a user are respectively defined as follows: $I_s = \{i | (s,i,t,e) \in A_s\}$, and $T_s = \{t | (s,i,t,e) \in A_s\}$. The set of tags assigned to a particular item i, T^i, and the set of items tagged with a particular tag t, I^t, are similarly defined.

Item relevance. We assume that, for an information seeker s, there is a probability mass function $p(\cdot|s)$ over the set of items in the system that the information seeker has not annotated yet (i.e., over $I - I_s$) that specifies the relevance of an item i given that information seeker s. Therefore, the set of items relevant to an information seeker can be defined as:

Definition 3 *Given an information seeker s, the set of items relevant to s is:* $\Gamma_s = \{i \in I - I_s | p(i|s) > 0\}$, *where* $p(i|s)$ *is the probability of relevance of an item i to an information seeker s.*

Note that Γ_s can be defined for different search tasks, such as exploring or including a user's own library (i.e., items already tagged by the user). The proposed method is general enough to work with alternative definitions of .

Modeling exploratory search. We model exploratory search as a communication channel between the search engine (the sender) and an information seeker (the receiver). Consider that the sender transmits items to the user, and the channel is characterized by the probability of item relevance $p(i|s)$ to the receiving user over the set of items Γ_s. In this context, a tag defines a filter that creates a new channel from the original one. The new channel is characterized by the probability of item relevance conditional both on a tag and on the information seeker: $p(i|t,s)$

Search space. Using a probabilistic interpretation where the items are assigned with a probability of relevance, a tag t reduces the search space if the probability mass function $p(\cdot|t,s)$ over the set of items Γ_s is more concentrated than the original probability mass function $p(\cdot|s)$ (see discussion below).

Probability estimation. It is worth highlighting that there are many ways to estimate the probabilities of relevance $p(\cdot|s)$ and $p(\cdot|t,s)$. More importantly, it is not our goal to advocate a particular estimator. In particular, the evaluation of our proposed method (Section 5.5) considers two estimators: *i*) a language model based on Bayes smoothing as described in Wang et al. [96]; and, *ii*) a topic model based on Latent Dirichlet Allocation (LDA) as proposed by Harvey et al. [35] [2].

5.3 A framework to assess the value of user contributions

Users in a tagging system are either information producers or information seekers, depending on the action they perform at a given moment. Information producers publish new items and/or annotate existing items. An information seeker navigates the set of items available in the system. To assess the value of a user's contribution in such system, one must combine the value of items and tags produced by the user.

More formally, let $v(t_u,s)$ and $r(i_u,s)$ be two functions that quantify the values of a tag t_u, and of an item i_u, respectively, produced by user u, from the perspective of an information seeker s. A function $K(u,s)$ should combine $v(t_u,s)$ and $r(i_u,s)$ for all tags and items produced by u.

In particular, the intuition behind computing $v(t,s)$ is that the value of a tag should be proportional to its ability to retrieve relevant items while reducing the search space (as the qualitative characterization of users' perception of tag value indicates).

Similarly, the value $r(i,s)$ of an item to an information seeker should be proportional to its 'relevance' and 'usefulness' to user s. This can be estimated di-

[2] In the spirit of enabling reproducible research, all of code used to estimate probabilities based on the work proposed in citeHarvey2011,Wang2010, together with the scripts to produce our presented results are available at: http://github.com/flaviovdf/tag_assess

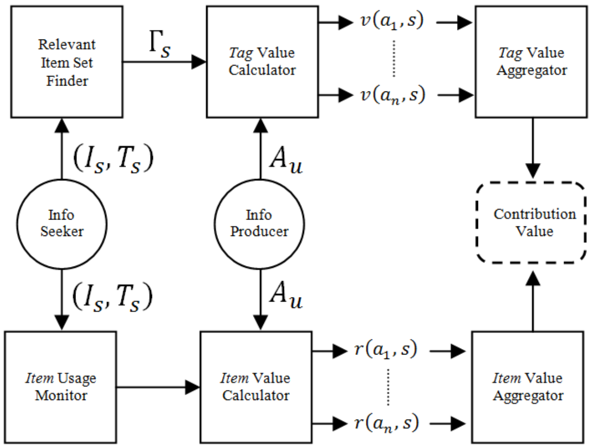


Figure 5.1: Components of a framework to quantify the value of user contributions.

rectly based on: (1) network analysis similar to that applied to the citation graph to find influential authors [52, 97]; (2) direct user feedback such as ratings; or (3) indirect user feedback such as the frequency an item is (re)visited.

Figure 5.1 presents a block diagram that illustrates the process to assess the value of user contributions. The top part of the diagram presents the flow to calculate the value of tags produced by user to an information seeker as a function of the tags' ability to lead to items relevant to her. The 'Tag Value Calculator' block combines the information seeker's set of relevant items Γ_s (produced by the relevant item set estimator, which can be based on an item recommendation engine) and the information producer's annotations to determine the value of tags (i.e., the tags extracted from the annotations produced by u) to s.

The bottom part of the diagram presents the flow to calculate the value of items produced by u that are used by s. The 'Item Value Calculator' box combines the information seeker's item usage statistics, represented by F_s (output from the item usage monitor), and the set of items originally published by u to estimate the value of these items. These usage statistics can be obtained via click traces, for example, that provide information about how often a user consumes a particular item.

Finally, the estimated values of tags and items are aggregated separately and then combined into the value of the contributions from u to s, $K(u,s)$.

It is important to highlight that the proposed framework 5.1 is generic. Each building block can be instantiated according to the specific characteristics of the system. For example, the availability of user activity data, such as records of tag assignments, click traces, item ratings, friendship links, or group co-membership information, can certainly drive the design of specific solutions for the value calculator and aggregator boxes.

The rest of this chapter focuses on assessing the value of tags from the perspective of an information seeker. In particular, it designs and evaluates an instance of the function $v(t,s)$

5.4 A Naive Method

A simple method that could capture both the ability of a tag to reduce the search space and to retrieve relevant items (i.e., the two most salient aspects that influence users' perception of tag value, as presented in Chapter 4) will include as inputs the number of retrieved items (relative to the number of items in the system) and some aggregation of the relevance score of items retrieved (e.g., the average). More formally, let us assume that the relevance score is given by the probability that an item is relevant to an information seeker s. Thus, a naïve method can define the value of tag t from the perspective of the information seeker s as:

$$v(t,s) = (|I| - |I^t|) \sum_{i \in I^t} \frac{p(i|s)}{|I^t|} \qquad (5.1)$$

where $p(i|s)$ is the relevance probability of item i given a user s and it is defined over Γ_s.

Although Equation 5.1 captures the reduction of search space and the relevance of items retrieved by the tag, this method fails to distinguish the value of two tags when they retrieve the same number of items, but the distribution of item relevance is different, yet the average relevance of items in I^t is the same. In this case, choosing the tag that is more valuable to user is simply arbitrary.

To illustrate this situation, suppose two distinct tags and that retrieve the same number of items – i.e., $|I^t| = |I^w|$. Now, consider that the probabilities of items in I^t are $(0.5, 0.2, 0.2, 0.1)$, and those for I^w are $(0.25, 0.25, 0.25, 0.25)$. In this case, the average relevance is the same for both tags, but the item relevance distribution for the items retrieved by t is more concentrated. In this case, tag t should be considered more efficient than w when used by a user to explore the set of item, as it reduces the search space (probabilistically) by concentrating the relevance distribution. However, this naïve method is unable to assign appropriate values to each tag. Therefore, it is important that a method takes into account the distribution of item relevance from the perspective of an information seeker, given that she uses a tag to prune the original search space of items I. The next section elaborates on this idea, and introduces our proposed method.

5.5 An Information-theoretical Approach

We split the presentation of our method into three parts: first, we present how we estimate the reduction of search space by a tag; second, we discuss an approach to estimate the relevance of the set of items retrieved by tag; and, finally, we combine these two components.

Estimating search space reduction. To estimate how much a tag reduces the search space for a given information seeker, the proposed method assumes a probabilistic interpretation (as opposed to assuming a deterministic approach that counts the number of filtered items by the tag). In our model of exploratory search a tag reduces the search space by *leading to a higher concentration* on the probability of relevance over the set of retrieved items.

More formally, given the distribution of probability of relevance $p(\cdot|s)$, and the conditional probability distribution $p(\cdot|t,s)$ over the set of relevant items Γ_s, to measure how much information one gains by using the channel defined by a tag to read the set of items , the proposed method uses the Kullback-Leibler divergence [21] of the two distributions, as defined below:

$$D_{KL}(p(\cdot|t,s)||p(\cdot|s)) = \sum_{i \in \Gamma_s} p(i|t,s) \log \frac{p(i|t,s)}{p(i|s)} \tag{5.2}$$

where $p(i|t,s)$ represents the probability that an item i is relevant to a given information seeker s when she uses a tag t to navigate the system; while $p(i|s)$ represents the probability that an item i is relevant to s.

Equation 5.2 measures the reduction in the item search space by a given tag t, as it quantifies how much the distribution of relevance conditional on a tag $p(i|t,s)$ diverges from the probability of relevance of an item i. The reduction in search space occurs, for example, when conditioning $p(i|s)$ to a tag t concentrates the probability of relevance over a smaller set of items. However, as conditioning to a tag may increase the concentration of the probability mass $p(\cdot|s)$ over fewer relevant items. Therefore, it is necessary to complement Equation 5.2 with a measure of relevance of items a tag t delivers to an information seeker s.

Estimating delivered relevance. To estimate the relevance of a set of items retrieved by tag t to a particular user s, we compare the set of items retrieved (ordered by probability of relevance) to a reference point – a subset with top items of Γ_s ordered by probability of relevance. The intuition is that the more items from the top of the ranked Γ_s the tag retrieves, the more valuable it will be. Note that according to this definition a tag maximizes its ability to retrieve relevant items by retrieving all items. This, however, does not necessarily maximize its value, as it will depend on how much of the search space the tag is able to reduce (as measured by Equation 5.2).

More formally, let I^t be the set of items retrieved by a tag t and *not* already annotated by the information seeker s (i.e., $I^t \not\subset I_s$). Also, let I^t be ordered by relevance to an information seeker s. Let $\Gamma_s^{[k]}$ be the set of top-k most relevant items to s from Γ_s when ordered according to $p(\cdot|s)$. We define the relevance delivered by a tag t to an information seeker s as:

$$\rho(t,s) = 1 - \tau\left(I^t, \Gamma_s^{[k]}\right) \tag{5.3}$$

where $\tau\left(I^t, \Gamma_s^{[k]}\right)$ is the generalized Kendall's τ distance [3] between I^t and $\Gamma_s^{[k]}$, and $k = |I^t|$. Kendall's τ distance measures the fraction of the number of changes needed (in regards to the maximum number of changes) to convert one rank (I^t) to the other ($\Gamma_s^{[k]}$). 0 distance means ranks are the same, while 1 states that the ranks are exact opposites. A penalty (which can be specified from 0 to 1) is incurred when items appear on one rank but not the other. The rationale is that the more relevant items a given tag retrieves, the smaller is the distance and the closer to 1, $\rho(t,s)$ gets.

Combining relevance and reduction of search space. The final step is to define the estimate of the value of tag t, from the perspective of an information seeker s, $v(t,s)$.

Definition 4 *Given an information seeker s and her set of relevant items Γ_s, the value of a tag t to s, is defined as:*

$$v(t,s) = \rho(t,s) D_{KL}(p(\cdot|t,s)||p(\cdot|s)) \tag{5.4}$$

The rationale behind this definition of tag value is that if a tag t retrieves only items with low relevance to s, the factor $\rho(t,s)$ penalizes the value, as it computes the distance from the retrieved set of items to the set of estimated relevant items to the user. Therefore, tag t has little value to the information seeker, even though it may reduce the search space towards a subset of Γ_s. On the other hand, if t leads the user to a subset of relevant items, its value is proportional to the reduction in search space, as the relevance of the items retrieved by t, which is represented by the coefficient $\rho(t,s)$, will be close to one and will have a smaller penalty effect.

5.5.1 Search Space Reduction Property

This section shows that the method we propose can indeed distinguish between two arbitrary tags, when they deliver different levels of reduction of search space.

[3] This quantity is also known as the Kendall distance with a penalty [25], as it introduces a penalty parameter to extend the original Kendall's tau distance to enable the comparison of top-k lists with different elements

As described in Section 5.5, we use a probabilistic interpretation of the search space, where the items are assigned with a probability of relevance. Therefore, a tag reduces the search space if the probability mass function $p(\cdot|t,s)$ over the set of items Γ_s is more concentrated than $p(\cdot|s)$.

The goal of this analysis is to show that our proposed method is able to distinguish between two tags that lead to different levels of search space reduction. More formally, we prove the following proposition:

Proposition 1 *Given an information seeker s, if a tag t reduces the search space more than another tag w by moving the probability mass towards more relevant items, then* $D_{KL}(p(\cdot|t,s)||p(\cdot|s)) > D_{KL}(p(\cdot|w,s)||p(\cdot|s))$.

Proof. The first condition in the proposition is such that given an information seeker s, if a tag t reduces the search space more than another tag w, we have that: $p(\cdot|t,s)$ is more concentrated than $p(\cdot|w,s)$, where the functions are defined over Γ_s. Therefore, $H(p(\cdot|t,s)) < H(p(\cdot|w,s))$, where H is Shannon's entropy [21].

Moreover, if according to the second condition in the proposition, tag t moves the probability mass towards more relevant items than the tag w does, this means that there are *at least* two items in $i,k \in \Gamma_s$ where $p(j|s) > p(k|s)$ and when conditioning the probability to each tag t and w, respectively, $p(j|t,s) > p(j|w,s)$ and $p(k|t,s) < p(k|w,s)$. Note that to conserve the probability mass, it is necessary that $|p(j|t,s) - p(j|w,s)| = |p(k|t,s) - p(k|w,s)|$.

Putting these two conditions together and applying Equation 5.2 to $p(\cdot|t,s)$ and $p(\cdot|w,s)$, we prove, by contradiction, that Proposition 1 holds:

$$
\begin{aligned}
D_{KL}(p(i|t,s)||p(i|s)) &< D_{KL}(p(i|w,s)||p(i|s)) \\
\sum_{i\ \in\Gamma_s} p(i|t,s)\log\left[\frac{p(i|t,s)}{p(i|s)}\right] &< \sum_{i\ \in\Gamma_s} p(i|w,s)\log\left[\frac{p(i|w,s)}{p(i|s)}\right] \\
\sum_{i\ \in\Gamma_s} p(i|t,s)\log[p(i|t,s)] & \\
-\sum_{i\ \in\Gamma_s} p(i|t,s)\log[p(i|s)] &< \sum_{i\ \in\Gamma_s} p(i|w,s)\log[p(i|w,s)] \\
& \quad -\sum_{i\ \in\Gamma_s} p(i|w,s)\log[p(i|s)]
\end{aligned}
$$

Replacing the first summations by the entropy term leads to:

$$
\begin{aligned}
-H(p(i|t,s)) & \\
-\sum_{i\ \in\Gamma_s} p(i|t,s)\log[p(i|s)] &< -H(p(i|w,s)) \\
& \quad -\sum_{i\ \in\Gamma_s} p(i|w,s)\log[p(i|s)]
\end{aligned}
$$

Next, by expanding the second summation on each side to isolate j and k, we have:

$$
\begin{aligned}
-H(p(i|t,s)) - \sum_{i\ \in\Gamma_s-\{j,k\}} (p(i|t,s)\log[p(i|s)]) & \\
-p(j|t,s)\log p(j|s) - p(k|t,s)\log p(k|s) &< -H(p(i|w,s)) \\
& \quad -\sum_{i\ \in\Gamma_s-\{j,k\}} (p(i|w,s)\log[p(i|s)]) \\
& \quad -p(j|w,s)\log p(j|s) - p(k|w,s)\log p(k|s)
\end{aligned}
$$

Cancelling the equal summations from each side leads to:

$$\begin{aligned} -H(p(i|t,s))+\log p(j|s) &< -H(p(i|w,s))+\log p(k|s) \\ H(p(i|w,s))-H(p(i|t,s)) &< \log p(k|s)-\log p(j|s) \end{aligned}$$

From the first condition set forth in the proposition, we know that $H(p(\cdot|w,s)) - H(p(\cdot|t,s)) > 0$, and from the second condition $p(k|s) - p(j|s) < 0$. Therefore, the last equation contradicts the original conditions, and the propositions holds.
□

5.5.2 Relevance Property

This section shows that the proposed method can distinguish between two arbitrary tags, when they deliver different relevance levels. In particular, the analysis show that, from the perspective of a given information seeker s, Equation 5.4 distinguishes two tags if they deliver two different levels of relevance. To show that our proposed method has this property, we prove the following proposition.

Proposition 2 *Given an information seeker s, if a tag t retrieves more relevant items than a tag w, it follows that* $\rho(t,s) > \rho(w,s)$.

Proof. If t retrieves more relevant items than w, we have that:

$$\tau\left(I^t, \Gamma_s^{[k]}\right) < \tau\left(I^w, \Gamma_s^{[k']}\right)$$

where $k = |I^t|$ and $k' = |I^w|$. By inverting the signs and adding 1 to both sides, we have:

$$1 - \tau\left(I^t, \Gamma_s^{[k]}\right) > 1 - \tau\left(I^w, \Gamma_s^{[k']}\right)$$

Therefore, $\rho(t,s) > \rho(w,s)$.
□

5.6 Evaluation

The previous section presents proofs that the proposed method can differentiate between two tags when they lead to different levels of search space reduction and relevance of retrieved items. This section complements these results by performing an experiment with real data to test the accuracy of our method. The method is accurate if the tag values it produces match users' perception of value.

Two hard constraints limit the validation experiments we can execute: we do not have access to browsing traces and we do not have access to a ground truth, that is, direct estimates of users' perception of value.

We have, however, access to tag assignment traces in a number of systems and we use them to estimate our method's accuracy based the following intuition: when a user assigns a tag to an item, this is the tag that had a high value for the user from the perspective of a future search for that particular object. Thus, if our method consistently estimates the value of the previously used tags higher than the value of random tags (that the user have not used before), then there is a strong indication that the method is accurate in quantifying tag value as perceived by users.

5.6.1 Experiment Design

To test the hypothesis that the proposed method passes this accuracy criterion, we collect tag assignments from a real social tagging system LibraryThing [4] (as described in Table 5.1).

Table 5.1: Data set used as input of an experiment that estimates the method's accuracy

	LibraryThing
# Users	7,279
# Items	37,232
# Tags (distinct)	10,559
# Tag Assignments	2,056,487

[4] LibraryThing data collected from: http://www.macle.nl/tud/LT/

The experiment consists of two major steps: *i*) finding the best probability estimator parameters (steps 2 to 4) to, which are used as inputs to our method; and, *ii*) for each user, computing the value of tags from two samples; a sample of tags from the user's vocabulary and a sample of tags not in the user's vocabulary (steps 5 and 6). These samples are denoted by $G_s \subset T_s$ and $R_s \subset T - T_s$. It is important to highlight that neither tags in G_s nor tags in R_s are used in the parameter estimation phase. Thus, the method has no information whether the user has annotated items with them before.

More formally, this experiment tests the hypothesis that the method is able to assign higher value to tags in G_s than to tags in R_s. The following steps provide a detailed set of steps we follow in the experiment:

1. First, we select a sample of users that use the system more than occasionally, that is, users with at least 50 annotated items. We denote this sample by S_{50}..

2. With the tagging trace sorted by annotation timestamp, we break the set of annotations A into three sets: A_{train}, $A_{validation}$, and A_{test}. The training set contains the first 80% (sorted by date) of items annotated for the users in the sample S_{50}. The validation and test set are each composed by 10% of the remaining annotations. We made sure that all tags/items on the validation and test sets, also appeared on training set.

3. We train the estimators (based on different parameters) for the probability distributions $p(\cdot|t,s)$ and $p(\cdot|s)$ on A_{train}. Models trained were based on language models [96] and topic models [35]. As in [38], we were unable to reproduce the results in [96], thus for the rest of this section we shall discuss results based on topic models only.

4. The set of items on the $A_{validation}$ are then used to measure average Success10 (as in [96]) of the estimator for each user. Success@10 captures the fraction of times at least one relevant item, that is, one item in the validation set, appeared in the first set of the first 10 items when sorted by

$p(i|t,s)$ or $p(i|s)$. Each probability distribution is evaluated independently of the other. This way, we pick the best estimator parameterization for both probability distributions. The best estimators reached Success@10 values of 0.05 and 0.06 for $p(i|t,s)$ and $p(i|s)$ respectively. Parameters used are $\alpha = 0.1/|I|, \beta = 0.1/|T|, \gamma = 0.001$. We refer the reader to [35] and our source code for more details on parameters and implementation issues.

5. With the best parameterization, we use A_{test} to perform our experiments. Recall that no parameter tuning is done on this set. Now, for each user $s \in S_{50}$, two sets of tags are constructed, namely: *hidden* and *random*. The hidden set, denoted by $G_s \subset T_s$, contains tags used by user s in the test set A_{test}. The random set, denoted by R_s, is comprised of 50 tags that are randomly selected from the trace and have not been used by the user on any of the train, test or validation sets.

6. Finally, we compare the distributions of tag value $v(t,s)$ for tags in $G = \cup_{s \in S_{50}} G_s$ to that of the tags in $R = \cup_{s \in S_{50}} R_s$.

5.6.2 Results

Figure 5.2 shows the results for the naïve method. The plot shows the cumulative distribution functions (CDF) of values for both tag sets from the perspective of all users in the LibraryThing data. The result shows that the naïve method is not efficient in distinguishing between tags that users find valuable (i.e., those part of the hidden set) and the others (i.e., those part of the random set).

In contrast, Figure 5.3 shows the CDF for tags values computed using our proposed method based on the information theoretical approach. The result shows that the distribution of tag values for tags in the hidden set G_s is concentrated over larger values than that of tags in the random set R_s (i.e., tags that are chosen at random and that have not used by the user).

To confirm that the tag values for tags in one sample are significantly larger than those from the other sample, we apply a Kolmogorov-Smirnov test. In fact,

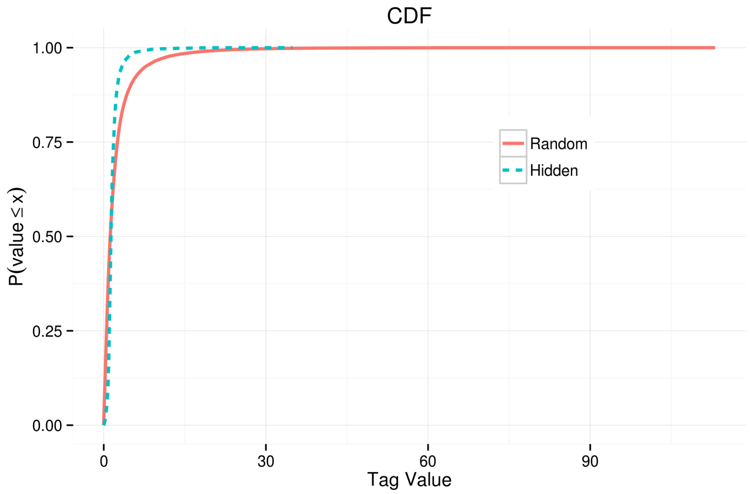


Figure 5.2: Comparison between the cumulative distribution functions (CDF) of tag values (naïve method), for tags in each set (Hidden and Random), from the perspective of each user in the LibraryThing data set.

the test allows the rejection of the null hypothesis that the values in the samples come from the same distribution, and accept the alternative hypothesis that the distribution of tag values for tags in the hidden set lies below that of random.

In particular, we observe that the D-statistic, which measures the distance between the two CDFs, for the Information Theoretical Method is 2.5 times larger th an that of the Naïve Method. The larger the difference the better is the method in distinguishing the valuable (hidden) from random tags. In fact, $D^{-} = 0.25$ $(p < 2.2 \times 10^{-16})$ for Naïve; $D^{-} = 0.64$ $(p < 2.2 \times 10^{-16})$ for our method.

Therefore, these experiments provide evidence that the proposed method (formalized by Equation 5.4) is accurate, as it is able to assign higher values to those tags that users perceive as more valuable.

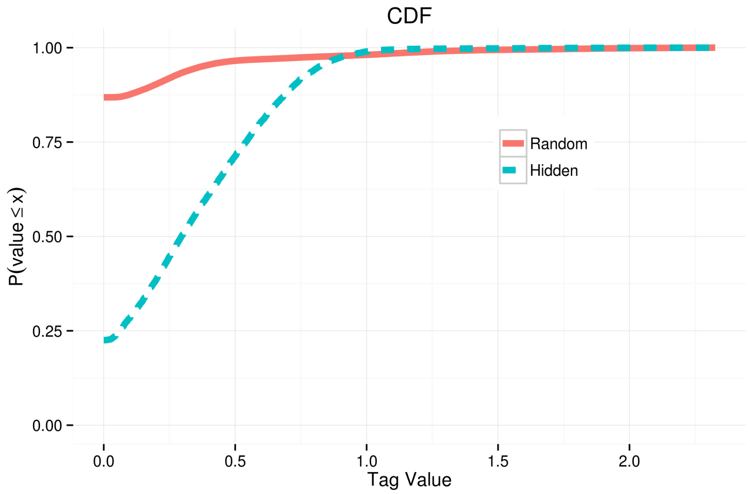


Figure 5.3: Comparison between the CDFs of tag values (our proposed method), for tags in each set (Hidden and Random), from the perspective of each user in the LibraryThing data set

5.6.3 Alternatives

This section discusses alternatives to the proposed method. Additionally, this section presents directions towards an experiment that compares methods to assess the value of tags regarding their ability to accurately capture the aspects that influence users' perception of value.

Alternatives to the proposed method can be divided into two classes: *i*) incorporation of user feedback; and, *ii* traditional information retrieval metrics. The next paragraphs discuss these in turn.

User feedback. In production, the method proposed in the previous section could be augmented to incorporate user feedback. In particular, the use of user click traces (i.e., a log of what tag users clicked when performing an exploratory search task) would not only enable better item relevance estimation, but also an important ground truth for evaluation purposes (as the tags clicked before finding the items of interest indicate what tags were valuable in that search task).

Traditional information retrieval metrics. It is natural to consider that other traditional metrics used in information retrieval tasks could be adapted to measure the value of tags. For instance, TF-IDF or *F-measure* could be used to assign value to tags. These metrics could provide a partial *ordering* on the set of tags in the system from the perspective of an information seeker. Similar to the proposed method, these metrics can be used to measure the ability of a tag to retrieve relevant items to a given information seeker. However, in contrast to the method proposed in this research, these traditional information retrieval metrics do not account explicitly for the reduction in search space that a particular tag can achieve.

Alternative experiment design. Assuming one can build a ground truth (e.g., based on click traces) that contains, for each information seeker, (at least) two classes of tags, namely: *more valuable* and *less valuable*. Therefore, given an information seeker, a tag, and a method, the effectiveness of methods can be compared by measuring their performance in classifying the tag correctly. More concretely, the experiment can be recast as a logistic regression problem with the class of tag as the output variable, while the value of the tag is considered the *feature* of a tag. The rationale behind this experiment is that if a method is accurate in capturing the value of tags from the perspective of an information seeker, using the value of the tag (measured by the method) as a feature in the classifier leads to an effective prediction of tag class (i.e., more valuable or less valuable).

5.7 Summary

This study focuses on the problem of quantifying the value of peer-produced tags for exploratory search. Informed by the qualitative analysis of these aspects, this study designs a method that quantifies tag value by considering the two most salient aspects identified by the qualitative analysis: reduction of search space and relevance of retrieved items (Chapter 4). Finally, an evaluation with real tagging data provides evidence that the proposed method is able to quantify and differentiate valuable tags from those less valuable.

It is also important to note that our qualitative analysis uncovers several aspects that influence the users' perception of tag value in exploratory search. The proposed method quantifies the two most salient ones. Therefore, the plan to extend this method to account for the other aspects is a future work. A larger evaluation of our methods using either a collected ground truth or click traces is also a natural extension of this study.

Finally, as the experiment design is subject to decisions that may threaten its validity, it is necessary to comment on the *Internal validity* of this study. One potential source of threats to internal validity is the interaction between data used in the probability estimators and the methods that assess the value of a tag, however we guarantee that this threat is removed by breaking the trace into three disjoint segments (training, test, and validation) to avoid using the same data in training (i.e., probability estimators) and testing (i.e., tag value computation).

Chapter 6

Assessing Value of Peer-Produced Information for Content Promotion

This chapter focuses on assessing the value of peer-produced information in a different context – *content promotion* [1].

Based on observations provided by the characterization of users' perception of tag value, where some participants state that one strong motivation produce tags is to promote their content. This observation together with the sheer volume content owners generate (e.g., *YouTube* receives 100 hours of video every minute [2]) turn creates the problem of optimizing tags to improve content viewership.

To cope with such high volume of content, it is common for large-scale content owners to offload online publication and monetization tasks to specialized content management companies. The job of content managers is to publish, monitor, and promote the owner's content, and usually there is a revenue sharing agreement between the content manager and the content owner (e.g., the content managers' revenue is directly related to the number of ad prints each piece of content receives).

Although viewers may reach a content item starting from many 'leads' (e.g.,

[1] The results presented in this chapter appeared at the following references: [81]

[2] http://www.youtube.com/yt/press/statistics.html

an e-mail from a friend or a promotion campaign in an online social network), a large portion of viewers relies on keyword-based search and/or tag-based navigation to find videos. An argument supporting this assertion is the fact that 10.5% of the unique visitors in *YouTube* come from Google.com searches [3]. With the integration of Google and *YouTube* search, one might expect that the volume of search traffic that leads to views on *YouTube* will only increase. Moreover, *YouTube* is the third most popular site on the web; behind Facebook.com and Google.com [4].

Consequently, the textual features of video content (e.g., the title, description, comments, and tags, in the case of *YouTube* videos) have a major impact on the view count of each particular content item and ultimately on the revenues of the content manager and content owner [44, 102].

Experts can produce these textual features via manual inspection of a content object (and our industry contacts confirm that this is a still current practice [5]). This solution, however, is manpower intensive and limits the scale at which content managers can operate. Therefore, mechanisms to support this process (e.g., automating tag and title suggestion) are desirable. It has been shown that simple suggestions of textual features produce positive results: for example, title suggestions in eBay have benefitted both sellers, who increased revenue, and buyers, who found relevant products faster [44].

With the ever increasing volume of user-generated content available on the Web, there is a plethora of sources from which an automated mechanism that suggests textual features, in general, and tags, in particular, could extract candidate terms. For example, Wikipedia (a peer-produced encyclopedia), MovieLens and Rotten Tomatoes (social networks where movie enthusiasts collaboratively catalog, rate, and annotate movies), New York Times movie review section (which includes over 28,000 movies) or even *YouTube* [102] comments are potential sources of candidate keywords to annotate user-generated video.

[3] http://www.alexa.com/siteinfo/youtube.com#keywords

[4] http://www.alexa.com/topsites/global

[5] This has been confirmed by companies who provide content management to large content producers such as NBA

This study primarily investigates the value of various information repositories like the above when used as data sources for tag recommendation algorithms that aim to boost video-content popularity. In particular, the data sources are categorized as *peer*- or *expert*-produced according to their production mode, and evaluate whether sourcing from one category or the other leads to better recommended tags. The following research questions drive the investigations presented in this chapter:

- **RQ5.1**. To what extent the tags that are currently associated with existing *YouTube* content are optimized to attract search traffic? Is there room for improvement using automated tag recommendation solutions?
- **RQ5.2**. How do peer- and expert-produced input data sources compare with regards to their impact on the performance of tag recommenders for boosting content popularity?
- **RQ5.3**. Do peer-production aspects such as the number of contributors to a data source influence the effectiveness of tag recommenders that aim to increase content popularity?

It is worth highlighting that this work uses recommender algorithms in a different context than many previous studies: the goal is *not* to design novel and more efficient recommendation algorithms but to explore the impact of input data source choice. While previous work proposes tag recommenders that aim to maximize, for instance, relevance or diversity [7, 32, 56, 62, 72], this study focuses on comparing the outcomes of using different sources of information (e.g., peer- and expert-produced) when recommending tags to boost video popularity.

In summary, the contributions of this work are:

- The evidence that the tags associated with a sample of trailers of popular movies currently available on *YouTube* can be optimized by an automated process: either by incorporating human computing engines (e.g., Amazon

Mechanical Turk) at a much lower cost than using dedicated ‘channel managers’ (the current industry practice), or, at en even lower cost, by using recommender algorithms to harness textual produced by a multitude of data sources that are related to the video content.

- A comparison of the effectiveness of using peer- and expert-produced sources of information as input for tag recommender that aim to boost content popularity.
- The production of a ground truth that is available to the community (together with the implemented tools).

It is worth noting that the quest to improve visibility of one’s content (e.g., a website, a video) is not new - the whole Search Engine Optimization segment has seen uninterrupted attention. Multiple avenues are available ranging from some that are viewed as abusive (e.g., link-farms) to perfectly legitimate ones (e.g., better content organization, good summaries in the title-bar of webpages). Our exploration falls into this latter category.

6.1 Related Work

The related literature falls into two broad categories: automated content annotation and tag value assessment. This section briefly discusses the previous works on each topic in turn, and positions this work among these previous efforts by highlighting the novel aspects of our comparison.

The majority of related work on automated content annotation (or tag recommendation) focuses on suggesting tags to annotate content items such that they maximize the relevance of the tag given the content [7, 17, 44, 60, 95] with a few exceptions where authors propose to leverage other aspect such as diversity [7].

Although finding tags that are relevant to a given content item is an important component of improving the tags assigned to the content, previous studies fail to account for the potential improvement on the view count of the annotated content an aspect which is valuable to content managers and publishers, as they monetize based on the audience that is able to find their content.

Zhou et al. [102] study is, to the best of our knowledge, the closest to our work. The authors focus on the same problem we use as a backdrop to investigate the value of data sources on improving content popularity via tag recommendation. However, contrary to our study that considers the search portion of traffic that reach videos, the authors approach the problem of boosting video popularity by proposing approaches to connect videos to other influential videos as a way to leverage the related video recommendations.

Our study is different on another axis as it concentrates on evaluating the impact of data source choice instead of aiming to design a new recommendation algorithm. This differentiates this work from other studies [32, 48, 62, 72, 84] as well. Therefore, our work complements and extends these previous efforts as recommender systems solutions could be designed combining the techniques proposed previously and the knowledge about the valuable data sources and their combination to improve content popularity such as the one proposed by Lipczak and Milios [59].

The problem of assessing the value of data sources for boosting content popularity via tag recommendation is related to assessing the value of individual tags in other contexts. In the context of exploratory search, for example, Santos-Neto et al. [65] (Chapter 5) pose the problem of assessing the value of contributions in social tagging systems. The authors argue that the value of collaboratively produced tags, in the context of exploratory search, is proportional to their ability to improve the efficiency of information seeking tasks from the perspective of a user.

In a different context, Gu et al. [31] propose a method to quantify 'tag confidence' in social tagging systems. Their approach quantifies the quality of tags produced collaboratively in social tagging systems by taking into account two aspects of a tag: *i*) the credibility of its producer; and, *ii*) the strength of its semantic relation to the tagged resource. Therefore, their work is an answer if one aims to optimize the tags of a content item that best

6.2 Context for Assessing the Value of Peer-Produced Information

This section describes the context in which we investigate the effectiveness of different information sources as inputs for tag recommendation algorithms and presents the formal statement of the recommendation problem used as a backdrop in our investigation.

Annotating a video with tags that match the terms users would use to search for it increases the chance that users view the video. Various textual sources that are related to the video and whose content can be automatically retrieved (e.g., movie reviews, comments, wiki-pages, news items) can be used as input sources for recommenders to suggest tags for these content items.

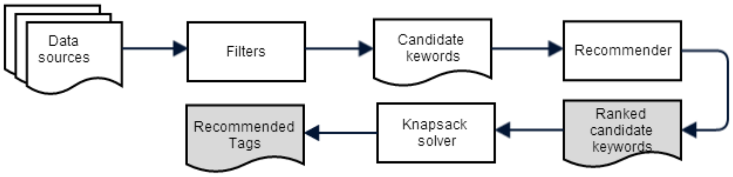


Figure 6.1: The recommendation pipeline.

A recommendation pipeline that implements the previous idea is schematically presented in Figure 6.1: data sources feed the pipeline with textual input data. Next, the textual data is pre-processed by filters to both clean and augment it (e.g., remove stopwords, detect named entities). This first processing step provides candidate keywords for the recommenders. The recommendation step uses the candidate keywords (and their related statistics, such as frequency and co-occurrence) to produce a ranked list according to a scoring function implemented by a given recommender algorithm. Finally, as the space available for tags provided by video sharing websites, such as *YouTube* or Vimeo, is limited, the selection of most valuable candidate keywords is constrained by a budget, often defined by the number of words or characters. Therefore, the final step consists of solving an instance of the 0-1-knapsack problem [20] that selects a set of recommended tags from the ranked-list produced by the recommender.

In summary, the recommendation pipeline is composed of four main elements: data sources, filters, recommender, and knapsack solver. The next paragraphs discuss each of these elements.

- **Data Sources**. This component provides the input textual data used by the recommenders. In particular, we are interested in peer-produced data sources such as Wikipedia and social tagging systems like MovieLens, as well as expert-produced data sources such as NYTimes movie reviews. We discuss in detail each of the data sources used in Section 6.4.1

- **Filters**. The raw textual data extracted from a data source is filtered to minimize noise. We consider simple filters such as stopwords and punctuation removal, lowercasing, and named entity detection [6] input data.

- **Recommender**. Starting from a set of candidate keywords together with relevant statistics (e.g., frequency, co-occurrence), a recommender scores the candidate keywords. Note that there are many ways of defining scoring functions; and, it is not our goal to advocate a specific scoring function or recommender. The intention is to investigate the influence of the choice of the data source on their performance. We discuss the recommenders used in this work in Section 6.4.2

- **Knapsack Solver**. Finally, after ranking candidate keywords, the final step is selecting the ones which best fit the budget. In this paper the budget is expressed in terms of the number of characters as done in video sharing systems such as *YouTube*, where the total number of characters one can use for tags is limited to 500. This step is formulated as the 0-1-knapsack problem, as follows:

  Let v be a video and $C\langle k_i \rangle, i = 1...n$ be a list of candidate keywords provided by a data source when used as input to a tag recommendation algorithm.

[6] We leverage OpenCalais web service to perform named entity detection. http://www.opencalais.com

Additionally, let us denote the length of a keyword k_i by w_i. Therefore, the problem of selecting the best tags that to improve viewership of the video v is equivalent to solving the following optimization [20]:

$$\begin{array}{rl} \text{maximize} & \sum_i^n f(k_i, v) x_i \\ \text{subject to} & \sum_i^n w_i x_i \leq B \end{array} \tag{6.1}$$

where B is the budget in terms of number of characters allowed in the tags field, $x_i \in \{0,1\}$ is an indicator variable, and $f(k_i, v)$ is a scoring function provided by the recommender for the keyword k_i with respect to the video v. Considering that the cost [7] (i.e., the keyword length) and scores are both nonnegative, we use a well-known dynamic programming algorithm [20] to solve this optimization problem.

6.3 Building the Ground Truth

The ideal ground truth would consist of experiments that vary the set of tags associate to videos and capture their impact on the number of views attracted. However, collecting this ground truth requires having the publishing rights for the videos and, even then, it implies executing experiments over a considerable duration.

After unsuccessful attempts to collaborate with content publishers to execute such an experiment we decided for an alternative solution: we built a ground truth by setting up a survey using the Amazon Mechanical Turk [8] a video and answer the question: *What query terms they would use to search for that video?* The rationale is that these terms would, if used as tags to annotate the video, maximize

[7] The budget can be defined in terms of number of tags as in *Vimeo* which restricts to 20 the number of tags a publisher can apply to an uploaded video. This study can easily be extended to consider this situation.

[8] www.mturk.com

its retrieval by *YouTube* search engine (and indirectly maximize viewership) while being relevant to the video.

The rest of this section presents the details of our methodology to build the ground truth and characterizes it.

Content Selection. Our study focuses on a specific type of content: movies [9] We ask turkers (i.e., the Amazon Mechanical Turk workers who accept to participate in the survey) to watch movie trailers, and not the actual movies. The reason is that the trailers are generally short (about five minutes or less), and this makes it possible to have the evaluation process more dynamic, encouraging 'turkers' to watch more trailers and associate more keywords to them.

In total, our dataset consists of 382 movies that were selected to meet two constraints: Firstly, their trailers must be available on *YouTube*; secondly, to enable comparisons, the movies selected had to have reviews available via the NYTimes movie reviews API [10], and records in the MovieLens catalog [11] the data sources used in our experiments in more detail.

Survey. First, we conducted a pilot survey by recruiting participants via our internal mailing lists and online social networks. This pilot highlighted two major issues: i) relying only on volunteerism to mobilize participants was insufficient (we were able to collect too few completed surveys); and, ii) quality control (e.g., typos in the keywords) is much harder as there was no automatic way to recruit only participants that are fluent in English (all videos in the survey are in English).

Therefore, we published a task [12] in the Amazon Mechanical Turk. The task requires the 'turkers' to watch trailers, and provide the query terms they would use to search for the videos they have just watched (Figure 6.2). For each video, we collected answers from three 'turkers'. Turkers who accept the tasks are required to associate at least 3 keywords (and at most 10 keywords) to each video,

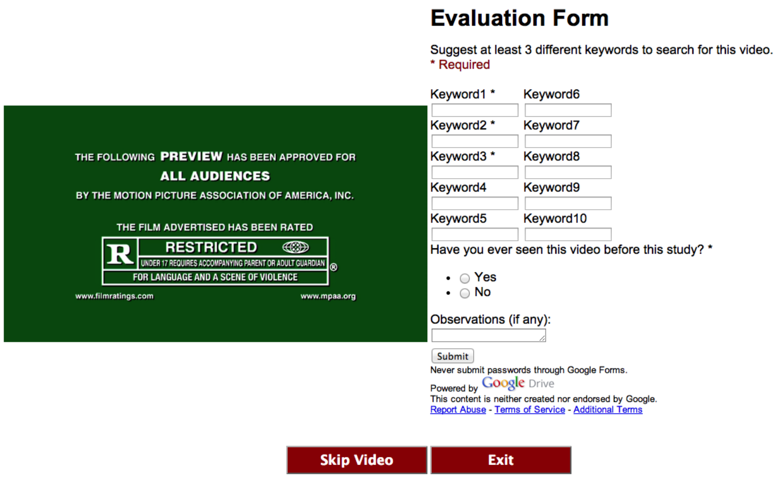
[9] Note that this work can easily be extended to other types of videos or content, as long as there is textual data available related to the content to produce candidate tags.

[10] developer.nytimes.com/docs

[11] movielens.org

[12] A similar form to that used in the AMT is available at: http://goo.gl/HZiUSw

as queries are typically of that length [36]. Each participant is paid $0.30 per task assignment with completion time of 6min (leading a total cost of $345 to conduct the survey. We followed AMT pay guidelines). This amounts to a hourly rate of $3/*hour*, which is way cheaper than the wage paid to dedicated ‘channel managers’.


THE FOLLOWING PREVIEW HAS BEEN APPROVED FOR
ALL AUDIENCES
BY THE MOTION PICTURE ASSOCIATION OF AMERICA, INC.
THE FILM ADVERTISED HAS BEEN RATED
R
RESTRICTED
UNDER 17 REQUIRES ACCOMPANYING PARENT OR ADULT GUARDIAN
FOR LANGUAGE AND A SCENE OF VIOLENCE
www.filmratings.com
www.mpaa.org
Skip Video
Exit

Evaluation Form

Suggest at least 3 different keywords to search for this video.
* Required

Keyword1 * Keyword6
Keyword2 * Keyword7
Keyword3 * Keyword8
Keyword4 Keyword9
Keyword5 Keyword10

Have you ever seen this video before this study? *

- ○ Yes
- ○ No

Observations (if any):

Submit

Never submit passwords through Google Forms.
Powered by Google Drive
This content is neither created nor endorsed by Google.
Report Abuse - Terms of Service - Additional Terms

Figure 6.2: A screenshot of the survey we set up on Amazon Mechanical Turk: turkers watch the video presented on the left side, enter the suggested keywords, answer the questions, and move on to the next video.

We also perform simple quality control by inspecting each answer to avoid accepting spam (which is expected to be low, due to the reputation mechanism adopted by the system). In fact, only one submission was rejected because the turker submitted URLs instead of keywords, and they had nothing to do with the video.

A brief characterization of the ground truth. In total, 33 turkers submitted solutions. Figure 6.3 shows the number of videos evaluated per turker: as we can observe, 58% of the turkers evaluated more than 5 videos, with the maximum reaching 333 videos evaluated by one turker. Figure 6.4 shows a histogram of the number of different keywords each video received. Even though we asked the turkers to associate at least 3 keywords to each video, 82% of the evaluations

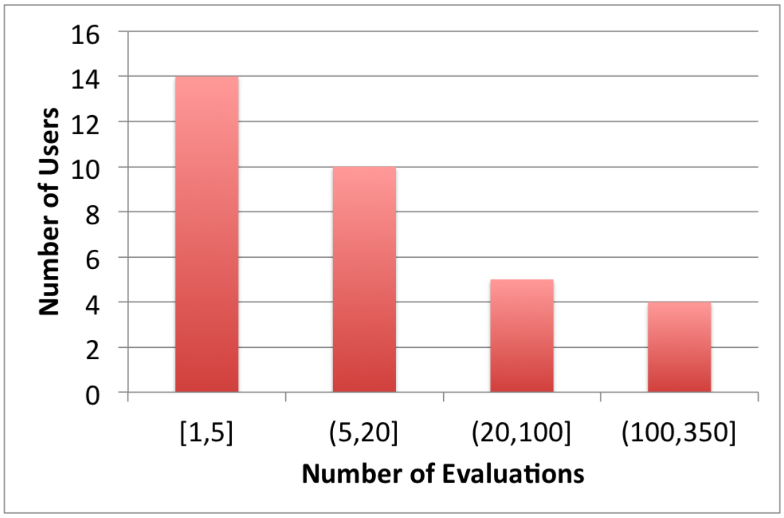


Figure 6.3: Histogram of the number of evaluations turkers have performed

provided more than the required minimum, which resulted in 96% of the videos with 10 or more different associated keywords.

Figure 6.5 presents the histogram of the total number of characters in the set of unique keywords associated to each video. The length of the ground truth varies from 51 (min) to 264 (max) characters; in fact, 32% of the videos have up to 100 characters. These values guide the budget parameter in our experiments as we explain in Section 6.5.

6.4 Experimental Setup

This section presents the instances of data sources and recommenders, and the metrics used in the evaluation.

6.4.1 Data Sources

We focus on comparing the effectiveness of using peer- and expert-produced data sources as input to recommender algorithms in the context of content promotion. The position of a data source in this spectrum (Figure 6.6) depends on whether

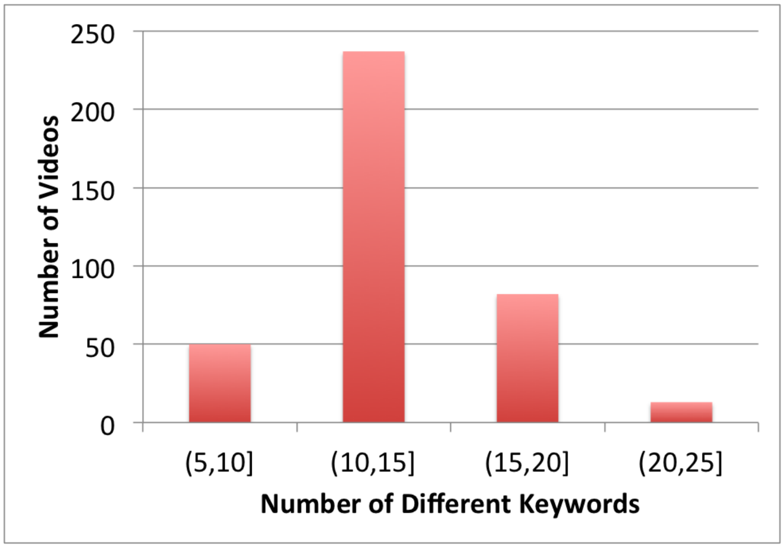


Figure 6.4: Histogram of the number of different keywords associated to a video by turkers

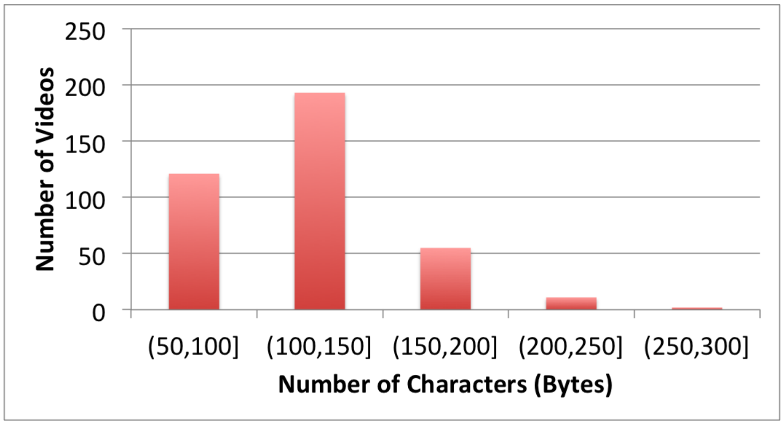


Figure 6.5: Histogram of the total length (in characters) for the set of distinct keywords associated to each video.

the data is produced collaboratively by non-experts or by a single expert user. For example, the page of a film in Wikipedia is likely edited by many non-expert users, while the reviews published by NYTimes are generally authored by a single movie critic. Next, we describe each of these data sources:

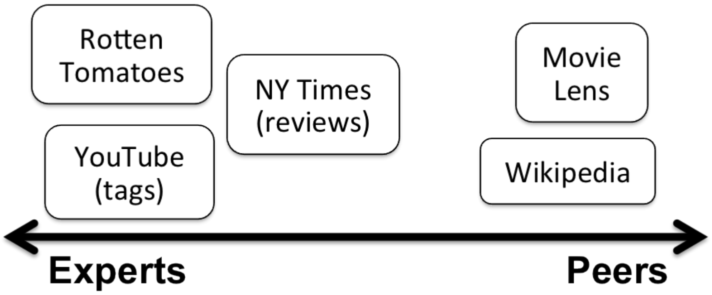


Figure 6.6: An illustration of the space of data sources we explore.

MovieLens is a web system where users collaboratively build and maintain a catalog of movies and their ratings. Users can create new movie entries; update existing entries; annotate movies with tags; review and rate movies. Based on previous users' activity and ratings, MovieLens suggests movies a user may like to watch.

For our evaluation we use only some of the data available in MovieLens: only the tags users produce while collaboratively annotate and bookmark movies. This data is a trace of tag assignments made available on the Web [13]

*Wikipedia*is a peer-produced encyclopedia where users collaboratively write articles about a multitude of topics. Users in *Wikipedia*also collaboratively edit and maintain pages for specific movies [14]. We leverage these pages as the sources of candidate keywords for recommending tags for their respective movies from our sample.

NYTimes reviews are written by movie critics who can be considered experts on the subject. Similar to the data provided by *Wikipedia*, we leverage the review page of a particular movie as the source of candidate keywords for the tag recommendation. The reviews are collected via the query interfaces [15] the New York Times API.

Rotten Tomatoes is a portal where users can rate and review movies, and, in

[13] http://www.grouplens.org/taxonomy/term/14

[14] E.g., http://en.wikipedia.org/wiki/Pulp_Fiction_(Film)

[15] http://developers.nytimes.com/

addition they have access to all credits information: actors and roles, directors, producers, soundtrack, synopsis, etc. The portal links to critics‘ reviews as well. The information about the credits of a movie and the critics’ reviews can be considered as produced by experts (likely the film credits are obtained directly from the producers, while the critics’ reviews are similar to those from NYTimes). While users can review the movies as well (and this qualifies as peer-produced information), these reviews are available on the website, but not accessible via the API at the time of our investigation. The rest of the information about the movies together with links to the experts’ reviews is available via the Rotten Tomatoes API [16]. Therefore, in this investigation, the data used from this data source lies in the expert end of the spectrum. In the experiments, we divide Rotten Tomatoes into two data sources: Rotten Tomatoes (with the credit information); and, RT Reviews (with the critics’ reviews).

YouTube. Finally, to test whether the tags already assigned to *YouTube* videos can be further optimized, we collect the tags assigned to the *YouTube* videos in our sample from the HTML source of each video’s page. The reason for using page scraping rather than API requests is that videos’ tags are accessible via the API only to the video publisher, even though these tags are still used by the search engine to match queries and are available in the HTML of the video page. *YouTube* data source figures in the expert-produced end of the spectrum, because only the publisher can assign tags to the video. Moreover, it is reasonable to assume that a video’s publisher is an expert on that video and aims to optimize its textual features to attract more views.

6.4.2 Recommenders

The experiments use two tag recommendation algorithms that process the input provided by the data sources. In particular, we use FREQUENCY and RANDOMWALK. We selected these two recommendation algorithms primarily because they harness some fundamental aspects of the tag recommendation problem

[16] http://developer.rottentomatoes.com/

that more sophisticated methods (e.g., [7, 53, 84]) also use (i.e., tag frequency, and tag co-occurrence patterns). Moreover, our goal is to understand the relative influence of the data sources on the quality of tags recommended. We note that, the methodology we describe and the ground truth can be used to evaluate other, more sophisticated, recommender algorithms as well.

The FREQUENCY recommender scores the candidate keywords based on how often each keyword appears in the data provided by a data source. Given the movie title, our pipeline finds the documents in the data source that match the title and extract a list of candidate keywords. For example, in *Wikipedia*, the candidate keywords for recommendation to a given movie are extracted from the *Wikipedia*page about the movie. Hence, the frequency of provided by a keyword is the number of times the keyword appears in that page. Similarly, in MovieLens, the frequency is the number of times a tag is assigned to a movie.

The RANDOMWALK recommender harness both the frequency and the co-occurrence between keywords. The co-occurrence is detected differently depending on the data source. In MovieLens, two keywords co-occur if they are assigned to the movie by the same user, for example, while in NY Times, Rotten Tomatoes, and Wikipedia two keywords co-occur if they appear in the same page related to the movie (i.e., review, movie record, and movie page, respectively). The RANDOMWALK recommender builds a graph based on keyword co-occurrence, where each keyword is a node and an edge connects two keywords if they co-occur. The initial score of each node is proportional to the individual frequency of each keyword as obtained from the data source. The RANDOMWALK is executed until convergence and the final node scores are used to rank the candidate keywords [18, 53, 95].

6.4.3 Budget Adjustment

To make the comparison fairer, for each movie, we adjust the budget to the size of the tag set in the ground truth. The knapsack solver uses this budget to select the recommended tags for a particular video. The reason for using a budget per video is that by using a budget larger than the ground truth for a video the F3-measure (see definition below), for example, is penalized by definition, as the number of recommended tags will be always larger than the ground truth size.

6.4.4 Success Metrics

The final step in the experiment is to estimate, for each video and for various input data sources and recommender algorithms, the quality of the recommended tag-set. To this end, we use multiple metrics to compare the ground truth with the recommended tag-set: the F3-measure, generalized distance [25], and the Normalized Discounted Cumulative Gain (NDCG). We present each of these in turn.

Let T_v and S_v be the set of distinct words in the ground truth and the recommended tag-set, respectively, for video v. The metrics are defined as follows:

- **F3-measure**. This metric is defined as $F_3(v) = \frac{10P(v)R(v)}{9P(v)+R(v)}$ for video v, where $P(v)$ is the precisions and $R(v)$ is the recall. This metric, however, weighs all tags in the ground truth equally, and thus ignores one important piece of information: in some cases multiple 'turkers' suggested the same tag, a strong indication that the tag has higher value. To account for this we use:

- **Generalized τ distance** [25]. This metric allows the comparison between two ranked lists. Given a video v, we use this metric to compare T_v (ground truth) sorted by frequency (i.e., number of turkers who assigned the tag to the video) and S_v (recommended set of tags) sorted by the recommender score function. Similar to the traditional Kendall τ distance, the generalized τ distance counts the number of permutations needed to transform one of the lists into the other one, while relaxing the constraint that the two list have to contain the same elements. The extension is done by introducing a penalty parameter to account for elements that are in one list, but absent in the other. This metric however, weighs equally all order inversions, regardless of whether they are at the top or at the bottom. To compensate for this we use:

- **Normalized Discounted Cumulative Gain (NDCG)**. This metric introduce a discount factor that penalizes order changes at the top of the ranked list. Given a video v, this metric is computed as follows:

  $$NDCG(T_v, S_v) = \frac{\sum_{j=1}^{|S_v|} \frac{2^{f(w_j,v)-1}}{log(j+1)}}{\sum_{i=1}^{|T_v|} \frac{2^{f(k_i,v)-1}}{log(i+1)}}$$

  where $f(\cdot, v)$ is the frequency of a tag in the ground truth (i.e., number of turkers who assigned the tag to the video); i and j are the positions of a tag in the ground truth and in the recommended set of tags, respectively. Note that if a tag $w \in S_v$ and $w \notin T_v$, we consider $f(w, v) = 0$.

6.5 Experimental Results

This section presents our experimental results to address the research questions that guide this study. First, we compare the performance of tags already assigned to *YouTube* videos in our sample to the effectiveness of both FREQUENCY and RANDOMWALK recommenders when using input from all data sources (Section 6.5.1). Next, we look into the performance of individual data sources to understand the influence that each one has in the recommendation performance (Section 6.5.2). To complement the comparison of individual data sources, we compare two sets of combined data sources that represent the two ends of the spectrum we study (Section 6.5.3). Finally, we perform a characterization to identify factors that may explain the observed performance of some peer-produced data sources (Section 6.5.4).

6.5.1 Are Tags Assigned to Videos Optimized?

The first experiment assesses the value of tags already assigned to videos on *YouTube* for boosting their popularity. To this end, we compare the tags to the ground truth of each video. If the tags are already optimized they should show a

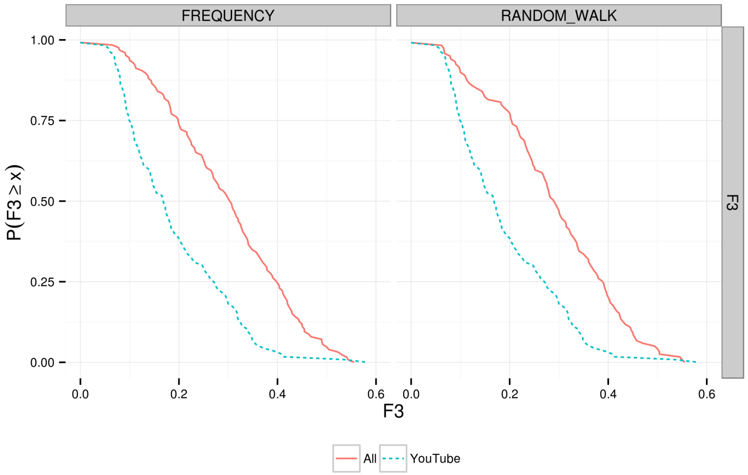


Figure 6.7: CCDF of F3-measure for *YouTube* tags (dashed line) compared to recommendation based on input from all other data sources combined (continuous line) using FREQUENCY (left) and RANDOMWALK (right) recommenders.

large overlap with the keywords in the ground truth.

Figure 6.7 shows the performance of tags previously assigned to the videos and the performance achieved by recommenders using input from all data sources combined (MovieLens, Rotten Tomatoes, Wikipedia, and NY Times) and the performance of tags already assigned to the *YouTube* videos. The curves represent the Complementary Cumulative Distribution Function (CCDF) of F3-measure. A point in the curve indicates the percentage of videos (y-axis) for which the F3-measure is larger than x. The more to the right the curve is, the more concentrated around larger values of F3-measure the recommendation performance is.

In fact, the Kolmogorov-Smirnov test of significance confirms that the performance of either recommender when using All data sources is significantly higher than that achieved by the *YouTube* tags (FREQUENCY: $D^- = 0.44, p$-value $= 3.9 \times 10^{-16}$; RANDOMWALK: $D^- = 0.43, p$-value $= 5.5 \times 10^{-15}$) . These results show that tags recommended by both methods are better than those currently assigned to the videos on *YouTube*. Therefore, the tags assigned to the *YouTube*

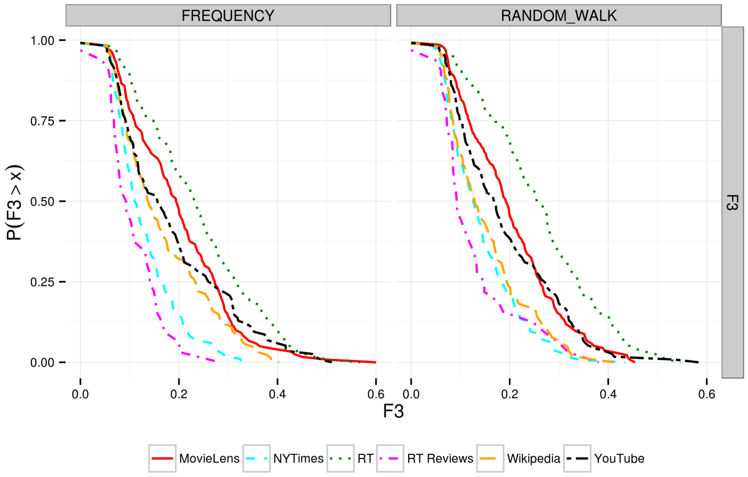


Figure 6.8: CCDF of F3-measure for each data source used as input for FREQUENCY (left) and RANDOMWALK (right) recommenders.

videos can still be improved towards attracting more search traffic, and, hence, more likely to boost popularity.

6.5.2 Is peer-produced information valuable?

The next experiment aims to assess the value of peer-produced versus expert-produced information, in the context of recommendation to improve content popularity. To this end, we compare the recommendation performance of different sources of candidate tags by fixing the recommender.

Figure 6.8 shows the CCDFs of the F3-measure for each individual data source as the input for the two recommenders. The first observation is that Rotten Tomatoes provide significant improvements over the existing tags on *YouTube*. Second, MovieLens is significantly better than the other three data sources NYTimes, RT Reviews, and Wikipedia, though MovieLens provides minor improvements on the currently assigned tags to videos on *YouTube*.

To put these results in perspective, we note that Rotten Tomatoes data source besides providing expert-produced information, it incorporates a schema for the information provided (i.e., actor names, character names, directors, i.e., named entities). Thus, one explanation for this good individual performance of Rotten Tomatoes is that users tend to use exactly names of entities related to the movie they are searching for. Therefore, by using an input that is rich of highly accurate named entities (i.e., entered by the movie producers), it is more likely that a recommender is successful.

In fact, we inspected the ground truth and after aggregating the top-5 most frequent keywords for each movie, around 50% of the top-10% most frequent keywords are named entities. Although it might be intuitive that accurate named entities improve recommendation, the observation that the MovieLens data source adds value, though minor, is an interesting observation. In particular, one would expect that the candidate keywords extracted from expert-produced reviews (NYTimes and RT Reviews) or peer-produced fact pages (Wikipedia) about the movie match what users would use to search. However, relative performance between MovieLens and the other three data sources suggests that candidate keywords produced via collaborative annotation is more effective than those produced by either collaborative text editing or produced by experts.

Figure 6.9 and Figure 6.10 show the performance of all data sources individually except the *YouTube* tags in terms of τ and NDCG metrics. The reason for removing *YouTube* tags is that the data obtained lacks ordering about the tags relative importance, and these two metrics compare the recommendations to the ground truth by considering the tags ranked according to the scores.

The results for the metrics τ and NDCG are qualitatively similar to those observed for F3-measure, as the relative order among data sources is kept unchanged – Rotten Tomatoes, *MovieLens*, *Wikipedia*, NYTimes, and RT Reviews, in order of the highest to the lowest performance. The only highlight that τ and NDCG bring is that the introduction of order in the comparison between the recommended tags and the ground truth widens the distance between Rotten Tomatoes and the other data sources.

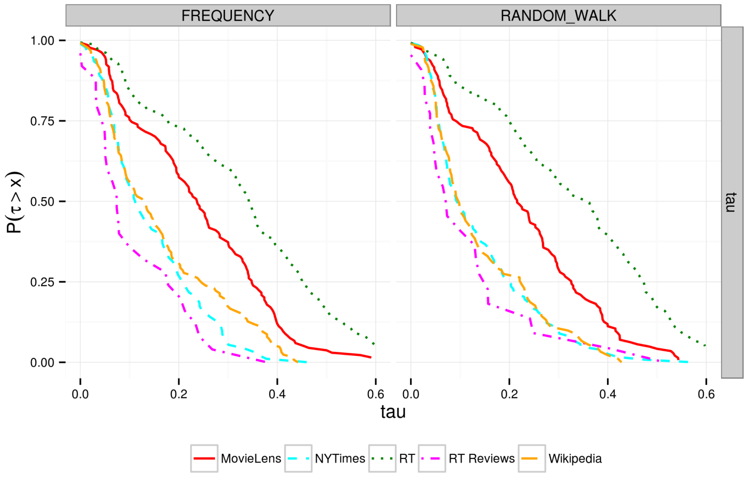


Figure 6.9: CCDF of τ for each data source used as input for FREQUENCY (left) and RANDOMWALK (right) recommenders.

6.5.3 Combining data sources

The previous experiments focused on the performance of individual data sources. In this section, we investigate the relative performance of combinations of data sources.

The goal is to understand whether each category of data sources leads to different performance levels. In particular, the experiment considers two groups – Peers: MovieLens + Wikipedia; Experts: NY Times Reviews + RT Reviews. Additionally, in these experiments, we use Rotten Tomatoes (which provides the credit information about the movies) and *YouTube* tags as baselines for comparison.

Figure 6.11 results lead to three observations: first, the CCDFs show that the performance of both recommenders using the Peers data source is significantly better than using the Experts data source; second, for the FREQUENCY recommender, the Peers data source performance is comparable to that of Rotten Tomatoes (which has the advantage of highly accurate named entity information to the movie, as discussed in the previous section); third, the Peers data source provides

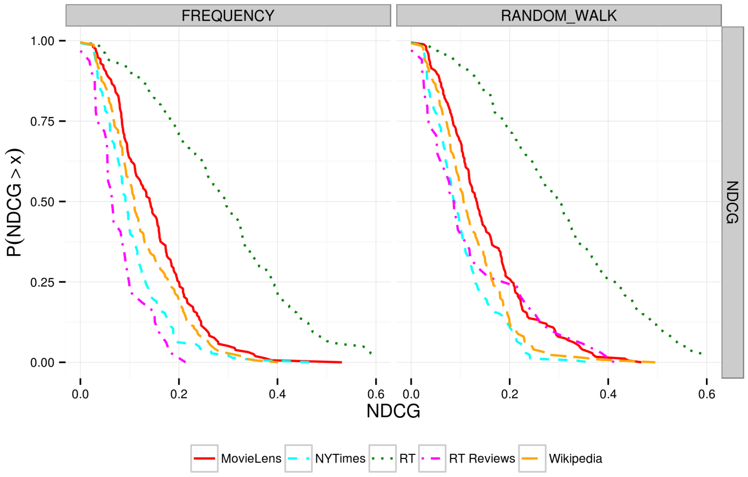


Figure 6.10: CCDF of NDCG for each data source used as input for FREQUENCY (left) and RANDOMWALK (right) recommenders.

significant improvement relative to the tags currently assigned to the *YouTube* videos, while for the RANDOMWALK recommender, there is no evidence of significant improvement.

Although Wikipedia alone leads the recommenders to poor performance, combining it with MovieLens seems encouraging, as the results for the FREQUENCY recommender using the Peers data source shows. The combination of candidate keywords produced by collaborative writing (Wikipedia) and collaborative annotations (MovieLens) seems to dilute important co-occurrence information that can be harnessed by the RANDOMWALK when using only MovieLens, as the relative performance between RANDOMWALK recommender with MovieLens (Figure 7) and the Peers compared to *YouTube* suggests (Figure 10).

Figure 6.12 and Figure 6.13 show similar results the performance of both recommenders using the Peers data source is significantly better than those using the Experts data source; and, while the performance of the FREQUENCY recommender using the Peers data source is comparable to Rotten Tomatoes, the RANDOMWALK performance is lower.

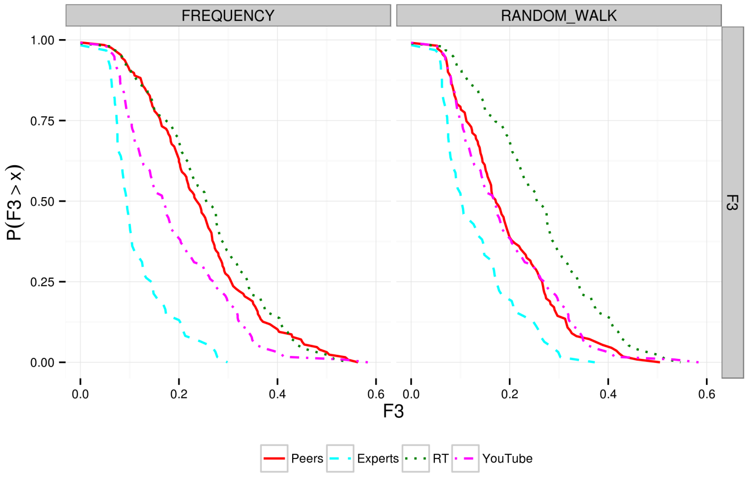


Figure 6.11: CCDF F3-measure performance comparison between combinations of groups of data sources Peers (MovieLens + Wikipedia) and Experts (NYTimes + RT Reviews) relative to *YouTube* and Rotten Tomatoes.

6.5.4 Is the number of contributors a factor?

As the previous result shows, the performance achieved by the peer-produced data sources vary widely across videos. This section investigates whether the number of peers that produce tags for a movie in the MovieLens data source has predictive power about the performance delivered by the data source in the recommendation (or how many peers is an expert worth?). To this end, we compute the correlation between the number of users who annotated a movie in MovieLens and the value of each quality metric for that video recommendation. The Spearman's rank correlation between the number of users and F3-measure of 0.31 indicates a mild positive correlation between these aspects.

Therefore, the number of contributors partially explains the value added by the MovieLens data source to the recommenders' performance. Yet, one potential reason for a lack of stronger correlation is that the motivation behind tagging a movie in MovieLens leads to drastically different terms from those used by users

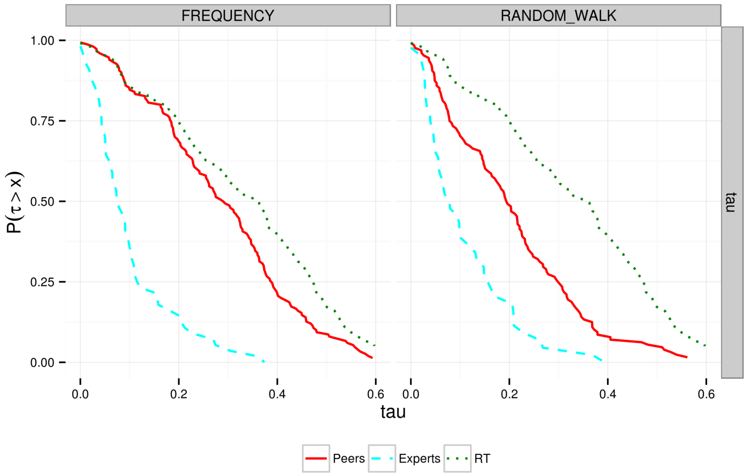


Figure 6.12: Tau performance comparison between combinations of data sources: Peers (MovieLens + Wikipedia) vs. Experts (NYTimes + RT Reviews).

searching for the video. For example, although ‘boring’ is a tag used to annotate movies in MovieLens as a way to express opinion about a movie to other users, it is unlikely that users searching for the same movie would use that term.

6.6 Summary

A large portion of traffic received by video content on the web is originated from keyword-based search and/or tag-based navigation. Consequently, the textual features of this content can directly impact the popularity of a particular content item, and ultimately the advertisement generated revenue. Therefore, understanding the performance of automatic tag recommenders is important to optimize the view count of content items.

First, this study confirms that tags currently assigned to a sample of *YouTube* videos can be further improved regarding their ability to attract more search traffic. Next, we perform comparisons between different types of data sources (peer- and expert-produced) with the goal of understanding the relative value of data sources

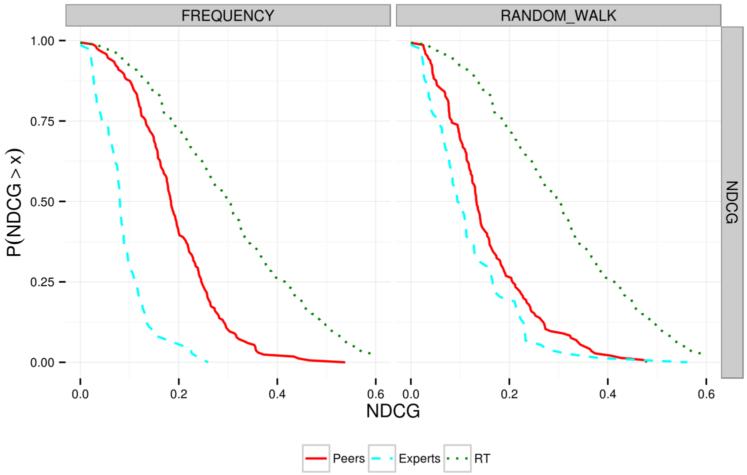


Figure 6.13: NDCG performance comparison between combinations of data sources: Peers (MovieLens + Wikipedia) vs. Experts (NYTimes + RT Reviews).

and combinations thereof, where we find that combinations of peer-produced data sources can add value compared to the best expert-based baseline. Finally, our experiments show that the number of contributors in a peer-produced data source partially explains its positive influence on the performance of tag recommendation for boosting content popularity.

Chapter 7

Conclusions

The study presented in this dissertation consists of a characterization of social tagging systems to inform the design of supporting mechanisms. In particular, this research focuses on the value of peer produced information in two distinct contexts: *exploratory search* and *content promotion*. In the former, the study presents the design and evaluation of a method to assess the value of tags from the information seeker standpoint. In the latter, this study investigates the value of peer-produced information in the context of content promotion in social media websites via tag recommendation.

To this end, a combination of quantitative and qualitative methods have been applied. The methodology consists of the following parts: *i)* characterization of traces of activity of social tagging systems with the focus on understanding usage patterns (Chapter 2 and Chapter 3); *ii)* qualitative analysis of users' perception of tag value when performing exploratory search tasks (Chapter 4); *iii)* design and evaluation of a method to assess the value of tags in exploratory search tasks (Chapter 5); and, *iv)* a study of the value of peer-produced information for content promotion in social media.

The main contributions of this research can be summarized as follows (grouped by their respective research question):

Characterization of Individual User Activity

- **RQ1.1.** *Users tend to reuse tags already present in the system more often than they repeatedly tag existing items* [77, 78]. This finding supports the intuition that tags are primarily a content categorization instrument. Additionally, the results show that the difference between the levels of tag reuse and repeated item tagging vary across different systems. This observation suggests that features such as tag recommendation and the type of content play a role in the patterns of peer production of information in tagging systems.

- **RQ1.2.** *The tag vocabulary of a user can be approximated by a small portion of her activity* [79]. The experiments on the evolution of user tag vocabularies show that only to accurately approximate the characteristics of a tag vocabulary, only a small percentage of the initial tag assignments performed by a user is necessary. These observed results can applied in the context of applications that rely on activity similarity scores between users, for example, as it provides a way to reason about the trade offs between accuracy of a user activity profile and the computational cost of updating the similarity scores.

Characterization of Social User Activity

- **RQ2.1.** *The strength of implicit social ties is concentrated over small portion of user pairs*. Moreover, the observed strength of activity similarity between pairs of users are the result of shared interest as opposed to generated by chance. The distributions of activity similarity strength deviate significantly from those produced by a Random Null Model (RNM) [71]. This suggests that the implicit ties between users, as defined by their activity similarity levels, capture latent information about user relationships that may offer support for optimizing system mechanisms.

- **RQ2.2.** *The average strength of implicit ties is stronger for user pairs with* *explicit ties* [78]. This investigation analyzes the similarity between users according to their tagging activity and its relation to explicit indicators of collaboration. The results show that the users' activity similarity is concentrated on a small fraction of user pairs. Also, the observed distributions of users' activity similarity deviate significantly from those produced by a *Random Null Model* [71]. Finally, an analysis of the relationships between implicit relationships based on activity similarity and other more explicit relationships, such as co-membership in discussion groups, shows that user pairs that tag items in common have in average higher similarity in terms of co-membership in discussion groups.

Characterization of Users' Perception of Tag Value

To complement the quantitative characterization and to inform the design of methods that assess the value of tags, this research conducts a qualitative characterization of user' perception of tag value. A summary of the major findings in this investigation is presented below:

- **RQ3.** *Users perception of tag value in exploratory search is multidimensional and the key aspects that influence users' perception are: relevance of items retrieved and reduction of search space [81].* Based on a qualitative characterization of users' perception of tag value in the context of exploratory search, this study finds that the two most salient aspects that influence users' perception of tag value are: *ability to retrieve relevant content items* and *ability to reduce the search space*. These findings inform the design of a method that quantifies the value of tags automatically by taking into account the important aspects, which are identified by the qualitative analysis.

Methods to Assess Value of Peer-Produced Information

Finally, this research proposes new techniques that exploit the usage characteristics of tagging systems to improve their design. The next paragraphs briefly describe the contributions related to studying social tagging as commons-based peer production systems and the design of methods to assess the value of user contribution in these collaborative contexts. Chapter 5 and Chapter 6 distills the proposed approaches and results in details.

Important to note that there are two perspectives to the problem of assessing the value of peer-produced information in tagging systems: the consumer and the producer. The goal is to design methods that cater for each of these perspectives. For consumers, assessing the value of tags are considered in the context of exploratory search, while for producers, the method takes into account the ability of a tag to improve the viewership of content (e.g., a YouTube video).

- **RQ4.** *An information-theoretical approach to assess the value of tags for exploratory search provides accurate estimates of value as perceived by users.* A method that automatically quantifies the value of tags that caters for the two desirable properties in the context of exploratory search, as identified by the qualitative user study. A proof shows that the proposed method has desirable theoretical properties while quantifying these two aspects. Additionally, an experiment using real tagging data that shows that the proposed method accurately quantifies the value of tags according to users' perception.

- **RQ5.** *Peer-produced information, though lacking formal curation, has comparable value to that of expert-produced information sources when used for content promotion.* An analysis of online videos provides evidence that the tags associated with a sample of popular movie trailers can be optimized further by an automated process: either by incorporating human computing engines (e.g., Amazon Mechanical Turk) at a much lower cost than using dedicated *channel managers* (the current industry practice); or, at an even lower cost, by using recommender algorithms to harness textual produced by a multitude of data sources that are related to the video content. To this end, I perform a comparison of the effectiveness of using peer- and expert-produced sources of information as input for tag recommender that aim to boost content popularity.

These contributions are key to the understanding user behaviour in social tagging systems. More importantly, the characterization study together with the design of methods that assess the value of tags (as proposed in this research) can help the design of incentive mechanisms that aim to boost user participation. In fact, a method to assess the value of users contributions in social tagging systems is a key building block in the design of incentive mechanisms. Therefore, this research provides an important contribution to future research that pursue this direction.

Bibliography

[1] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman. Search in power-law networks. *Physical Review E*, 64(4), 2001. → pages

[2] G. Adomavicius and A. Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17 (6):734–749, 2005. → pages

[3] M. Ames and M. Naaman. Why we tag. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07*, CHI '07, page 971, New York, New York, USA, 2007. ACM Press. ISBN 9781595935939. → pages

[4] P. André, M. Bernstein, and K. Luther. Who gives a tweet? In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*, page 471, New York, New York, USA, Feb. 2012. ACM Press. ISBN 9781450310864. doi:10.1145/2145204.2145277. URL http://dl.acm.org/citation.cfm?id=2145204.2145277. → pages

[5] R. Baeza-Yates and L. Rello. On measuring the lexical quality of the web. In *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality - WebQuality '12*, page 1, New York, New York, USA, Apr. 2012. ACM Press. ISBN 9781450312370. doi:10.1145/2184305.2184307. URL http://dl.acm.org/citation.cfm?id=2184305.2184307. → pages

[6] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999. ISBN 020139829X. → pages

[7] F. Belém, E. Martins, J. Almeida, and M. Gonçalves. Exploiting Novelty and Diversity in Tag Recommendation. In P. Serdyukov, P. Braslavski,

S. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, and E. Yilmaz, editors, *Advances in Information Retrieval SE - 32*, volume 7814 of *Lecture Notes in Computer Science*, pages 380–391. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-36972-8. doi:10.1007/978-3-642-36973-5_32. URL http://dx.doi.org/10.1007/978-3-642-36973-5_32. → pages

[8] Y. Benkler. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press, May 2006. ISBN 0300110561. → pages

[9] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu. Can all tags be used for search? In *CIKM*, CIKM '08, pages 193–202, Napa Valley, California, USA, 2008. ACM. ISBN 978-1-59593-991-3. → pages

[10] A. Budanitsky and G. Hirst. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47, 2006. ISSN 0891-2017. → pages

[11] A. Capocci, A. Baldassarri, V. D. Servedio, and V. Loreto. Statistical properties of inter-arrival times distribution in social tagging systems. In *20th ACM International Conference on Hypertext*, pages 239–244, Torino, Italy, 2009. ACM. ISBN 978-1-60558-486-7. → pages

[12] C. Cattuto, A. Baldassarri, V. D. P. Servedio, and V. Loreto. Vocabulary growth in collaborative tagging systems. 2007. → pages

[13] C. Cattuto, A. Barrat, A. Baldassarri, G. Schehr, and V. Loreto. Collective dynamics of social annotation. *Proceedings of the National Academy of Sciences*, 106(26):10511–10515, June 2009. ISSN 1091-6490. → pages

[14] E. Chi, P. Pirolli, and S. Lam. Aspects of Augmented Social Cognition: Social Information Foraging and Social Search. In D. Schuler, editor, *Online Communities and Social Computing*, volume 4564 of *Lecture Notes in Computer Science*, pages 60–69. Springer Berlin Heidelberg, 2007. → pages

[15] E. H. Chi. Information Seeking Can Be Social. *Computer*, 42(3):42–46, Mar. 2009. ISSN 0018-9162. → pages

[16] E. H. Chi and T. Mytkowicz. Understanding the efficiency of social tagging systems using information theory. In *19th ACM International Conference on Hypertext*, pages 81–88, Pittsburgh, PA, USA, 2008. ACM. ISBN 978-1-59593-985-2. → pages

[17] P. A. Chirita, S. Costache, W. Nejdl, and S. Handschuh. P-TAG. In *Proceedings of the 16th international conference on World Wide Web - WWW '07*, page 845, New York, New York, USA, 2007. ACM Press. ISBN 9781595936547. doi:10.1145/1242572.1242686. URL http://portal.acm.org/citation.cfm?doid=1242572.1242686. → pages

[18] M. Clements, A. P. de Vries, and M. J. T. Reinders. Optimizing single term queries using a personalized markov random walk over the social graph. Mar. 2008. → pages

[19] M. Clements, A. P. de Vries, and M. J. Reinders. The task dependent effect of tags and ratings on social media access. In *ACM TOIS*, 2010. → pages

[20] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press, third edit edition, July 2009. ISBN 0262033844. → pages

[21] T. M. Cover and J. A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2 edition, July 2006. ISBN 0471241954. → pages

[22] M.-P. Dubuisson and A. Jain. A modified Hausdorff distance for object matching. In *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*, volume 1, pages 566–568, Jerusalem, Israel, 1994. IEEE Comput. Soc. Press. ISBN 0-8186-6265-4. → pages

[23] F. Eggenberger and G. Polya. Ueber die Statistik verketteter Vorgaenge. *Zeit. Angew. Math. Mech*, 3(4):279–289, 1923. → pages

[24] B. M. Evans, S. Kairam, and P. Pirolli. Exploring the cognitive consequences of social search. CHI EA '09, pages 3377–3382, Boston, MA, USA, 2009. ACM. ISBN 978-1-60558-247-4. → pages

[25] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. In *SODA'03*, pages 28–36, Baltimore, Maryland, 2003. Society for Industrial and Applied Mathematics. ISBN 0-89871-538-5. → pages

[26] F. Figueiredo, F. Belém, H. Pinto, J. Almeida, M. Gonçalves, D. Fernandes, E. Moura, and M. Critso. Evidence of Quality of Textual Features on the Web 2.0. In *CIKM*, 2009. → pages

[27] F. Figueiredo, H. Pinto, F. Belém, J. Almeida, M. Gonçalves, D. Fernandes, and E. Moura. Assessing the quality of textual features in social media. *Information Processing & Management*, 49(1):222–247, Jan. 2013. ISSN 03064573. doi:10.1016/j.ipm.2012.03.003. URL http://dx.doi.org/10.1016/j.ipm.2012.03.003. → pages

[28] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Commun. ACM*, 30 (11):964–971, 1987. ISSN 0001-0782. → pages

[29] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, Apr. 2006. ISSN 0165-5515. → pages

[30] O. Görlitz, S. Sizov, and S. Staab. PINTS: Peer-to-Peer Infrastructure for Tagging Systems. In *Procdings of the 7th International Conference on Peer-to-Peer Systems*, 2008. → pages

[31] X. Gu, X. Wang, R. Li, K. Wen, Y. Yang, and W. Xiao. Measuring Social Tag Confidence: Is It a Good or Bad Tag? 6897, 2011. → pages

[32] Z. Guan, J. Bu, Q. Mei, C. Chen, and C. Wang. Personalized tag recommendation using graph-based ranking on multi-type interrelated objects. SIGIR '09, pages 540–547, Boston, MA, USA, 2009. ACM. ISBN 978-1-60558-483-6. → pages

[33] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *16th International World Wide Web Conference*, WWW '07, pages 211–220, Banff, Alberta, Canada, 2007. ACM. ISBN 978-1-59593-654-7. → pages

[34] T. Hammond, T. Hannay, B. Lund, and J. Scott. Social bookmarking tools (i): A general review. *D-Lib Magazine*, 11(4), April 2005. ISSN 1082-9873. → pages

[35] M. Harvey, I. Ruthven, and M. J. Carman. Improving social bookmark search using personalised latent variable language models. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 485–494, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0493-1. → pages

[36] B. He and I. Ounis. Query performance prediction. *Information Systems*, 31(7):585–594, Nov. 2006. ISSN 03064379. doi:10.1016/j.is.2005.11.003. URL http://linkinghub.elsevier.com/retrieve/pii/S0306437905000955. → pages

[37] M. Heckner, M. Heilemann, and C. Wolff. Personal information management vs. resource sharing: Towards a model of information behaviour in social tagging systems. In *Proc. the Third International AAAI Conference on Weblogs and Social Media (ICWSM 09)*, 2009. → pages

[38] D. Helic, M. Strohmaier, C. Trattner, M. Muhr, and K. Lerman. Pragmatic evaluation of folksonomies. WWW '11, pages 417–426, Hyderabad, India, 2011. ACM. ISBN 978-1-4503-0632-4. → pages

[39] M. Hennik, I. Hutter, and A. Bailey. *Qualitative Research Methods*. SAGE Publications, 1 edition, 2011. → pages

[40] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 195–206, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-927-2. → pages

[41] P. Heymann, A. Paepcke, and H. G. Molina. Tagging human knowledge. In *WSDM*, WSDM '10, pages 51–60, New York, New York, USA, 2010. ACM. ISBN 978-1-60558-889-6. → pages

[42] J. Hirshleifer. Where Are We in the Theory of Information? *The American Economic Review*, 63(2):31–39, 1973. ISSN 00028282. → pages

[43] A. Hotho, R. Jschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In *Proceedings of the 3rd European*

Semantic Web Conference, volume 4011 of *LNCS*, pages 411–426, Budva, Montenegro, June 2006. Springer. ISBN 3-540-34544-2. → pages

[44] S. Huang, X. Wu, and A. Bolivar. The effect of title term suggestion on e-commerce sites. WIDM '08, pages 31–38, Napa Valley, California, USA, 2008. ACM. ISBN 978-1-60558-260-3. → pages

[45] A. Iamnitchi, M. Ripeanu, and I. Foster. Small-world file-sharing communities. In *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, volume 2, pages 952–963, Hong Kong, China, 2004. → pages

[46] A. Iamnitchi, M. Ripeanu, E. S. Neto, and I. Foster. The Small World of File Sharing. *IEEE Transactions on Parallel and Distributed Systems*, 22: 1120–1134, 2011. ISSN 1045-9219. → pages

[47] P. Jaccard. The Distribution of the Flora in the Alpine Zone. *New Phytologist*, 11(2):37–50, 1912. → pages

[48] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag Recommendations in Folksonomies. In J. Kok, J. Koronacki, R. Lopez de Mantaras, S. Matwin, D. Mladenic, and A. Skowron, editors, *Knowledge Discovery in Databases: PKDD 2007*, volume 4702 of *Lecture Notes in Computer Science*, book part (with own title) 52, pages 506–514. Springer Berlin / Heidelberg, Warsaw, Poland, 2007. ISBN 978-3-540-74975-2. → pages

[49] Y. Kammerer, R. Nairn, P. Pirolli, and E. H. Chi. Signpost from the masses: learning effects in an exploratory social tag search browser. In *CHI*, CHI '09, pages 625–634, Boston, MA, USA, 2009. ACM. ISBN 978-1-60558-246-7. → pages

[50] S. Kashoob and J. Caverlee. Temporal dynamics of communities in social bookmarking systems. *Social Network Analysis and Mining*, 2:387–404, 2012. ISSN 1869-5450. → pages

[51] M. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1/2): 81–93, June 1938. ISSN 00063444. → pages

[52] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse. Identifying influential spreaders in complex networks. *http://arxiv.org/abs/1001.5285*, 2010. → pages

[53] I. Konstas, V. Stathopoulos, and J. M. Jose. On social networks and collaborative recommendation. In *SIGIR*, SIGIR '09, pages 195–202, Boston, MA, USA, 2009. ACM. ISBN 978-1-60558-483-6. → pages

[54] G. Koutrika, F. A. Effendi, Z. Gyöngyi, P. Heymann, and H. G. Molina. Combating spam in tagging systems: An evaluation. *ACM Trans. Web*, 2 (4):1–34, 2008. ISSN 1559-1131. → pages

[55] B. Krause, C. Schmitz, A. Hotho, and G. Stumme. The anti-social tagger: detecting spam in social bookmarking systems. pages 61–68, Beijing, China, 2008. ACM. ISBN 978-1-60558-159-0. → pages

[56] R. Krestel and P. Fankhauser. Language Models and Topic Models for Personalizing Tag Recommendation. pages 82–89, Toronto, AB, Canada, Aug. 2010. → pages

[57] C. Lampe, J. Vitak, R. Gray, and N. Ellison. Perceptions of facebook's value as an information source. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, page 3195, New York, New York, USA, May 2012. ACM Press. ISBN 9781450310154. doi:10.1145/2207676.2208739. URL http://dl.acm.org/citation.cfm?id=2207676.2208739. → pages

[58] P. Li, B. Wang, W. Jin, J. Y. Nie, Z. Shi, and B. He. Exploring categorization property of social annotations for information retrieval. CIKM '11, pages 557–562, Glasgow, Scotland, UK, 2011. ACM. ISBN 978-1-4503-0717-8. → pages

[59] M. Lipczak and E. Milios. Efficient Tag Recommendation for Real-Life Data. *ACM Transactions on Intelligent Systems and Technology*, 3(1): 1–21, Oct. 2011. ISSN 21576904. doi:10.1145/2036264.2036266. URL http://dl.acm.org/citation.cfm?doid=2036264.2036266. → pages

[60] D. Liu, X. S. Hua, L. Yang, M. Wang, and H. J. Zhang. Tag ranking. WWW '09, pages 351–360, Madrid, Spain, 2009. ACM. ISBN 978-1-60558-487-4. → pages

[61] C. Lu, J.-R. Park, and X. Hu. User tags versus expert-assigned subject terms: A comparison of LibraryThing tags and Library of Congress Subject Headings. *Journal of Information Science*, 6(6):763–779, Dec. 2010. → pages

[62] L. B. Marinho and L. Schmidt-Thieme. Collaborative Tag Recommendations. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, editors, *Data Analysis, Machine Learning and Applications*, book chapter/section Chapter 63, pages 533–540. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-78239-1. → pages

[63] C. Marlow, M. Naaman, D. Boyd, and M. Davis. HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *Proceedings of the seventeenth conference on Hypertext and hypermedia - HYPERTEXT '06*, page 31, New York, New York, USA, 2006. ACM Press. ISBN 1595934170. → pages

[64] E. S. Neto, S. A. Kiswany, N. Andrade, S. Gopalakrishnan, and M. Ripeanu. enabling cross-layer optimizations in storage systems with custom metadata. In *International Conference on High Performance Computing - HotTopics*, pages 213–216, Boston, MA, USA, 2008. ACM. ISBN 978-1-59593-997-5. → pages

[65] E. S. Neto, F. Figueiredo, J. Almeida, M. Mowbray, M. Gonçalves, and M. Ripeanu. Assessing the Value of Contributions in Tagging Systems. *Social Computing / IEEE International Conference on Privacy, Security, Risk and Trust, 2010 IEEE International Conference on*, 0:431–438, Aug. 2010. → pages

[66] O. Nov, M. Naaman, and C. Ye. What drives content tagging: the case of photos on Flickr. In *26th Annual SIGCHI Conference on Human factors in computing systems*, pages 1097–1100, Florence, Italy, 2008. ACM. ISBN 978-1-60558-011-1. → pages

[67] E. Ostrom. *Governing the Commons : The Evolution of Institutions for Collective Action*. Political Economy of Institutions and Decisions. Cambridge University Press, Nov. 1990. ISBN 0521405998. → pages

[68] T. Pedersen, S. Patwardhan, and J. Michelizzi. WordNet::Similarity: measuring the relatedness of concepts. In *In Demonstration Papers At*

HLT-NAACL on XX Human Language Technology Conference. Association for Computational Linguistics, pages 38–41, Boston, Massachusetts, 2004. → pages

[69] P. L. T. Pirolli. *Information Foraging Theory: Adaptive Interaction with Information (Oxford Series in Human-Technology Interaction).* Oxford University Press, USA, 1 edition, Apr. 2007. ISBN 0195173325. → pages

[70] R. Ramakrishnan and A. Tomkins. Toward a PeopleWeb. *Computer*, 40 (8):63–72, 2007. → pages

[71] J. Reichardt and S. Bornholdt. Market Segmentation: The Network Approach. In *Managing Complexity: Insights, Concepts, Applications*, pages 19–36. 2008. → pages

[72] S. Rendle and L. S. Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. pages 81–90, New York, New York, USA, 2010. ACM. ISBN 978-1-60558-889-6. → pages

[73] A. J. Repo. The dual approach to the value of information: An appraisal of use and exchange values. *Information Processing and Management*, 22 (5):373 – 383, 1986. ISSN 0306-4573. URL http://www.sciencedirect.com/science/article/pii/0306457386900725. → pages

[74] B. D. Ruben. *Information as an economic good: a reevaluation of theoretical approaches*. Information and behavior series, 3. Transaction Publ., New Brunswick, NJ [u.a.], 1990. ISBN 0887382789 9780887382789. → pages

[75] E. Santos-Neto. Characterizing and harnessing peer-production of information in social tagging systems. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 761–762, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-0747-5. → pages

[76] E. Santos-Neto, M. Ripeanu, and A. Iamnitchi. Tracking User Attention in Collaborative Tagging Communities. In *International ACM/IEEE Workshop on Contextualized Attention Metadata: Personalized Access to Digital Resources*, pages 11–18, Vancouver, June 2007. CEUR-WS.org. → pages

[77] E. Santos-Neto, M. Ripeanu, and A. Iamnitchi. Content Reuse and Interest Sharing in Tagging Communities. In *Proceeding of the AAAI Spring Symposium on Social Information Processing (AAAI-SIP 2008)*, Stanford, Mar. 2008. → pages

[78] E. Santos-Neto, D. Condon, N. Andrade, A. Iamnitchi, and M. Ripeanu. Individual and social behavior in tagging systems. In *Hypertext*, HT '09, pages 183–192, Torino, Italy, 2009. ACM. ISBN 978-1-60558-486-7. → pages

[79] E. Santos-Neto, D. Condon, N. Andrade, A. Iamnitchi, and M. Ripeanu. Reuse, Temporal Dynamics, Interest Sharing, and Collaboration in Social Tagging Systems. Jan. 2013. URL http://arxiv.org/abs/1301.6191. → pages

[80] E. Santos-Neto, F. Figueiredo, N. Oliveira, N. Andrade, J. Almeida, and M. Ripeanu. Assessing value of peer-produced information for exploratory search. In *Technical Report. Submitted to WWW'2014*, October 2013. → pages

[81] E. Santos-Neto, T. Pontes, J. Almeida, and M. Ripeanu. How many peers is an expert worth? on the value of information sources for boosting content popularity via tag recommendation. In *Technical Report. Submitted to WSDM'2014*, October 2013. → pages

[82] K. Seki, H. Qin, and K. Uehara. Impact and prospect of social bookmarks for bibliographic information retrieval. JCDL '10, pages 357–360, Gold Coast, Queensland, Australia, 2010. ACM. ISBN 978-1-4503-0085-8. → pages

[83] S. Sen, S. K. Lam, A. M. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, F. M. Harper, and J. Riedl. tagging, communities, vocabulary, evolution. In *15th International World Wide Web Conference*, CSCW '06, pages 181–190, Banff, Alberta, Canada, 2006. ACM. ISBN 1-59593-249-6. → pages

[84] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *17th International World Wide Web Conference*, WWW '08, pages 327–336, Beijing, China, 2008. ACM. ISBN 978-1-60558-085-2. → pages

[85] H. A. Simon. On a Class of Skew Distribution Functions. *Biometrika*, 42 (3/4):425–440, 1955. ISSN 00063444. → pages

[86] J. Sinclair and M. Cardew-Hall. The folksonomy tag cloud: when is it useful? *Journal of Information Science*, 34(1):15–29, Feb. 2008. ISSN 1741-6485. → pages

[87] M. A. Stephens. EDF Statistics for Goodness of Fit and Some Comparisons. *Journal of the American Statistical Association*, 69(347): 730, Sept. 1974. ISSN 01621459. doi:10.2307/2286009. URL http://www.jstor.org/stable/2286009?origin=crossref. → pages

[88] G. J. Stigler. *The Organization of Industry*. University Of Chicago Press, Mar. 1983. ISBN 0226774325. → pages

[89] J. Stoyanovich, S. A. Yahia, C. Marlow, and C. Yu. Leveraging Tagging to Model User Interests in del.icio.us. In *Proceeding of the AAAI Spring Symposium on Social Information Processing (AAAI-SIP 2008)*, Stanford, 2008. → pages

[90] M. Strohmaier, C. Krner, and R. Kern. Understanding why users tag: A survey of tagging motivation literature and results from an empirical study. *Web Semantics: Science, Services and Agents on the World Wide Web*, 17 (0), 2012. ISSN 1570-8268. → pages

[91] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web - WWW '07*, WWW '07, pages 697–706, Banff, Alberta, Canada, 2007. ACM. ISBN 978-1-59593-654-7. → pages

[92] F. M. Suchanek, M. Vojnovic, and D. Gunawardena. Social tags: meaning and suggestions. In *CIKM*, CIKM '08, pages 223–232, Napa Valley, California, USA, 2008. ACM. ISBN 978-1-59593-991-3. → pages

[93] P. K. Vatturi, W. Geyer, C. Dugan, M. Muller, and B. Brownholtz. Tag-based filtering for personalized bookmark recommendations. pages 1395–1396, Napa Valley, California, USA, 2008. ACM. ISBN 978-1-59593-991-3. → pages

[94] P. Venetis, G. Koutrika, and H. G. Molina. On the selection of tags for tag clouds. WSDM '11, pages 835–844, Hong Kong, China, 2011. ACM. ISBN 978-1-4503-0493-1. → pages

[95] C. Wang, F. Jing, L. Zhang, and H. J. Zhang. Image annotation refinement using random walk with restarts. MULTIMEDIA '06, pages 647–650, Santa Barbara, CA, USA, 2006. ACM. ISBN 1-59593-447-2. → pages

[96] J. Wang, M. Clements, J. Yang, A. P. de Vries, and M. J. Reinders. Personalization of tagging systems. *Information Processing & Management*, 46(1):58–70, Jan. 2010. ISSN 03064573. → pages

[97] J. Weng, E. P. Lim, J. Jiang, and Q. He. TwitterRank: finding topic-sensitive influential twitterers. In *WSDM*, WSDM '10, pages 261–270, New York, New York, USA, 2010. ACM. ISBN 978-1-60558-889-6. → pages

[98] S. A. Yahia, M. Benedikt, L. V. S. Lakshmanan, and J. Stoyanovich. Efficient network aware search in collaborative tagging sites. *Proc. VLDB Endow.*, 1(1):710–721, 2008. ISSN 2150-8097. → pages

[99] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka. Can social bookmarking enhance search in the web? JCDL '07, pages 107–116, Vancouver, BC, Canada, 2007. ACM. ISBN 978-1-59593-644-8. → pages

[100] D. Yin, Z. Xue, L. Hong, and B. D. Davison. A probabilistic model for personalized tag prediction. pages 959–968, Washington, DC, USA, 2010. ACM. ISBN 978-1-4503-0055-1. → pages

[101] D. Zhou, J. Bian, S. Zheng, H. Zha, and C. L. Giles. Exploring social annotations for information retrieval. In *17th International World Wide Web Conference*, pages 715–724, Beijing, China, 2008. ACM. ISBN 978-1-60558-085-2. → pages

[102] R. Zhou, S. Khemmarat, L. Gao, and H. Wang. Boosting video popularity through recommendation systems. In *Databases and Social Networks on - DBSocial '11*, pages 13–18, New York, New York, USA, June 2011. ACM Press. ISBN 9781450306508. doi:10.1145/1996413.1996416. URL http://dl.acm.org/citation.cfm?id=1996413.1996416. → pages

Appendix A

Contextual Interview Guideline

The contextual interview guide consisted of the questions below. Note that although the interviews help to collect data that enable us to confirm previous studies about both motivations to use tags and the types of use, the primary goal of this investigation is to understand what aspects influence the users' perception of value when choosing tags during information seeking tasks:

1. Why do you use tags? Why do you use each of these specific systems you mentioned?

2. What's the perceived value of tags produced by other users to you?

3. Can you describe search interfaces/systems of your choice that you use when looking for a set of items related to the same topic? For example, to explore a given topic of interest. (*Probes*: to find articles related to a topic of interest)

4. What are the situations where you feel the search interfaces mentioned above are more adequate to perform your search tasks, as opposed to other alternatives? (Probes: traditional keyword-based search vs. AND-search navigation)?

5. Please, describe/show us (in as much details as possible) the process you follow when using exploratory search. You can recount your last experience, for example.

6. Consider a scenario where you are looking for content on a given topic of interest. How do you choose among tags when navigating (i.e., performing information seeking tasks)?

7. Can you show us an example of an exploratory search where you had to choose among tags to proceed?

8. Why did you choose these tags while looking up these content items (from question 7)?

9. How does the partial search results influence the tags you choose to proceed with the navigation?

10. Let's talk about a different use of tags: annotation instead its use in search/-navigation. How did you choose the tags when annotating content?

11. Do you speak/write/read more than one language? If so, how do these multiple languages influence your choice of tags?

12. How the intended use of content you found during your search/navigation influence your choice of tags to annotate it?

In the second part, users are requested to 'solve' the following navigation tasks:

- **Task 0** (tutorial). Find articles related to cooking. (The goal is to get the user acquainted to the Getboo interface and enable her to perform task 1 and 2 without much intervention).

- **Task 1**. Find articles related to your work that are interesting (and new) to you.

- **Task 2**. Find articles related to your hobbies that are interesting to you.

We note that the search tasks are deliberately vague. The reason is that such tasks are the ones that motivate users to go into exploratory search mode [86] rather than trying to locate a single specific answer to an information need (e.g., what is a factotum? What is the blindekuh restaurant's location in Zrich?)

Appendix B

Codebook

Code	Type	Description	Example from data		
A - Validate previous work on motivation for tag usage					
Motivation to use systems	Inductive	Users have different motivations to use a specific system that offers tagging features.	The process of making sense is faster. Instead of reading a page... I just look for one or two sentences gives me an idea of what does it means and then I go for something else. when I do it traditionally.... Sorry... when I do it traditionally, the way it works is I have to go over everything I have to read everything more carefully. I think that in this sense it's way better.	P1	P1
Tag production	Inductive	Describes the aspects related to users' tag production habits	With Flickr I think I used more of tagging (I used it awhile ago), but [I used] tagging more as a content producer, in Twitter more as consuming conten	P1	P1
Categorization (tag production)	Inductive	Tags can be used to annotate items with the purpose of grouping these items into different groups.	Flickr I use to put my photos in different groups	P1	P1
Content promotion (tag production)	Inductive	Tags can be used to annotate items with the purpose of increasing the chance of its discovery by others.	I try to tag them in a better way to discover them so people can come and comment in my photos.	P1	P3
Join a trend (tag production)	Inductive	Refers to situations where tags are used to create trends (like promotions or games) or join a trend.	The only thing that I might use a hashtag, if something is trending in Twitter, and I want to join the trend, I add a hashtag to it.	P1	P1
Target audience (tag production)	Inductive	Refers to the fact users may consider their audience when choosing tags to annotate content	Maybe, because the target audience, or the place where I used the content from, or I generate the content for, are not persian. that's one issue.	P4	P2
Language choice (tag production)	Inductive		Maybe, because the target audience, or the place where I used the content from, or I generate the content for, are not persian. that's one issue.	P3	P3
Descriptive tags (types of tags)	Deductive	Tags may vary widely in its type (or function), this code represent evidence of tags that are used to describe the items.	I think it was location that I paid attention to - kind of - describe the location. And... I think that feature of the photo as far I can remember. If I was interested in a low light photography I put a "low light" tag, if it was HDR I put an "HDR" tag, if it was a wide angle photography...	P1	P1
Task-organization tags (Types of tags)	Deductive		outra parte das tags era pra dizer o que eu queria com aquilo, por exemplo, tinha muita coisa do tipo: eu achava que tinha um livro bacana e eu achava que deveria comprar ele ou ler ele depois	P3	P3
Source of tags	Inductive	Tags can be created by the user or extracted from social norm, for example.	In some conferences, they say: here is a hashtag of a conference, I usually tend to, I don't tweet, but I search for the hashtag to see things about the paper that just have been presented, or anything specifically that have been going inside the conference.	P1	P6

(SUB) Aggregated items evaluation	Inductive	One of the motivations to use social tagging is that items are directly (via tags or other system provided tool) or indirectly (amount of user who added the item) evaluated by the users	Eu acho que quando o cara se dava o trabalho de escrever um resumo do negócio ou era muito ruim ou era muito bom, entendeu?		
B - Build a theory on how users consume tags via exploration					
Types of searches	Inductive	We can build a taxonomy for types of searches in the context of our study. Users have mentioned serious and less serious type of searches, which seems to affect their expectations about the relevance of results, and level of refinement.	I wanted to look for one page websites, I am not sure I used two tags, or not. So, again, it was not a very serious search. So, I search for a website, I browsed through and tried to look for different things. But, I think I could use, I think, "website" and "one page", I think there would be a "onepage" tag.	P1	P4
Tag consumption	Inductive	Describes aspects related to tag use for search or other activities, such as aspects that influence the choice of tags that should be used.		P5	P5
Semantic similarity (tag consumption)	Inductive	Refers to questions related to choice of tags based on their (a)similarity	So, if we show, type and typography both of them point to the same thing, web and website, icon and icons, it's a bit of useless to have these two similar, very similar tags together, this is something that impacts the value, icons have zero value here because you have icon here.	P3	P3
Relevance of Results (tag consumption)	Deductive	Refers to the user's action of considering the results of a query to take a decision about the next step of search.	as I couldn't find lots of stuff in 'hci' 'research' 'usability' then I tried to remove the tags to find more interesting stuff with a broader... 'hci'	P3	P3
Search space reduction (Tag consumption)	Deductive	Refers to the fact that tags can be used to "cut" desired parts of the search space	as I couldn't find lots of stuff in 'hci' 'research' 'usability' then I tried to remove the tags to find more interesting stuff with a broader... 'hci'	P3	P3
Source of tags (Tag consumption)	Inductive	The producer of the information influences the perception of value of information goods.	Sometimes it happens that some of my friends... they retweet others tweet... the probability that I follow those tweets is less than their own tweets... because I'm much more interested in to see what they are thinking what they are doing than what other are thinking... but this is different from search.	P3	P3
Information foraging behaviour	Inductive		. If I feel that these first results are not the ones I'm looking for... I refine the search...	P3	P4
Search mechanism	Deductive	Users have different understanding/perceptions about how the system works.	I think of Twitter of this full text search feature that it has, that doesn't matter if your search it's specific a hashtag or not.	P1	P1
Diversity (tag consumption)	Deductive	Users may value tags conditionaly to the other tags already selected. This code highlights the aspect where users value tags more if they have diversity among themselves.	Yeah... diversity, exactly! it did not add much information. probably,, hci... then I search for something like more orthogonal tags.	P1	P1

Appendix C

Thick Descriptions

Aspects	Context	Motivation	System	What and How is it discussed?	Relations	Codes	Examples	About emotion and some text	Influence on Value
Diversity	work exploration	complete a task	Getboo	Suggesting too similar tags decreases the value of them	- Related tags	- Semantic similarity	P4	- Negative. "Eu acho confuso (...) Acho ruim. (...) por que não bota aqui “portfolio”, “design”, “musics”?"	Increase
Diversity	fun exploration	complete a task	Getboo	Using synonyms tags in a search DOESN'T make sense		- Semantic similarity	P6	- Neutral. "tem dois “fun(ny)” aqui, num faz sentido né. Vou tirar esse aqui"	Increase
Diversity	fun exploration	find resource	Dribble	Suggesting too similar tags decreases the value of them	- Related tags	- Diversity	P1	- Negative. "it’s a bit of useless to have these ... very similar tags together. This is something that impacts the value. 'Icons' have zero value here because you have 'icon' here... (I prefer to have) meaninful diversity within the tags!"	Increase
Known vocabulary	work exploration	complete a task	Getboo	When differenciating similar tags the one that is known/used have more value		- Semantic similarity	P3	- Neutral. "'computer_science' era uma tag que eu usava mais basicamente.”	Increase
Known vocabulary	work exploration	complete a task	Getboo	Tag composition have less value than the tag with the words	- Category definition	- Tag consumption	P4	- Negative. "'performance_art' é uma categoria específica que não, 'performance' sozinha"	Increase
Known vocabulary	fun exploration	complete a task	Getboo	When differenciating similar tags the one that is known/used have more value		- Semantic similarity	P8	- Neutral. "Sei lá! Quando eu procuro relacionado a games eu uso no plural."	Increase

Aspects	Context	Motivation	System	What and How is it discussed?	Relations	Codes	Examples	About emotion and some text	Influence on Value
Known vocabulary	stream/info exploration	find work related pointers	Twitter	Content and related tags associated with known vocabulary have more value.	- Related tags - More like this!	- Relevance of results	P8	- Positive. "eu escolheria as hashtags associadas que sejam semelhantes ou iguais a termos que eu escuto ... esse aqui provavelmente eu não clicaria porque eu não sei o que é esse “DAADC13”"	Increase
Making sense	stream/info exploration	been up-to-date	Twitter	Hashtags creates an easier space to be processed/evaluated. (More about Twitter?)		- Types of search	P2	- Neutral. "The process of making sense is faster."	
More like this!	stream/info exploration	job search	Twitter	Indexation by hashtags can generate a "more like this" search. It is faster for this kind of search	- Categorization	- Types of search	P2	- Positive. "In this kind of stuff, find it usually easier and faster."	Increase?
More like this!	photo exploration	find image to use	Flickr	Even as a "secondary" way of doing it, clicking on tags may help you to explore a "theme"	- Categorization	- Tag consumption	P5	- Negative. "I didn’t mean to use it. I just see a tag and I click on it... and I may find more types of trees."	Increase?
More like this!	photo exploration	find image to use	Flickr	When the results are not good enough, you may use similar tags to explore a "near space"	- Related tags - Near space?	- Semantic similarity	P7	- Neutral. "a que mais se aproxima seria “dark room”, certo? Essa, esse seria meu próximo alvo."	Increase?
More like this!	work exploration	complete a task	Getboo	When the results are not good enough, you may use similar tags to explore a "neighbour space"	- Related tags - Near space?	- Semantic similarity	P2	- Neutral. "the next one is gonna be user experience... there might be some results that they used these term instead of ux"	Increase?

Aspects	Context	Motivation	System	What and How is it discussed?	Relations	Codes	Examples	About emotion and some text	Influence on Value
More like this!	stream/info exploration	job search	Twitter	One tag has more value than the other if it describes better what I want (will retrieve more relevant results)!!!	- Related tags - Describability	- Relevance of results	P2	- Neutral. "If the results were not that much related to what I'm looking for I start looking for new hashtags ... (they might be) really more close to what I'm looking for."	Increase
More like this!	photo exploration	find image to use	Flickr	After start with a more specific tag (s) and finding that the results are from "another category" she decided to make the space broader.	- Search space - Space definition	- Relevance of results	P7	- Negative. "Isso, eu queria uma sala de cinema real. Tá, minha próxima alternativa seria apagar o 'room' e tentar só 'cinema'. Obviamente pode aparecer bem mais coisa, por exemplo, o exterior de cinemas."	NULL
Narrower space	fun exploration	complete a task	Getboo	**A more specific tag (that result in a more focused space) have more value**	- Space size - Space definition	- Relevance of results	P1	- Neutral. "I might look at 'typography'... it might give more focused result... something in this area"	Increase
Narrower space	work exploration	complete a task	Getboo	Results are used to identify if the search is going in the right direction	- Space size - Space definition	- Relevance of results	P3	- Neutral. "parece que tem muito a ver com 'storage' e tal, mas nada a ver com pesquisa... tem que refinar mais."	Increase
Narrower space	work exploration	complete a task	Getboo	Suggested tags would have more value if helped to cut more the space, make it more focused.	- Search space - Related tags	- Search space reduction	P3	- Neutral. "É a impressão que eu tenho é que.. é.. essas tags que aparecem aqui no.. na caixinha da direita são muito gerais, sabe? ... Eu ía mais na coisa mais restrita"	Increase

Aspects	Context	Motivation	System	What and How is it discussed?	Relations	Codes	Examples	About emotion and some text	Influence on Value
Narrower space	fun exploration	find link to content	Twitter	A tag is more valuable when it cuts the space more precisely.	- Space definition	- Proportion of relevant items - Search space reduction	P6	- Positive. "Sozinha não. Não, porque eu não estava atrás de conteudo do ufc, eu estava atrás do link do UFC 160, entendeu? Certamente essa palavra “link” estaria no meio."	Increase
Narrower space	fun exploration	find link to content	Twitter	A tag is more valuable when it cuts the space more precisely.	- Space definition	- Proportion of relevant items - Search space reduction	P6	- Positive. "Essa aqui #UF160 (é mais valiosa). Porque está especificando bem"	Increase
Narrower space	fun exploration	find link to content	Twitter	A tag is more valuable when it cuts the space more precisely.	- Space definition	- Proportion of relevant items - Search space reduction	P6	- Neutral. "Eu combino 2 hashtags ou 3, quando faz sentido, para tentar refinar mais."	Increase
Narrower space	work exploration	complete a task	Getboo	A tag have more value when it defines a narrower search space	- Space definition	- Search space reduction	P7	- Neutral. "Como 'opensource' já é um subconjunto de desenvolvimento/pro então eu vou começar clicando em 'opensource'"	Increase
Narrower space	work exploration	complete a task	Getboo	A tag have more value when it defines a narrower search space	- Space definition	- Search space reduction	P7	- Positive. "Então eu vou adicionar mais uma tag pra restringir mais ainda e ver se eu acho o que eu quero."	Increase
Narrower space	work exploration	complete a task	Getboo	**Two things improve the value of the tag: space definition, and space reduction.**	- Space definition	- Search space reduction	P7	- Neutral. "como eu acho que estas três tags aqui já definem bem, eu vou adicionar uma coisa bem específica pra ver se tem no sistema alguma coisa relacionada."	Increase

Aspects	Context	Motivation	System	What and How is it discussed?	Relations	Codes	Examples	About emotion and some text	Influence on Value
Narrower space	work exploration	complete a task	Getboo	**A tag have more value when it helps to define a search space**	- Space definition	- Search space reduction	P2	- Neutral. "I used 'development'... you can find different things. 'ux+development'... that was kind of development... used for user experience. if I had seen this before I would start with ux"	Increase
Narrower space	fun exploration	complete a task	Getboo	**- Two things improve the value of the tag: space definition, and space reduction.** **- A tag have more value when it defines a narrower search space**	- Space definition	- Search space reduction	P11	- Neutral. "Eu acho que ela é uma tag muito clara assim, mas muito abrangente né. Então ela com certeza não vai ser suficiente"	Increase
Narrower space	stream/info exploration	find work related pointers	Twitter	**A tag is more valuable when it cuts the space more precisely.**	- Known vocabulary	- Relevance of results	P8	- Neutral. "A segunda forma seria mais valiosa porque eu já relaciono isso aqui como uma palavra-chave. Se eu colocar separado vai me trazer resultado aqui relacionado a “máquinas” e a “social”"	Increase
Narrower space	work exploration	complete a task	Getboo	A tag have more value when it defines a narrower search space	- Search results	- Search space reduction	P8	- Neutral. (after seen that there were some pages of results): Deixa eu filtrar mais. Deixa eu ver aqui se aparece alguma coisa se botar "maps"	NULL

Aspects	Context	Motivation	System	What and How is it discussed?	Relations	Codes	Examples	About emotion and some text	Influence on Value
Proportion of relevant items	work exploration	complete a task	Getboo	**If the proportion of relevance in the first results is low the filter is not good enough**	- Search results	- Relevance of results	P7	- Neutral. “eu vou olhar as primeiras cinco, dez, eh, entradas aqui e ver mais ou menos como é que tá o meu, como é que estão os meus resultados. (...) os resultados ainda tão com muito ruído. Então eu vou adicionar mais uma tag.	Increase
Proportion of relevant items	work exploration	complete a task	Getboo	**If the proportion of relevance in the first results is low the filter is not good enough**	- Search results	- Relevance of results	P11	- Neutral. "se num tá na primeira página logo, nas primeiras, eu, eu acho que não vai ser do meu interesse"	Increase
Proportion of relevant items	stream/info exploration	general search	Twitter	**If the proportion of relevance in the first results is low the filter is not good enough**	- Search results	- Relevance of results	P11	- Negative. "o primeiro resultado é em espanhol, o segundo também... o quarto já é em espanhol também. ... muito conteúdo em espanhol que não me interessava"	Increase
Proportion of relevant items	work exploration	complete a task	Getboo	**If the proportion of relevance in the first results is low the filter is not good enough**	- Search results	- Relevance of results	P8	- Neutral. "Ainda não é suficiente isso aqui. (...) Eu olhei de início assim, os primeiros num me interessaram não."	Increase
Proportion of relevant items	work exploration	complete a task	Getboo	**If the proportion of relevance in the first results is low the filter is not good enough**	- Space definition	- Relevance of results	P8	- Neutral. "sou analista de requisitos, então, eu cliquei em “software” ... ele me retornou coisas relacionadas ao desenvolvimento em si ... nada ao que interessa."	Increase

Aspects	Context	Motivation	System	What and How is it discussed?	Relations	Codes	Examples	About emotion and some text	Influence on Value
Proportion of relevant items	photo exploration	find image to use	Flickr	**Tags that have a higher probability to return relevant items have more value**	- Search results	- Search space reduction	P7	- Neutral. "eu manteria “cinema” porque eu acho que é a palavra principal aqui... Eu tento manter aquele que eu tenho mais certeza que vão me levar ao objeto que eu quero encontrar”	Increase
Proportion of relevant items	work exploration	complete a task	Getboo	**One tag has more value than the other if it will retrieve more relevant results!!!**	- Space definition	- Relevance of results	P7	- Neutral. "buscando inicialmente por 'software' ... (vs 'programming) eu acharia mais coisas que não são relacionadas ... ao meu trabalho."	Increase
Proportion of relevant items	work exploration	complete a task	Getboo	A set of tags used to search is considered ok when the result is revelevant (Even if the created space is "too short/not meaninful?")	- Space definition - Space size	- Relevance of results	P3	- Neutral. "Aí aparentemente quando eu refino as coisas não significa quase nada. Mas o que fica é uma coisa bacana"	Increase
Proportion of relevant items	stream/info exploration	general search	Twitter	**When the tag retrieves too much (irrelevant) content this tag loses values.**	- Space size	- Search space reduction	P11	- Negative. " o que me desapontou um pouco na época foi que veio muito conteúdo. Ou seja, é uma tag tão clara e tão abrangente que veio muito conteúdo indesejado."	Increase
Related tags	fun exploration	complete a task	Getboo	If after reading related tags the user uses his own vocabulary this means that suggestions had low value!?	- Search results	- Relevance of results	P6	- Neutral. "(Murmurando analisa os resultados da tag “videos”! Faz a leitura das tags relacionadas... dizendo algo como: “É eu não usaria nenhuma daqui!”)"	Decrease

Aspects	Context	Motivation	System	What and How is it discussed?	Relations	Codes	Examples	About emotion and some text	Influence on Value
Related tags	fun exploration	complete a task	Getboo	Tags can have their value decremented if related with irrelevant content???	- Search results - Content Relevance	- Relevance of results	P6	- Negative. "(Olhando para as tags relacionadas com o link que não gostou!) Mas tá classificado aqui ó: “free, funny, humor”. Então vamos tentar “fun”)"	Decrease
Related tags	work exploration	complete a task	Getboo	- Content relevance can be judged based on related tags: if there are irrelevant tags the content loses value.	- Search results - Resulting category - Content Relevance	- Relevance of results	P8	- Negative. "Talvez esse seja mais relevante do que este aqui! Porque esse aqui trouxe mais assim: “python”, eh “php”, que já são tags que não me interessam."	NULL
Related tags	work exploration	complete a task	Getboo	Tags can have their value incremented if related with relevant content	- Search results - Content Relevance	- Relevance of results	P8	- Neutral. "se eu achei um link interessante, ... eu tento ver quais são as tags que ele tá usando pra fazer possíveis pesquisas relacionadas"	Increase
Space definition	work exploration	complete a task	Getboo	A tag that better describes a desired theme will probably retrieve more relevant results.	- Search results - Search space - Describability	- Relevance of results	P11	- Neutral. "Mas a que eu vou clicar é a “web2.0” porque eu acho que ela é mais, é, representativa da rede social na web"	Increase
Space definition	fun exploration	find link to content	Twitter	A tag that better describes a desired theme will probably retrieve more relevant results.	- Search space - Known vocabulary - Describability	- Search space reduction	P6	- Neutral. "se eu quisesse todo o conteúdo do UFC160, eu acho legal que o cara, por exemplo, toda vez que tweetasse alguma coisa ele botasse UFC160."	Increase

Aspects	Context	Motivation	System	What and How is it discussed?	Relations	Codes	Examples	About emotion and some text	Influence on Value
Space definition	work exploration	complete a task	Getboo	One tag has more value than the other if it describes better what I want (will retrieve more relevant results)!!!	- Search space - Describability	- Search space reduction	P6	- Neutral. “Então como era “tutorial”, para aprender, a palavra “tutorial” estava dentro da tag “Ruby” aqui, então para mim faria mais sentido.”	Increase
Space definition	bookmark exploration	complete a task	Getboo	The search was simpler because the space was "well defined" (in a small amount of tags)		- Types of search	P2	- Neutral. "I mean it was simpler... there are not so many options of what I'm looking for."	Increase
Space definition	fun exploration	complete a task	Getboo	If the space is already defined using more tags will have a lower value.	- Search space	- Relevance of results	P3	- Neutral. "eu não usaria isso aqui (TAGS RELACIONADAS) pra buscar aqui. Eu já sabia que, por exemplo 'scifi' definiria isso aqui pra mim o suficiente"	Decrease
Space definition	fun exploration	complete a task	Getboo	If the space is already defined using more tags will have a lower value.	- Space size?	- Search space reduction	P3	- Positive. "Bacana. Parece que colocando o tema direto é mais bacana do que (combinar com outras tags) talvez seja porque é uma coisa.. sei lá, muito específica, que pouca gente gosta. Não sei."	Decrease
Space definition	fun exploration	complete a task	Getboo	**- A tag have more value when it helps to define a search space**		- Search space reduction	P8	- Neutral. (after asked about 'entertainment' tag): "porque eu tava vendo muita coisa relacionada a notícia aqui."	Increase

Aspects	Context	Motivation	System	What and How is it discussed?	Relations	Codes	Examples	About emotion and some text	Influence on Value
Space size	work exploration	complete a task	Getboo	Sometimes to decide between two similar spaces is better to take a look at both and evaluate the results		- Semantic similarity	P7	- Neutral. "acho que pode ajudar bastante quando você precisa filtrar por tags relacionadas, é a quantidade de resultados"	Mediate
Space size	work exploration	find work references	general	The number of results is an aspect to decide when to continue to search.		- Search space reduction	P4	- Neutral. "Aí aparece muitas palavras aí eu coloco um nome que tenta identificar"	Decrease
Space size	fun exploration	complete a task	Getboo	The number of results is an aspect to decide when to continue to search.		- Search space reduction	P4	- Negative. "(PORTLAND) apareceu demais aqui! Eu acho que não vou confiar ... (PORTLAND+MUSIC	Decrease
Space size	fun exploration	complete a task	Getboo	There is an ideal size of a search space: not too many BUT with options!		- Search space reduction	P11	- Negative. "muitos resultados confundem e você não consegue achar aquilo que você quer... um número de resultados que nem preenche a primeira página do sistema é um pouco frustante."	Mediate
Space size	work exploration	complete a task	Getboo	When the space is too small the related tags have a smaller value?	- Search space	- Search space reduction	P1	- Neutral. "It might have filtered too much... or I was really specific... to very focused... as I couldn't find lots of stuff in 'hci' 'research' 'usability' then I tried to remove the tags to find more interesting stuff with a broader... 'hci'"	Increase

Aspects	Context	Motivation	System	What and How is it discussed?	Relations	Codes	Examples	About emotion and some text	Influence on Value
Search space	photo exploration	general	Flickr	When you use tags to search you create different spaces to deal with		- Tag consumption	P5	- Neutral. "expect to see things more in a category. … I think I just expected different sets."	
Search space	stream/info exploration	explore for information	Twitter	Hash tags have more value in Twitter because they define a narrower search space		- Search space reduction	P7	- Neutral. "o hash marca o tópico. Normalmente isso ajuda. Se você não achar com o tópico pode voltar e tentar sem eles."	
Search strategy	work exploration	complete a task	Getboo	After defining a good space to search it is valuable to test more specific tags	- Search space	- Relevance of results	P3	- Neutral. "Aí eu vou colocar (filesystem) ... mais ligado a pesquisa... termo menor popular."	
Search strategy	work exploration	complete a task	Getboo	Cutting around is good enough for a first step	- Search space	- Search space reduction	P6	- Neutral. "Obviamente é uma linguagem, não é o framework, mas eu iria ver do que é que se trata."	
Search strategy	fun exploration	complete a task	Getboo	One tag has more value than the other if it will retrieve more relevant results!!!	- Search space	- Relevance of results	P8	- Neutral. "também é interessante você tirar uma tag ... ver o que tem a outra. ... 'não, isso aqui não interessa não, vou tirar que tá vindo muita besteira'	
Types of tags	stream/info exploration	general search	Twitter	Groups/Sources of tags that are previouly judged as having low relevance will contribute negatively for non knowing tags.	- Search results	- Relevance of results	P11	- Negative. "as tags das trends geralmente... têm um caráter que não me atraem nos tweets... essa coisa aqui que eu nem sei o que é, oh!"	NULL

Aspects	Context	Motivation	System	What and How is it discussed?	Relations	Codes	Examples	About emotion and some text	Influence on Value
Types of tags	work exploration	complete a task	Getboo	- To decide between two tags the expected result is considered. - A "known tag" will have more value!?	- Known vocabulary - Describability	- Relevance of results	P4	- Neutral. "eu acho que.. esse tipo de tag não seria usado pela pessoa que postou"	NULL