BUILDING AND EXPLOITING CONTEXT ON THE WEB

Von der Fakultät für Elektrotechnik und Informatik
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des Grades

Doktor der Naturwissenschaften

Dr. rer. nat.

genehmigte Dissertation von

M. Sc. Ricardo Robalinho Kawase

geboren am 7. Januar 1981, in São Paulo, Brasilien

Hannover, Deutschland, 2014

Referent: Prof. Dr. techn. Wolfgang Nejdl
Ko-Referent: Prof. Dr. Daniel Schwabe
Tag der Promotion: 16. Oktober 2014

ZUSAMMENFASSUNG

Das World Wide Web ist ein wichtiger Teil unseres Lebens geworden. Wir nutzen das Internet für eine Vielzahl verschiedener Tätigkeiten wie arbeiten, lernen, einkaufen oder spielen.

Der Austausch von Informationen im Web folgte lange Zeit nur eine Richtung: Autoren veröffentlichten Seiten für Leser, die nur begrenzt mit den Texten interagieren konnen. Das Web 2.0 bietet engagierten Nutzern deutlich mehr Möglichkeiten zur Interaktion. Viele Nutzer liefern explizite Informationen wie Kommentare, Annotationen, Tags oder Bewertungen. Zusammen mit Aufzeichnungen des Nutzerverhaltens bilden diese Aktivitäten den Kontext für Web-Ressourcen. Dieser Kontext kann genutzt werden, um das Klassifizieren, Finden, Verarbeiten und Verstehen von Informationen zu ermöglichen oder zu verbessern.

In dieser Arbeit stellen wir eine Vielzahl von Studien vor, in denen wir verschiedene Arten kontextueller Informationen ausgenutzt haben, um die Nutzung des Webs für Benutzer zu verbessern.

Zunächst analysieren wir den Einfluss von kontextabhängigen Annotationen in schriftlicher Form und auf Webseiten. Zu diesem Zweck implementierten und evaluierten wir ein Tool für kontextuelle Annotationen, welches das Teilen und Finden von Informationen unterstützt und erleichtert. Die Ergebnisse dieser Arbeit führten zu einer verbesserten Unterstützung von Annotationen in einer Online-Lernumgebung.

Desweiteren stellen wir ein System vor, bei dem Kontext in Folksonomie-Systeme integriert wird. Basierend auf diesem erweiterten Folksonomie-Model zeigen wir Strategien, die kontextabhängige Informationen nutzen um Profile von Nutzern und Ressourcen zu verbessern. Die durchgeführten Experimente zeigen, dass kontextabhängige Ranking-Algorithmen das Information Retrieval in Folksonomie-Systemen signifikant verbessern.

Abschlieend werden kontextabhängige Vorhersagemethoden für das Surfverhalten von Nutzern vorgestellt und evaluiert. Durch eine Reihe von Analysen und Experimenten zeigen wir, dass wiederkehrende Besuche eine zentrale Rolle beim Verhalten von Nutzern im Internet spielen, und dass dieses Verhalten gut vorhersagbar ist. Unsere kontextsensitiven Methoden erlauben signifikante Verbesserungen bei der Vorhersage der nächsten besuchten Seite. Zusätzlich zeigen wir eine Klassifikation für Online-Aktivitäten von Nutzern. Diese liefert wichtige Informationen für die Weiterentwicklung von kontextualisierter Unterstützung bei der Navigation im Netz.

Schlagworte: Contextualization, Annotations, Information Refinding, Contextualized Profiles, Revisitation.

ABSTRACT

The World Wide Web has become an important part of our lives. We use the Web for a whole range of diverse activities, including working, learning, dating, shopping, and gaming.

Whereas the exchange of ideas on the Web used to be mostly one-way - that is, authors publish and the viewers have a limited means of interacting with information - with the Web 2.0 new means of interactions have given more power and more influence to the more engaged user. Explicit user input - such as comments, annotations, tags and ratings - and implicitly recorded interaction data provide contextual information for Web resources. This context is potentially useful for enabling and improving features that are essential for supporting information classification, retrieval, processing and understanding.

In this thesis, we present several studies in which we analyze and exploit different kinds of contextualized information in order to improve users' Web experience.

First, we study the effects of contextualized annotations on paper and on the Web. We implemented and evaluated an online contextualized annotation tool that support and improve information sharing and re-finding. The outcomes guided us in improved annotation support in an online learning environment that support users in their learning activities.

Second, we propose a model that incorporates context in folksonomies systems. Based on this extended folksonomy model, we propose strategies for exploiting the contextualized information in order to improve profiling of users and resources. Our experiments show that context-based ranking algorithms significantly improve information retrieval in folksonomy systems.

Finally, we propose and evaluate several contextual prediction methods that exploit the user browsing context. We demonstrate through a series of analyses and experiments that revisitation plays a major role in Web users' activities, and that this recurrent behavior is highly predictable. Our context-sensitive methods significantly improve the next-page prediction task. Additionally, we provide a sense-making classification of users' online activities that provides important pointers for the further development of contextualized browsing support.

Keywords: Contextualization, Annotations, Information Refinding, Contextualized Profiles, Revisitation.

FOREWORD

The studies presented in this thesis have been published at various conferences or journals, as follows.

In Chapter 2, we describe contributions included in:

- Ricardo Kawase, Eelco Herder and Wolfgang Nejdl. A Comparison of Paper-Based and Online Annotations in the Workplace. In ***EC-TEL**: Proceedings of the 4th European Conference on Technology Enhanced Learning: Learning in the Synergy of Multiple Disciplines*, pages 240-253, 2009 (Full Paper). [KHN09]

- Ricardo Kawase and Wolfgang Nejdl. A Straightforward Approach for Online Annotations: SpreadCrumbs - Enhancing and Simplifying Online Collaboration. In ***WEBIST**: Proceedings of the Fifth International Conference on Web Information Systems and Technologies*, pages 407-410, 2009 (Poster). [KN09]

- Ricardo Kawase, Eelco Herder and Wolfgang Nejdl. Annotations and Hypertrails with Spreadcrumbs - An Easy Way to Annotate, Refind and Share. In ***WEBIST**: Proceedings of the 6th International Conference on Web Information Systems and Technologies, Volume 2*, pages 5-12, 2010 (Full Paper). [KHN10]

- Ricardo Kawase, George Papadakis, Eelco Herder and Wolfgang Nejdl. The impact of bookmarks and annotations on refinding information. In ***HT**: Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, pages 29-34, 2010 (Short Paper). [KPHN10]

- Bernardo Pereira Nunes, Ricardo Kawase, Stefan Dietze, Gilda Helena Bernardino de Campos and Wolfgang Nejdl. Annotation Tool for Enhancing E-Learning Courses. In ***ICWL**: Proceedings of the 11th International Conference on Advances in Web-Based Learning*, pages 51-60, 2012 (Full Paper). [NKD+12]

Chapter 3 is built upon the work published in:

- Fabian Abel and Ricardo Kawase and Daniel Krause and Patrick Siehndel. Multi-faceted Tagging in TagMe!. In ***ISWC**: 8th International*

Semantic Web Conference (ISWC '09), Poster Track, 2009 (Poster). [AKKS09]

- Fabian Abel, Ricardo Kawase, Daniel Krause, Patrick Siehndel and Nicole Ullmann. The Art of Multi-faceted Tagging - Interweaving Spatial Annotations, Categories, Meaningful URIs and Tags. In ***WEBIST**: Proceedings of the 6th International Conference on Web Information Systems and Technologies*, pages 21-28, 2010 (Full Paper). [AKK+10b]

- Fabian Abel, Nicola Henze, Ricardo Kawase and Daniel Krause. The Impact of Multifaceted Tagging on Learning Tag Relations and Search. In ***ESWC**: Proceedings of the 7th Extended Semantic Web Conference*, pages 90-105, 2010 (Full Paper). [AHKK10]

- Fabian Abel, Ricardo Kawase and Daniel Krause. Leveraging multi-faceted tagging to improve search in folksonomy systems. In ***HT**: Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, pages 299-300, 2010 (Poster). [AKK10a]

- Ricardo Kawase, George Papadakis and Fabian Abel. Generating Resource Profiles by Exploiting the Context of Social Annotations. In ***ISWC**: Proceedings of the 10th International Semantic Web Conference*, pages 289-304, 2011 (Full Paper). [KPA11]

Finally, in Chapter 4 we include our research presented in:

- Ricardo Kawase, George Papadakis and Eelco Herder. How Predictable Are You? A Comparison of Prediction Algorithms for Web Page Revisitation. In ***ABIS**: Proceedings of 18th International Workshop on Personalization and Recommendation on the Web and Beyond*, 2010 (Full Paper). [KPH10]

- Ricardo Kawase, George Papadakis, Eelco Herder, Wolfgang Nejdl. Beyond the usual suspects: context-aware revisitation support. In ***HT**: Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*, pages 27-36, 2011 (Full Paper - **Douglas Engelbart Best Paper Award**). [KPHN11]

- Ricardo Kawase, George Papadakis and Eelco Herder. Supporting revisitation with contextual suggestions. In ***JCDL**: Proceedings of the 2011 Joint International Conference on Digital Libraries*, pages 227-230, 2011 (Short Paper). [KPH11]

- Ricardo Kawase and Eelco Herder. Classification of user interest patterns using a virtual folksonomy. In ***JCDL**: Proceedings of the 2011*

Joint International Conference on Digital Libraries, pages 105-108, 2011 (Short Paper). [KPH11]

During the stages for my Ph.D. studies, I have also published a number of papers investigating different areas of Web Science. Not all researched areas are touched in this thesis due to space limitation, but the complete list of publications follows:

- Ricardo Kawase, Enrico Minack, Wolfgang Nejdl, Samur Araújo and Daniel Schwabe. Incremental End-user Query Construction for the Semantic Desktop. In ***WEBIST***: *Proceedings of the Fifth International Conference on Web Information Systems and Technologies*, pages 270-275, 2009. [KMN+09]

- Fabian Abel, Nicola Henze, Ricardo Kawase, Daniel Krause and Patrick Siehndel. TagMe!: Enhancing Social Tagging with Spatial Context. In ***WEBIST (Selected Papers)***: *Web Information Systems and Technologies, Lecture Notes in Business Information Processing*, pages 114-128, 2010. [AHK+10]

- Ricardo Kawase, Eelco Herder, George Papadakis and Wolfgang Nejdl. In-Context Annotations for Refinding and Sharing. In ***WEBIST (Selected Papers)***: *Web Information Systems and Technologies, Lecture Notes in Business Information Processing*, pages 85-100, 2010. [KHPN10]

- Katja Niemann, Uta Schwertel, Marco Kalz, Alexander Mikroyannidis, Marco Fisichella, Martin Friedrich, Michele Dicerto, Kyung-Hun Ha, Philipp Holtkamp, and Ricardo Kawase. Skill-based scouting of open management content. In ***EC-TEL***: *Proceedings of the Fifth European Conference on Technology Enhanced Learning*, pages 632-637, 2010. [NSK+10]

- Fabian Abel, Matteo Baldoni, Cristina Baroglio, Nicola Henze, Ricardo Kawase, Daniel Krause and Viviana Patti. Leveraging search and content exploration by exploiting context in folksonomy systems. In ***The New Review of Hypermedia and Multimedia***, pages 33-70, 2010. [ABB+10]

- George Papadakis, Ricardo Kawase, Eelco Herder and Claudia Niederée. A Layered Approach to Revisitation Prediction. In ***ICWE***: *Proceedings of the 11th International Conference on Web Engineering*, pages 258-273, 2011. [PKHN11]

- Ernesto Diaz-Aviles, Marco Fisichella, Ricardo Kawase, Wolfgang Nejdl and Avaré Stewart. Unsupervised Auto-tagging for Learning Object Enrichment. In ***EC-TEL***: *Proceedings of the 6th European Conference of Technology Enhanced Learning*, pages 83-96, 2011. [DAFK+11]

- Ernesto Diaz-Aviles and Ricardo Kawase. Exploiting Twitter as a Social Channel for Human Computation. In ***CrowdSearch***: *Proceedings of the First International Workshop on Crowdsourcing Web Search*, pages 15-19, 2011. [DAK12]

- Marco Fisichella, Ricardo Kawase, Juri Luca De Coi, and Maristella Matera. User profile based activities in flexible processes. In ***WIMS***: *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, Article 33, 2012. [FCKM12]

- George Papadakis, Ricardo Kawase and Eelco Herder. Client- and server-side revisitation prediction with SUPRA. In ***WIMS***: *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, Article 14, 2012. [PKH12]

- Patrick Siehndel and Ricardo Kawase. TwikiMe! - User Profiles That Make Sense. In ***ISWC***: *Proceedings of the 11th International Semantic Web Conference, Posters & Demonstrations Track*, 2012. [SK12]

- Ricardo Kawase, Patrick Siehndel, Eelco Herder and Wolfgang Nejdl. Hyperlink of Men. In ***LA-WEB***: *Proceedings of the 8th Latin American Web Congress*, pages 9-16, 2012. [KSHN12]

- Bernardo Pereira Nunes, Alexander Arturo Mera Caraballo, Marco A. Casanova and Ricardo Kawase. Automatically generating multilingual, semantically enhanced, descriptions of digital audio and video objects on the Web. In ***KES***: *Proceedings of the 16th Annual KES Conference on Advances in Knowledge-Based and Intelligent Information and Engineering Systems*, pages 575-584, 2012. [NCCK12a]

- Bernardo Pereira Nunes, Alexander Arturo Mera Caraballo, Marco A. Casanova, Ricardo Kawase. Boosting Retrieval of Digital Spoken Content. In ***KES*** *(Selected Papers): Proceedings of the 16th Annual KES Conference on Advances in Knowledge-Based and Intelligent Information and Engineering Systems*, pages 153-162, 2012. [NCCK12b]

- Ricardo Kawase, Patrick Siehndel, Bernardo Pereira Nunes, Marco Fisichella and Wolfgang Nejdl. Towards Automatic Competence Assignment of Learning Objects. In ***EC-TEL***: *Proceedings of the 7th European Conference of Technology Enhanced Learning*, pages 575-584, 2012. [KSN+12]

- Ricardo Kawase, Marco Fisichella, Katja Niemann, Vassilis Pitsilis, Aristides Vidalis, Philipp Holtkamp and Bernardo Pereira Nunes. OpenScout: harvesting business and management learning objects from the web of data. In ***WWW**: 22nd International World Wide Web Conference, WWW '13, Companion Volume*, pages 445-450, 2013. [KFN+13a]

- Patrick Siehndel, Ricardo Kawase, Asmelash Teka Hadgu and Eelco Herder. Finding relevant missing references in learning courses. In ***WWW**: 22nd International World Wide Web Conference, WWW '13 Companion Volume*, pages 425-430, 2013. [SKHH13]

- Ricardo Kawase, Marco Fisichella, Bernardo Pereira Nunes, Kyung-Hun Ha and Markus Bick. Automatic classification of documents in cold-start scenarios. In ***WIMS**:3rd International Conference on Web Intelligence, Mining and Semantics, WIMS '13, June 12-14, 2013*, page 19, 2013. [KFN+13b]

- Bernardo Pereira Nunes, Stefan Dietze, Marco Antonio Casanova, Ricardo Kawase, Besnik Fetahu and Wolfgang Nejdl. Combining a Co-occurrence-Based and a Semantic Measure for Entity Linking. In ***ESWC**: The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013, May 26-30, 2013*, pages 548-562, 2013. [NDC+13]

- Ricardo Kawase, Bernardo Pereira Nunes, Eelco Herder, Wolfgang Nejdl and Marco Antonio Casanova. Who wants to get fired? In ***WebSci 2013**: Paris, France, May 2-4, 2013*, pages 191-194, 2013. [KNH+13]

- Bernardo Pereira Nunes, Ricardo Kawase, Besnik Fetahu, Stefan Dietze, Marco A. Casanova and Diana Maynard. Interlinking Documents based on Semantic Graphs In ***KES 2013**: Kitakyushu, Japan, September 9-11, 2013*, pages 231-240, 2013. [KNS13]

- Ricardo Kawase, Patrick Siehndel, Bernardo Pereira Nunes and Marco Fisichella. Automatic Competence Leveling of Learning Objects In ***ICALT 2013**: Beijing, China, July 15-18, 2013*, pages 149-153, 2013. [KSNF13]

- Bernardo Pereira Nunes, Ricardo Kawase, Patrick Siehndel, Marco A. Casanova and Stefan Dietze. As Simple as It Gets - A Sentence Simplifier for Different Learning Levels and Contexts In ***ICALT 2013**: Beijing, China, July 15-18, 2013*, pages 128-132, 2013. [NKS+13]

- Ricardo Kawase, Bernardo Pereira Nunes and Patrick Siehndel. Content-Based Movie Recommendation within Learning Contexts In ***ICALT 2013**: Beijing, China, July 15-18, 2013*, pages 171-173, 2013. [KNS13]

- Bernardo Pereira Nunes, Stefan Dietze, Marco Antonio Casanova, Ricardo Kawase, Besnik Fetahu and Wolfgang Nejdl. Combining a Co-occurrence-Based and a Semantic Measure for Entity Linking In ***ESWC 2013****: Montpellier, France, May 26-30, 2013*, pages 548-562, 2013. [NDC+13]

- Bernardo Pereira Nunes, Stella Pedrosa, Ricardo Kawase, Mohammad Alrifai, Ivana Marenzi, Stefan Dietze and Marco Antonio Casanova Answering Confucius: The Reason Why We Complicate In ***EC-TEL 2013****: Paphos, Cyprus, September, 17 - 21, 2013*, pages 496-501, 2013. [NPK+13]

- Kaweh Djafari Naini, Ricardo Kawase, Nattiya Kanhabua, Claudia Niedere. Characterizing high-impact features for content retention in social web applications. In ***WWW (Companion Volume)*** *Seoul, Korea, April, 7-11, 2014*, pages 559-560, 2014. [NKKN14]

- Ricardo Kawase, Patrick Siehndel, Eelco Herder: Haters Gonna Hate: Job-Related Offenses in Twitter. In ***WWW (Companion Volume)*** *Seoul, Korea, April, 7-11, 2014*, pages 2014: 557-558, 2014. [KSH14]

ACKNOWLEDGMENTS

First, I would like to thank Prof. Dr. Wolfgang Nejdl for giving me the opportunity to achieve my Ph.D. and to establish the L3S Research Center as a unique, creative and international environment that I am pleased and proud to be part of.

I thank Prof. Dr. Daniel Schwabe for introducing me to the academic life, as well as for the important role he played during my graduate years and my Master studies, for his time invested in reading and reviewing this thesis, and above all, for his friendship.

I thank Dr. Eelco Herder for his brilliant support and guidance during my Ph.D. years. As a mentor, he taught me how to turn ideas into research. Additionally, I thank him for being the main critic of my work. His meticulous criticism challenged me to go one step beyond in each research project, and helped me incredibly improve my creative thinking.

I thank my colleague and friend Fabian Abel who was the first to welcome me in Germany. His support in my early days in this country was essential for my adaptation. Also, he was a role model for my Ph.D work. He showed me how to do research efficiently, set up collaboration, and engage in an infinity of research activities.

I thank Patrick Siehndel and Bernardo Pereira Nunes for the fruitful brainstorm sessions, the priceless collaboration, and for sharing with me the same challenging spirit.

I thank Dr. Stefan Dietze for the collaborations and his valuable feedback of this thesis.

I thank all my talented colleagues with whom I spent the first and last year at the L3S Research Center, especially Mohammad Alrifai, Kerstin Bischoff, Gianluca Demartini, Julien Gaugaz, Tereza Iofciu, Benjamin Köhncke and Ivana Marenzi. Their companionship, on many occasions, went much beyond the work environment.

I thank my new friends, particularly Marco Fisichella, George Papadakis and Gian Luca Volpato. Besides being my closest collaborators at work, they were the foundation of my social life in Germany, and I will always cherish their friendship, beyond time and geography.

I thank my dear old friends from Brazil, Eduardo, Ian and Karina for their omnipresence in my life during my Ph.D. years, despite the 10,000km distance.

I thank Agata, for being my touchstone, my safe-haven and for taking part in my Ph.D. life.

I thank each and every one who crossed my path during my Ph.D years. No matter how fast or how short was our contact, I believe each of them had a significant influence in my life and the person I became, changing my way of thinking, thus impacting, to different extents, the outcome of my work.

Last but not least, I thank my family, especially my father and my mother for the unconditional love and support. Thank you. This is for you.

Contents

List of Figures

1 Introduction

The World Wide Web is arguably the biggest source of information nowadays. Whereas the exchange of ideas on the Web used to be mostly one-way - that is, authors publish and the viewers have a limited means of interacting with information - with the Web 2.0 new means of interactions have given more power and more influence to the more engaged user. However, there are still a number of features missing that are essential for supporting information classification, retrieval, processing and understanding.

Most of these issues have been already reported during the early inception of the Web, mainly from the hypertext community [WDBG+02, VB99]. In particular, frequently mentioned are the lack of typed or annotated links, the absence of hypertrails, limited browser history mechanisms and the lack of support for annotations.

During the last decade, the Web 2.0 became the most popular Web setting. The most successful websites strongly depend on the contents and interactions produced by their users. In this setup of user contributed environments, *tags* emerge as the simplest form of user generated content.

In fact, the tagging paradigm attracted much attention in the Web community. More and more Web systems allow their users to annotate content with freely chosen keywords (*tags*). The tagging feature helps users to organize content for future retrieval [MNBD06b]. Resource sharing systems like Del.icio.us[1], Flickr[2], or Last.fm[3] would not work without the users, who assign tags to the shared bookmarks, images, and music respectively, because tag assignments are used as information source to provide diverse features such as recommendation, search, or exploration features. For example, tag clouds, which depict the popularity of tags within the system, intuitively allow users to explore a repository of tag-annotated resources, just by clicking on tags.

Beside search algorithms that simply detect resources directly annotated with

[1] http://delicious.com/
[2] http://flickr.com/
[3] http://last.fm/

the *search tag*, there exist more advanced algorithms that exploit the full structure of the *folksonomy* [Wal07]. A folksonomy is basically a collection of all tag assignments (user-tag-resource bindings) in the system. It can be modeled as graph which makes it possible to apply graph-based search and ranking algorithms according to the paradigm of PageRank [BP98]. Such ranking algorithms like FolkRank [HJSS06b], which is based on PageRank and applicable to folksonomies, not only allow to rank resources but also *tags* and *users*. This feature expands the scope of applications to tag recommendations, user/expert search, etc.

Hence, ranking algorithms play a central role in a multitude of applications, however all ranking algorithms have to face the problem of ambiguity. For example, the tag "java" might be assigned to resources related to programming or the island of Indonesia. Another problem is caused by tags that are re-used on various occasions with different (though implicit) meaning. For instance, the tag "to-read" might be added by a same user at different times to scientific papers that are relevant for a research work or to websites that explain what to see in some location the user would like to visit on holidays. If the tag "to-read" would be used in a query, likely the ranking algorithm outcome would not satisfy the user because such algorithms lack the means to contextualize the ranking.

Correspondingly, for broad tags like "music" or "web", which are assigned to a huge amount of resources, it is difficult to compute a ranking that fits to the actual desires of the user.

One could think that ambiguity could be reduced by adopting personalization strategies, so to produce personalized rankings. The problem is that personalization techniques are currently limited by their need of time to build adequate user models: The user has, in fact, to register to the system and work long enough to allow the system itself to collect a sufficient amount of data to provide personalization.

More than a decade ago, Nielsen claimed that, rather than investing time and energy on trying to predict individual user's needs, it would be more fruitful to enhance the overall system design[4]. In contrast to his assertion, we share the vision of the adaptive hypermedia community, supporting the idea that "one size does not fit all" [Bru01]. Much has changed since Nielsen's declaration, with the majority of contemporary systems (especially web-based ones) incorporating recommendation mechanisms to suggest resources (e.g., web pages, files or products) to their users according to an underlying prediction model.

Indeed, many applications can benefit from effective methods of user modeling, like Web search, where predictive models have improved the ranking of search engine results [BP98].

For example, navigational information is actually considered more important than text keywords, since relevant web pages are typically re-ranked according to the distribution of visits over them. Hence, the more accurate the predictive models are,

[4] http://www.useit.com/alertbox/981004.html

the better search results they yield. Similarly, individual users can benefit to a large extent from methods predicting and recommending their next page request. Both in their working and in their personal environment, they usually have to handle repetitive but infrequent tasks, revisiting pages after a considerable amount of time [CM01]. Although users typically employ bookmarks to facilitate such activities, the usability of their bookmark declines rapidly with the constant increase of its size [CM01].

In this light, we divide this thesis in three main distinct chapters. In each of them, we will approach the aforementioned problems that involves annotations on the Web, profiling in folksonomy systems and surfing the Web. The pivot element that binds together this thesis is the use of contextualized information that regards each of the topics.

A generic definition of context is:

Definition 1 *[Context] 'The circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood and assessed'.*

Here, we will demonstrate that context is, in fact, a generic definition that can be broken down into fine-grained interpretations and, we will prove that each piece of context can be used to improve user experience on the Web.

1.1 Thesis Structure

The main contributions of this thesis are described in Chapters 2-4. Chapter 2 presents the research on top of contextualized annotations on the Web. Chapter 3 contains the research around the development of contextualized profiles. Chapter 4 presents the research regarding browsing context and predictive models. Each of these chapters will start with an introduction, which motivates the corresponding research questions by referring to related work, and will conclude with a summary of main findings and contributions. Chapter 5 concludes with a summary of main findings and contributions:

Chapter 2: This chapter begins with a thorough analysis of paper-based annotations, where a field-study was performed to collect enough evidence of annotations used in real reading/learning activities. Later, we expose the development of a contextualized Web annotation system, followed by several evaluations of its benefits. The chapter finally presents the implementation and outcomes of a Web annotation system in real education scenarios.

Chapter 3: In this chapter, we propose strategies for deducing contextual information form social tagging processes. We introduce a generic context folksonomy model that integrates such information. Further, we define strategies to exploit this information in order to build context-based resource profiles. In addition to that, we demonstrate the applicability of these profiles for the task of tag recommendations in two different folksonomy systems.

Chapter 4: This chapter focus on the extraction and use of user browsing context. It begins with a anaylisis of users browsing behavior followed by the proposal of different predictive models for exploit the context of the users' browsing history. The models are the result of combination of propagation methods with ranking methods. We present two user evaluations of contextual recommendations through the development and use of a dynamic contextual bookmark plug-in. The lessons learned during the first evaluation implicate the study on recommending pages versus sites and the subsequent evaluation. This chapter finaly presents an additional study to provide a sensemaking classification of users' tasks interests.

Chapter 5: This chapter concludes this thesis by summarizing our main findings and contributions. Further, we outline future work made possible by the findings of this thesis and discuss open research challenges.

1.2 Contributions of this Thesis

Our contributions are summarized as follows:

- We provide a thorough understanding of annotations in paper-based and web-based scenarios.
- We develop a solution to support in-context Web annotations.
- We develop a solution to support in-context annotations in educational scenarios.
- We propose a model that incorporates context in folsonomies.
- We propose strategies that exploit the contextualized folksonomies in order to improve profiling of users and resources.
- We propose and evaluate several contextual prediction methods that exploit the user browsing context.
- We develop a tool to collect contributions of browsing user data and publicly provide the dataset for future research.
- We evaluate all proposed ideas in this thesis with user studies.

2

Web Annotations in context

In this chapter, we will study the problem of shifting paper-based annotations to the digital environment. Ever since a great deal of reading activities occur in digital format, it is expected that annotations take place in the same environment. The study presents a thorough data collection, comparison, user-study and validation. At first, in order to understand the differences between environments, we present a comparison between paper-based and digital annotations. The idea is to comprehend the different forms and goals of annotations. In sequence, we propose and validate the benefits of a contextualized Web annotation tool that supports collaboration and information refinding. Finally, we validate the usefulness of contextualized digital annotations in a real learning scenario, introducing a contextualized annotation tool to online courses.

2.1 Introduction

We understand annotation as some extra information attached to a resource, that can assume many different forms. In-context annotations may not only help the annotator later but may be useful as well for other future readers; indeed, scribbling is extremely common during reading activities. In some user driven tests O'Hara and Sellen [OS97] demonstrated that most of the subjects used annotations to help understand the text and to aid in the future task of writing. In an impressive field study on annotations in college textbooks, Marshall [Mar97, Mar98] managed to identify patterns in annotations, statistics and further more describing and classifying the many forms of annotations such as: signaling for future attention, memory aiding, problem-working, interpretation, progress tracking in narrative and so on.

Given that the Internet is the largest source of information, it is expected that a lot of the readings occur online; consequently Web annotation would be an expected feature on the Internet. However, no annotation system so far has shown nimbleness, perspective or has survived the first years of existence. Nevertheless, it has

been widely discussed the importance of annotations for comprehension and also the benefits for reading/writing proposes [OS97]. Given the absence of any dominant mature annotation system, it appears that there is still no generally accepted, concrete method for straightforward online annotation. In order to understand such problem and the user's preferences for tagging and bookmarking systems over annotation systems - we have developed a simple, easy to use and straightforward system that supports in-context Web annotation with basic features of annotation, bookmarking and social navigation support.

This system, namely SpreadCrumbs, was developed with the intention of supporting our research and validating how users interact with such systems, the benefits of contextualized annotations for re-finding information and for learning activities. The SpreadCrumbs tool was designed after a thorough study to understand the differences between paper annotations and digital ones. The evaluation of the tool in refinding tasks gave us further insights to install an in-context annotation support system in real learning scenarios.

In this light, the research questions we address in this chapter are:

- What are the main differences between paper based and web based annotations?
- Can spatial context be exploited in digital environments?
- Do spatial contextualized annotations improve refinding information tasks?
- Can spatial contextualized annotations support learners?

In the reminder of this chapter we answer these questions and provide the following contributions:

- We provide a thorough understanding of annotations in paper-based and web-based scenarios.
- We develop a solution to support in-context Web annotations.
- We demonstrate the effectiveness of in-context digital annotations in supporting information refinding tasks.
- We develop and evaluate a solution to support in-context annotations in educational scenarios.

2.2 Related Work

The first group of related works is the existing and past commercial tools for web annotation. ThirdVoice[1] was probably the first expressive commercial Web annotation tool. It was a plug-in for Internet Explorer 4 and Netscape Web browsers which

[1] http://www.ThirdVoice.com (March, 2000)

allowed the users to publicly annotate any webpage. The 'in margin' written annotations were visible to any user of the application that accesses the site. It is not completely transparent where ThirdVoice failed but the service was discontinued on April 2001. Some other old discontinued commercial systems Hypernix, NovaWiz, utok, Zadu followed the same steps with less public attention. More up-to-date systems Fleck[2], SharedCopy[3], Diigo[4] bring a new air for the annotations scenario. They grant tagging, re-finding, collaboration, social navigation and annotation itself working in the same way as the old tools - a plug-in for the browsers. Still, none of them reached a considerable impact level in the Web as it happened in social networks, folksonomies and tagging systems.

In addition to the commercial tools, several research projects aim to enhance Web collaboration by providing annotation capabilities. We have investigated some of these works to try to understand the evolution in the Web annotation scenario.

The Anchored Conversations system [CTB+00] provides a synchronous text chat window that can be anchored to a specific point within a document. It is presented as a post-it note and can also be used for re-finding by the system search option. In this case, the collaboration occurs during a synchronous chat. Like the Anchored Conversations, we understand that the most appropriate metaphor for transient annotations is the post-it notes.

Fluid Annotations [ZBJM01] supports in-context annotations and it is an extension of the open hypermedia Arakne Environment [Bou99]. But different from other researches, the studies and evaluations are mostly presentation of the annotations, as seen in [ZRMC00, ZBJM01] in terms of visual cues, interactions and animated transactions. Their evaluations give valuable material for annotations manipulation and usability, however, their approach of 'between lines' annotations disrupts the original layout of the annotated content besides the distractive animation transactions.

In the end, all attempts, projects and commercial tools aim to enhance communication and collaboration among the users independently of the task. Putting together all those systems there is a common understanding of the potential value provided by annotations nevertheless few has been used in large scale to gather enough data to understand the user's behaviors and existing patterns during online task-free annotation practice.

2.2.1 Paper Annotations

We adopt the definition of annotations as set forth by MacMullen [Mac05] and Marshall [Mar97] - as *any additional content that is directly attached to a resource and that adds some implicit or explicit information in many different forms.* Annotations

[2] http://www.fleck.com
[3] http://sharedcopy.com
[4] http://www.diigo.com

may serve different purposes, such as: signaling a foreshadow, aiding memory and interpretation or triggering reflection. Additionally, annotations may occur in many different forms; for example: by highlighting, encircling or underlining text, we emphasize the importance of a certain part of the document; a strikethrough indicates that something is wrong, misplaced or not relevant; arrows signal relations between two or more elements.

Interacting with a document is known to stimulate critical thinking and reflection, a process that can be called 'active reading' [AD72], which is in contrast to passive consumption of text. In particular, text in the margin of a document may support a better understanding of the topic during later reading.

In [MR06], the authors draw a comparison between the early Hypertext pioneers visions and the present-day Web applications, commonly known as Web 2.0. The results of their analysis show that most of these systems support both private and public annotations and provide support for collaboration. Even though these features are identical with the first ideas of the Hypertext, the annotations are limited, because they reside exclusively bound to individual Web 2.0 services providers and they are not 'in-context'. More specifically, they are not visualized together and associated with the annotated content (the topic of interest), whose the benefits will be exposed later.

2.2.2 Social navigation

Social navigation support (SNS) describes techniques for guiding users through specific chosen resources [Bru01]. In AnnotatEd [FRBP08], the authors introduce two types of SNS: traffic-based and annotation-based. Our model is more related to the annotation-based style, in that every annotated page becomes a step in a trail.

Annotation-based social navigation support has been shown to be more proficient and reliable than traditional footprint-based social navigation support [FB05]. When the annotated resource reflects the interest of the annotator, it appends more value to the SNS. Annotation based SNS assists users in gathering information by making it easier to re-access the information and by showing the collective wisdom of the collaborators.

Allowing users to 'attach' their personal insights to a resource increases the reliability of annotation-based navigation support. Previous study of annotation-based SNS shows that users are particularly interested in being informed about resources annotated by others. Annotated resources are significantly more likely to be visited by users, specifically after being annotated [FB05].

2.3 Understanding Annotations

In order to understand how to better support active reading and annotations in the digital context, we carried out a study to compare how people annotate online with how people create paper-based annotations. Specific attention is given to the type of annotations, their function and perceived difficulties in creating and using these annotations. Before presenting the comparative study, we present some theoretical underpinnings.

2.3.1 Annotations in Learning

In this section we provide an overview on the role of annotations in learning. First we discuss a classification of different forms of annotation. We continue with a categorization of reasons why people annotate while learning. At the end of this section we explore various impediments for the take-up of annotation in the online context. Based on an extensive field research on textbooks, Marshall [Mar97] categorized the different kinds of annotations by forms and its functions. Below, we will discuss the forms of annotation that are relevant for learning purposes and their functions during the learning process:

- underlining or highlighting titles and section headings: this kind of annotation serves as signaling for future attention. Drawing an asterisk near a heading or highlighting it will remind the reader that there is something special about that topic, something to be considered or explored in more detail.
- highlighting and marking words or phrases and within-text markings: similar to above, the main goal is signaling for future attention - from themselves or from collaborators. The annotated pieces of text typically carry important and valuable observations. The act of highlighting text also helps in memorizing it.
- notation in margins or near figures: any kind of diagrams, formulas and calculations that structure and elaborate the document contents. This type of annotation is specifically meant to serve comprehension. An example is a calculation near an equation or theorem presented in a text, to quickly check its meaning and correctness.
- notes in the margins or between lines of text: these descriptive annotations are usually interpretations of the document's contents. These can be phrases in the margin that summarize or comment upon a section or a page. Single words are typically general terms, keywords and classification of a section.

Such annotations help the interpretation of the whole text where the reader better establishes the topic of the content of each part of the text creating his own mental

structure and decreasing the overall cognitive load. In all of these cases the value of annotations are for both annotators and future readers. Memory aiding, signaling attention, problem working and interpretation annotations definitely benefit the annotator but may also benefit other readers - provided that the annotations are explicit, readable and understandable. In collaborative group work, students typically work on the same content, but this content is extracted from different resources: for example, they all have their own copies of the obligatory textbook. This is a limitation inherent to paper-based annotations. Even though the annotations are still useful for personal use, they fail to play a role in the communicative and collaborative learning processes, which is a barrier for the leverage of learning by social constructivism [VC78]. Web 2.0 technologies explicitly facilitate these processes and their benefits on knowledge gathering and construction have been lately discussed [UBL+08]. Moreover, the exchange of documents, including annotations, remarks and insights, does not only serve the direct, content related goals, but also contributes to motivation and enjoyable professional relationships [LFK88].

Despite the many potential benefits of online collaborative environments in comparison with traditional paper-based annotation, there are several issues related to migrating reading and annotation to the computer. There is a vast body of research [Dil92, Has96, OS97, SH97] that discusses the many issues when moving from paper based reading to screen display reading:

- tangibility: in contrast to a text displayed on a computer screen, paper offers physical tangibility. Readers can hold the paper as they like, they can move it around to adjust their perspective and distance [Has96] in order to improve legibility [Dil92] and even to facilitate handwriting [Gui87]. Paper is also superior to electronic devices in terms of legibility. Further, while reading one page, readers can use another page for writing notes.

- orientation: paper documents give readers a better sense of location within the text, by physical cues, such as the thickness on the sides of a book or different paper materials in a magazine [OS97]. These cues support text skimming and cross-reading and they are instrumental when trying to relocate some text [Dil04, MW87]. Digital documents do not hold these characteristics [Dil92, OS97], an issue that needs to be overcome by increased attention for usability in device design and interface design.

- multiple displays: paper provides a single canvas for each page of text [Dil04]. Each one holds unique properties of physical tangibility, text content, modifications and additions from the readers. The virtual pages simulate this on the single device screen, but in some cases supporting concurrence reading from several documents turns to be an unwieldy task [OS97].

- cooperative interaction: by circulating a piece of paper, more than one person can interact with the content and build upon each others' annotations [SH97].

> Whereas groupware facilitates simultaneous revisions, versioning and collaboration, it does not yet reach the intuitive interaction as provided by circulating paper-based documents [MW87].

In addition to these usability issues, there are several technical issues that have been examined [MW87] to understand the challenge of digital reading. Here, we are mainly concerned with the implications for annotations. A major question is whether given the required progress in terms of technology and interface design - electronic annotations will be used in the same manner as the traditional paper-based annotations. From the above there is evidence that due to inherent differences when moving from the paper-based world to electronic devices, the character of annotations will necessarily change. Paper-based annotations have been used for centuries and can therefore be considered a highly developed activity, one that represents an important part of reading, writing, and scholarship. Annotation occurs in a wide variety of forms and it is applied for many different purposes. Annotations not only add substance to the text but also implicitly may reveal the reader's engagement with the material [Mar97]. Previous research has verified that no matter the form or purposes of the annotations, the benefits are immediately clear to the future reader [AGH+98]. Further, some researchers state that people's needs for making annotations in the Web environment do not differ significantly from their needs in the paper environment [GG99].

2.3.2 Web Annotations in e-Learning

The benefits and opportunities of electronic and automatic annotations, elaborating on their paper-based counterparts, have long ago envisioned by Vanevar Bush in the Memex [Bus45]. Bush envisaged that by relating all documents that users have read and attaching their annotations to these documents, individuals could organize and re-find information resources in an associative manner, together with any earlier annotations. Whereas the original rich forms of annotations in Hypertext systems with different categories, directions and even multi-links allowed for these associative trails, in the Web as it is today this functionality is not totally fulfilled, as readers have limited possibilities for sharing comments or questions by writing back to the pages. As a result, users spend a lot of effort trying to comprehend the different formats of how people comment on-line resources using coping strategies such as sending comments via e-mail [FCMS05].

Recent Web 2.0 technologies provide an open resource environment where individuals can freely collaborate. Nevertheless, these technologies typically only cover just a slight portion of the Web or one specific kind of annotation. These technologies are typically implemented as Web servers or browser enhancements. The basic idea of a Web annotation system is that the user has the ability to change, add or attach any type of content to any online resource, similar as she would do it with a paper

document. An application (usually a browser plug-in) enables the user to modify the Web pages, highlight parts of it and add tags or comments, while the back-end of the system just need to check these annotations and associate them with the specific user and the specific URL.

As discussed in the previous section, by actively being involved with the text, users can better memorize and understand it. By contrast, annotating on a computer-screen is an activity that competes with the reading itself, due to the lack of direct manipulation. However, users will do so when the benefits are higher than the costs in terms of effort. These benefits may include the saving of time needed for re-finding, summarizing, organizing, sharing and contributing online annotations. A rather economical view on the balance between the drawbacks and benefits has been given by Pirollis's information foraging theory [Pir07], in which the author described the above activities as information enrichment.

Today, both companies and academic institutions train learners to complete tasks and solve problems through project-centered learning. Since it may not be feasible for all participants involved in the projects to meet on a regular basis, they must be assisted by information and communication technology. To support this collaboration there are specific methods for Computer Supported Collaborative Learning (CSCL) provided by learning environments and other platforms can be adapted to fit this need. For the best results of the learning process, the methods should help each learner to act individually to reach her own goals and to cooperate by sharing and discussing ideas to accomplish an assignment.

As discussed in the previous section, in the same way annotations contribute for memory aiding, text interpretation and information re-finding, Web annotations provide the same functionality in the online environment. Web annotations are accessible anytime and anywhere, with diverse sharing possibilities, clearly enhancing workgroup collaboration [FB08] for cooperative tasks and learning processes. However it is important to remark that the full richness of paper annotations will only be achieved if the digital annotations hold the same beneficial feature of being 'in-context'. 'In-context' annotations are visible within the original resource, enhancing it with the observations and remarks of the annotator, which are likely to help in individual tasks in similar ways as is the case with paper documents [OS97]. Despite the limitations in terms of usability and tangibility, advantages of Web annotation tools go far beyond the advantages of regular paper annotations. In addition to the sharing capabilities within online communities, digital annotations can be indexed, ordered, rated and searched. These benefits are confirmed by several studies on annotations tools (e.g. [GG99]), in which participants have remarked that search the annotations is a very desirable feature.

Even though there are currently systems that support annotations, studies have shown that users often resort to different strategies for simulating annotation tools, making use of e-mails and messages to self and separated text documents. The main reason for this phenomenon lies mainly in the necessary effort required for creating

and organizing annotations: 'If it takes three clicks to get it down, it's easier to e-mail' [BVKKS08]. As users will inevitably resort to other strategies if annotation tools require too much effort, it is necessary to have a lightweight capture tool, with flexible organizational capacity, visibility and practical reminding. In particular if one takes into account that many annotations are primarily meant as temporary storage, or a means for cognitive support or as reminders, it becomes clear that these factors need to be better taken into account in annotation tools for personal information management and learning systems.

2.4 SpreadCrumbs: A tool for Web annotations

SpreadCrumbs is an in-context Web annotation system which has been implemented as an extension of the Mozilla Firefox Web browser[5]. The underlying assumption of SpreadCrumbs is that users can annotate Web resources with keywords or sentences and create hypertrails through a set of annotations. These annotations can not only be used for one's own reference, but can also be shared within a social network. The design of SpreadCrumbs has deliberately been kept minimalistic. Following the approaches seen in related work, we chose the basic visual metaphor for the annotations: Post-it notes. The Post-it representation has an optimized approach to simulate the most common paper based annotations forms namely underlining, highlighting and notation in margins. The idea is not to mimic different representations but to provide a way to achieve the same goals: signaling for future attention, comprehension and summarization. In addition post-it notes are extremely efficient as 'in-context' landmarks which are the main purpose of the research. Furthermore, by bringing the annotation behavior to the digital online environment we also add valuable features that are not applicable in the paper-based scenarios. The most prominent are the re-finding and the social sharing possibilities. The content of an annotation is easily searchable within the tool and shareable with other users.

2.4.1 The Browser Add-on

The SpreadCrumbs Browser add-on is a Javascript implementation based on AJAX principles. We used the AJAX and Javascript library from Yahoo, The Yahoo! User Interface Library (YUI) [6]. The library provides functionalities for drag & drop and other manipulations used in SpreadCrumbs. A simple client server architecture stores all the data on the server providing the user the possibility to access her data anytime from any computer where the client application is installed. Once the client add-on is installed to the browser, the user can access the sidebar. Through the sidebar the users have access to straightforward ordinary actions like creating account, profile

[5] http://www.mozilla.org/en-US/firefox/
[6] http://yuilibrary.com/

Figure 2.1 Web page annotated with SpreadCrumbs

management, login and logout. Additionally, the user has direct access to a contact managing webpage and a tabbed annotation-browser-window. From the right-click context menu an option is available to annotate the page, the same as from a small annotation button near the address bar.

2.4.2 Networking

As a non-mandatory step, new users may add their social network contacts to become collaborators in SpreadCrumbs. From the sidebar the users have access to the 'contact manager' webpage, from which they can import their contacts from their Facebook [7] Network using Facebook Connect technology. Once the contacts are imported they become part of the user's SpreadCrumbs network and the user is able to share annotations with her contacts. If at some point these contacts join SpreadCrumbs and grant permission to Facebook Connect; their accounts will be synchronized and all the annotations previously shared by some other user will be retrieved.

2.4.3 Annotating

Annotations (which we will refer to as 'crumbs') are added via the right-click context menu by the option 'Add Crumb', which results in the opening of a pop-up window that contains three fields: the receivers of the annotations, a topic and the content. By default, annotations are private. An auto-completion drop-box helps the user in adding receivers from her contact list. Once the annotation is created, a post-it note appears in the screen, originally on the clicked spot but easily relocated by drag and drop (Figure 2.1).

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

When any of the involved users in the annotation accesses the annotated website, the post-it note will be displayed. Additionally, if the user keeps her connection to Facebook through SpreadCrumbs, the receivers of the annotation will get a notification on Facebook and a notifying e-mail about the new annotation.

2.4.4 Reacting

Each annotation is an entity in a thread (a crumb in a trail) and diverse actions can be taken over it. When visualizing an annotation, any of the involved users has the ability to interact with it: moving it around, closing it, following trails and replying.

2.4.5 Connect and disconnect

Each user has her individual status in the context of one annotation. The status 'Connected' is the normal status to visualize the annotations; 'Disconnected' means that she will not visualize the annotation anymore once she comes back to the website; and 'Stand by' means that she will not visualize the annotation again until some modification has occurred in the annotation thread.

2.4.6 Replying

The reply link on an annotation brings up the same window pop-up as adding an annotation offering to the user just the content field to be filled. Once confirmed, the reply is attached to the first post-it note and the same notifications actions are triggered. Any user involved in the annotation is able to add a reply to the running thread, which is visible to all participants. This action simulates a micro in-context forum on each annotated web page.

2.4.7 Following trails (SNS)

What makes SpreadCrumbs unique is that the annotated pages are not simply a loose collection, but the resources become interconnected. Each annotation is associated with links that can be followed from the crumb: the user trail and the topic trail. Near the name of each user who annotated the page and near the topic text there are two small linked arrows indicating the path to the previous and to next annotation in the hypertrail. Following the previous/next link next to the name of a user will redirect the current user to the next/previous annotated page where both users share another annotation.

Following the topic trail will lead the user to web pages on which the user has annotations with the same topic description. A simple illustrative example: one user privately annotates five different pages with the topic 'Conference' adding specific

content for each annotation. Once it is done, each conference page annotated has a link connecting to each other. A temporal defined (and connected) collection of web resources was created and at any time the user is able to remove, edit or add new stop points in this trail. The final output is a simulation of the Memex [Bus45] idea where the resources are now annotated and associated in accordance with the user's preferable organization.

Providing sharing capabilities of these trails, SpreadCrumbs provides Social Navigation Support in a very concrete and defined manner. Differently from others SNS systems, the resources are not only a collection of links but they have a well-defined temporal order, each resource becomes interconnected and they hold in-context insights from the annotation authors.

2.4.8 Browsing Annotations

The SpreadCrumbs' sidebar contains a browser pane with three different tabs that shows the three facets of the organizational dimensions of a trail: *topics*, *pages*, *people*. Additionally, a small pane in the bottom shows detailed information on the selected trail.

The tab *topics* shows the trails grouped by topic description. The user visualizes distinct items that represent the different trail-topics she created. From this pane, the user is able to access the annotated page, edit the topic description and change her status in the topic. By clicking or selecting one of the topic-trails, the bottom pane loads and displays all the crumbs belonging to this trail assembled by page. In this pane, the user has the same possibilities to directly access the annotated page, to edit the crumb and to reply it.

The second tab, *page*, shows the trails grouped by the resource annotated. The visualization has the title extracted from the Webpage and the trail last modified date as well. The user has the possibility to edit the name of the page, if she wants to. It is important to notice that, although trails mainly contain the same page title, in this facet they will not be grouped together, since the grouping is based on the URL location of the annotation. By clicking or selecting one of the page-trails, the bottom pane loads and displays all the crumbs belonging to this trail, assembled by the different existing topics on the selected page, with same management capabilities.

Finally, the *people* tab shows items that represent the trails from the user's contacts. The *item* visualization shows the name of the contact and her last activity on the trail. It also indicates whether the contact is already connected to SpreadCrumbs' network or not (due to the fact that is possible to share annotations to imported contacts that are not subscribed to SpreadCrumbs). By clicking or selecting one of the people-trails, the bottom pane works in the way as the *topics* tab previously described.

Table 2.1 Annotations found by type.

Annotation types	Total	%
Highlighting/Mark sections headings	153	8.6%
Highlighting/Mark text	1297	73%
Problem solving	2	0.1%
General notes (Notes in the margins)	326	18.3%

2.5 A Comparative Study on Paper-Based and Online Annotations

The main goal of this study is to investigate the types of annotations encountered online and on paper, and to find differences between these two situations. The results of this study are expected to provide insight in differences between these two situations and to provide design guidelines for the design of annotation tools and the way they are used.

2.5.1 How People Annotate on Paper

To compare annotations in the online context with paper-based annotations, we visited the working place of 22 PhDs students and pos-Docs. We asked each one of them to take a look at the last 3 research papers or articles that they have printed and read. In total, we have collected 66 articles covering a total of 591 pages of text. We found 1778 annotations and an average of 3.08 annotations per page. Table 2.1 shows the average of each type of annotation per page.

The far majority of the annotations (73%) involved the highlighting and marking of text. Some participants had the tendency to only highlight main words within a sentence or paragraph. In these cases we counted the collection of highlighted words belonging to a continuous block of text as one piece of annotation. 9% of the documents discussed with the participants turned out to be part of collaborative work in which two or more people were involved. All except two participants reported that they shared their comments via email or some online communication tool; only two participants shared the same sheet of paper, which contained annotations from both parties. Another valuable observation is that all of the participants who share annotations said that they do annotate in a different (more careful) way when they annotate concerning another reader.

To examine in more detail the annotation strategies, we asked our participants to classify the goal of reading the paper. We distinguished between the following categories: *reading for writing, reading for learning, reviewing and other. Reading for writing* is the common activity of reading related articles to extract ideas and references specifically for purpose of writing. *Reading for learning* includes the act

Table 2.2 Results by reading goal.

	Writing	**Learning**	**Review**	**Other**
Articles	31	23	9	3
Articles annotated	28	16	7	3
Annotations/Page	2.36	4.7	1.11	6.3
Annotation types				
Highlighting/Mark sections headings	10.5%	7.5%	9.4%	4.8%
Highlighting/Mark text	66.0%	82.9%	40.6%	72.2%
Problem solving	0.1%	-	0.9%	-
General notes (Notes in the margins)	23.3%	9.6%	49.1%	23.0%

of getting updated in some particular field, read about new publications or learning some new approaches to apply in some other activity, such as solving math problems or implementing algorithms. *Reviewing* consist exclusively of reading papers to give feedback to the author. Finally, any other type of reading was categorized as *other*. Table 2.2 shows some numbers of the field research by the type of reading activity and Figure 2.2 depicts some annotations collected.

In addition to comments directly put on paper, three participants also used the technique of attaching annotations to the original document with post-its that were attached to the paper. From the 66 articles analyzed, 10 (15%) did not contain any annotation. One participant that did not have any annotation in any printed paper said that she keeps her annotations in a separated file in her computer for each digital article. Two other participants said that they first do a very quick reading on the computer to check the relevance of the text, and if it looks relevant they print it. In their own words: 'First I read on the computer to see if I really need to print'. We have noticed that in many cases participants also used different marking colors for highlighting with the purpose of attributing different levels of importance. From the annotations we identified many different ways of signaling important parts on the text. As an example, one participant created her own symbology for annotating: *squares* around the terms means new terminology, *underline* means definitions and *circles* means open question or issues over some topic. Those annotations symbols were used combined with highlighting (importance) and many times they even overlapped. One last interesting observation was the behavior of one of the participants who keeps two printed versions of every paper: one with annotations and one clean print. As stated, the clean print is for a future reading when she may want to get the idea without influence of her previous readings. Although the vast number of highlighting annotations on the papers, none of the participants use such mechanisms that allow persistent highlighting on digital documents or web resources.

In summary, from the observations we identified two main clusters of annotations: *relevance adjustment annotations* where implicit highlight and signaling indicate dif-

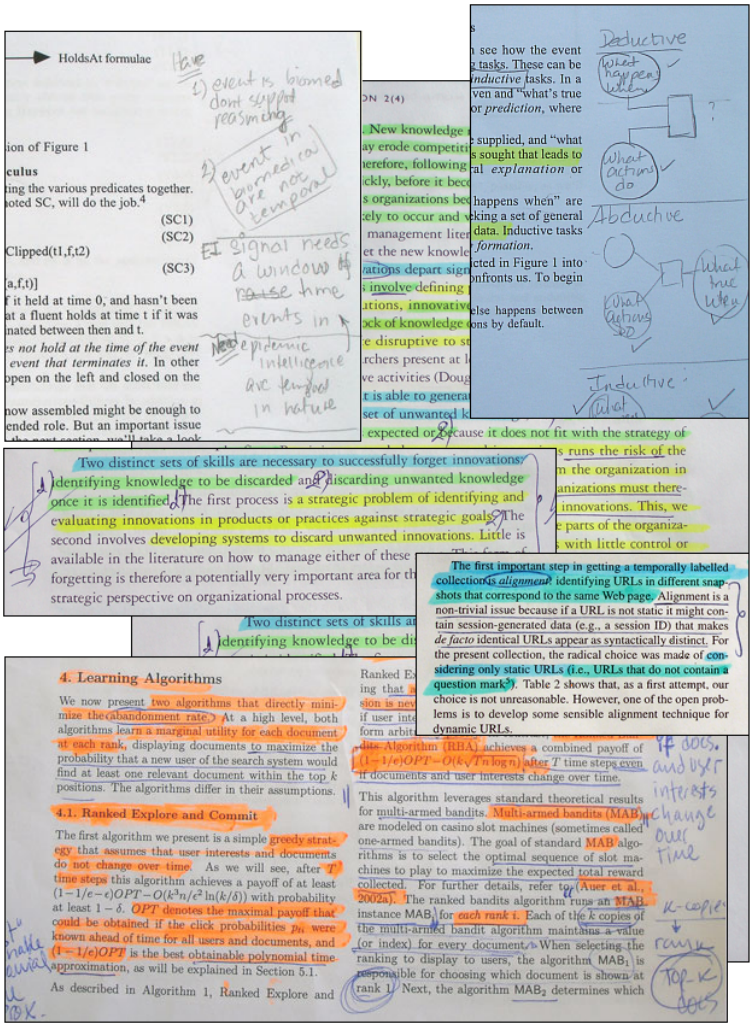


Figure 2.2 Examples of annotated papers examined during the field research

ferent levels of importance in the text and *contributive annotations* where explicit readable remarks are added attached to the text.

As a last part of our interviews, we asked the subjects to describe how they arrange their papers that lay on their desktops. The relevant categories described were topic, quality, importance, date of reading and task. This simple observation may guide us to design better metaphors of the possible dimensions when trailing online resources.

2.5.2 How People Annotate on the Web

The experiments with our annotation tool were conducted with 18 participants, who all stated to be very proficient working with computer and internet technology. From those, 16 are working in the field of computer science.

At the beginning of each session, in which only the participant and the experimenter were present, the tool was introduced to the participant by giving a brief overview its usage. Following the introduction, we asked the participants to answer a set of 10 questions by writing down the answer and annotating the resource. These questions were specific information finding tasks that could be solved by a brief internet search with any popular search engine. We ensured that most of the questions were very specific domain questions or numerical in nature to reduce the possibility of the participants to know the answers an example: 'What is the estimate percentage of Chinese among the population of Brunei?'. The experiment setup enforced the participant to annotate useful but hard to memorize information for future reference.

During the experiment, the participants created a total of 207 annotations, covering 81 different Web resources. The average number of words per annotation was 4.1. An important observation was that the participants in general carefully positioned the annotations in the context of the Web page: from the 18 participants using SpreadCrumbs, 16 placed the annotations of each question near the text, table, or paragraph where they found the answers. This type of behavior is not supported by the simple bookmarking functionality of regular browsers.

We noticed that out of the 18 participants who used SpreadCrumbs, only six of them included the answers in the annotations while the majority opted for using keywords of the respective question. Just one participant typed explicit full sentences when annotating the pages: 'There seem to be different walks - I'm not sure whether the 9.4km walk brings us to the top, but I think so.' ; '.. made 35 homeruns in 2005. Yes, I think this should be the right answer.'

Although the participants were very proficient with the computer, all of them stated that they regularly print digital documents for reading, even when these documents are relatively short (up to 8 pages). All of them confirmed that they usually annotate those printed documents in one way or another, by means of highlighting text and adding their own comments or insights in the margin.

This somehow contradicts a very interesting observation during the experiment. One of the answers consisted of a short passage from a book (2 sentences with less than 40 words). However, all of the participants demonstrated laziness when having to write down the quote on paper. All of them asked the same question: 'Do I have to write the whole sentence?'. We allowed them to write down only the reference for the passage (page and paragraph), a suggestion that was followed by all of the participants. The contradiction arises since the participants do not desire to write if they have the option of typing (or copy and paste). However, they keep their behavior of annotating with the pen even though several means of digital annotation exist.

None of the users demonstrated problems regarding the usage of the tool. After the short introduction, all of them performed the tasks of annotating and consulting annotated resources without any effort or mistake. The participants demonstrated enjoyment with the tool interface and functionalities. The direct manipulation and the 'in-context' features were the most appreciated. After having conducted the tasks,

Table 2.3 Example of personal and shared Web annotations.

Personal	**Shared**
'Conference Deadline: October 29'	'All artists are from Sweden, I think, and do Jazz music (quite soft) but nice...'
'Flat 64m 2 rooms windthorststr. 8'	'Let me know if there's anything else to be done.'
'TO DO!'	

the participants were handed over a questionnaire in which they had to choose terms from a list of adjectives. This gave us a dataset of the user perspective over the tool. This questionnaire [8] measures usability and satisfaction with a list of 118 adjectives, positives and negatives. This methodology gives the participants more confidence to be critical to the system choosing negative terms. The top 10 terms chosen were: *Easy to use, Usable, Useful, Collaborative, Helpful, Convenient, Connected, Friendly, Innovative, Straight Forward*. These results sctrongly suggests us that the participants would be willing to use such tool on a more regular basis.

Regular use of SpreadCrumbs

In addition to the laboratory study, we collected and analyzed log files from users that were not involved in the experiments. The results show some interesting differences that distinguish two behaviors when annotating. Examining 177 shared annotations, we identified an average length of 10.35 words per annotation, whereas from 371 personal annotations we found an average of 4.56 words per annotation. With the permission of the users we extracted some examples of annotations that illustrate these numbers and the difference between the linguistic structures of the notes - see Table 2.3.

The examples of personal notes show that these private annotations in many cases contain a rather short, cryptic message. These annotations typically just consist of keywords or some sort of reminders for the authors, of which the purpose often is only understandable by the users themselves. It should be noted that these keywords should not be mistaken for tags. While tags have a descriptive nature, these keyword-based annotations carry additional (sometimes implicit) information. By contrast, shared annotations are very explicit and well-described with full meaningful sentences, in form similar to chat or text messages.

2.5.3 Discussion

From the results presented above, we can sketch some impressions on some user's behaviors. Apparently, the high amount of highlighting/marking signifies 'laziness'

[8] http://www.userfocus.co.uk/articles/satisfaction.html

of the annotators. This laziness is in fact a way to reduce cognitive overload (because of switching between tasks) and to keep focused on the main task (the reading itself) while still providing meaningful cues. The higher amount of annotations per page (for the 'learning' papers) shows that these annotations have a clear function for memorizing certain parts of the text (by actively doing something with it).

The category of 'review papers' shows a higher frequency of notes in the margin comparing to the other categories. These are almost certainly comments to be included in the review. Additionally, the low number of highlights clearly shows that the readers are not concerned about signaling for future attention. Out of this we draw the conclusion that there is indeed a significant difference between the goals and behaviors of paper and digital online annotations. The papers that had higher amount of notes and the lower number of highlights (as explained before, an action that means signaling for future attention) indicate a non-concern of the reader about a future reading. On the other hand, online annotations (notes in the margin as used in the experiment) are mostly used on resources that are meant to be reused and found in a future work session. We conclude that, although online annotations are similar in its structure to margin notes, its scope is more comparable to highlighting where the real main goal remains in signaling for future attention and facilitation for re-finding.

Within the collected data of online annotations, the average number of words (4.56) in private annotations does not cover the average length of short sentences while the shared annotations (average of 10.35 words per annotations) fit the average of short and medium sentences statistically measured in plain text document [Alt88]. We deduce that private annotations, in general, don't contain full sentences and as in the paper based texts they are just a perspective over the topic context or keywords and classification of a section (or resource) - in the digital environment mostly used for re-finding. The shared online annotations clearly hold more explicit meanings where the authors tend to be clearer when sharing their thoughts. This evidently shows the different behavior and concerns of the individual when writing personal or shared annotations. Although differences have been found between paper and digital annotations, if we use the same reading goals classification for online readings and translate the annotations meanings, we find out that in-context notes annotations are the optimized form for attention signaling, summarization, interpretation and improving bookmarks search, in both personal and shared environments.

The sum of these two studies suggests some design implications for annotation systems. First of all the annotation action must be effortless in all senses - easy to access and visualize, as few interactions as possible and in-context interactions to minimize the lose focus. Online resources can be used for all sorts of reading tasks, thus annotation systems must supply all forms of annotations, not by similar representations but by providing the means to achieve the same goals. The necessary effort still requires some engagement from the user, however the benefits discussed should overcome and become in hand to the users: re-finding tools, easy manipulation

and organization of the annotations and resources and sharing capabilities.

2.6 A Comparative Study for Refinding Information

In order to validate the usefulness of in-context annotations, we set up an evaluation that compares different means to refind information. The goal of this study is to quantitatively estimate the efficiency and ease-of-use of three different refinding methodologies: web search, bookmarks and in-context annotation. Regarding the tools instantiating these methodologies, we selected the Del.icio.us social bookmarking service and for in-context annotations we used the SpreadCrumbs tool. For web search, the participants were free to make use of the search engine they were more familiar with; all participants turned out to prefer the Google[9] search engine.

2.6.1 Systems

Del.icio.us is a popular online social bookmarking system. The system combines bookmarking and tagging with social networking, features that turned it into one of the most successful social bookmarking services. With Del.icio.us a user can annotate bookmarked Web pages by tagging and adding comments to them, while also having the additional option of sharing them with her contacts. A crucial difference with annotation tools is that the information added or tagged by users can only be viewed and managed through their personal home page in the Del.icio.us' web site; therefore, they are far from being in-context. Figure 2.3 shows a sample of the Del.icio.us' bookmarks and its comments collected by of one of the participants in the user studies.

On the other hand, the main idea behind SpreadCrumbs is that a user can annotate Web resources with keywords or sentences. These annotations can then be used not only for her own reference, but also for sharing with her social network. The interface of SpreadCrumbs has deliberately been designed in a minimalistic way, so that users get easily acquainted with it. Figure 2.4 illustrates a page annotated in-context by one of the participants using SpreadCrumbs. Spreadcrumbs uses the basic visual metaphor of Post-it notes for the annotations following the best approaches proposed by relevant annotation tools, including Anchored Conversations [CTB+00], StickyChats [CTBN00], MapChat [CGO08], Keyholes [NSC08]and the Fluid Annotations Project [ZBJM01, ZCM98, ZRMC00]. As Post-its are quite popular in real life, SpreadCrumbs offers a very familiar way of adding inscriptions and remarks to web content. Moreover, it is transient in the sense that it is easily replaceable and not disruptive. The following subsections elaborate on the two sessions of our experiments in greater detail.

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

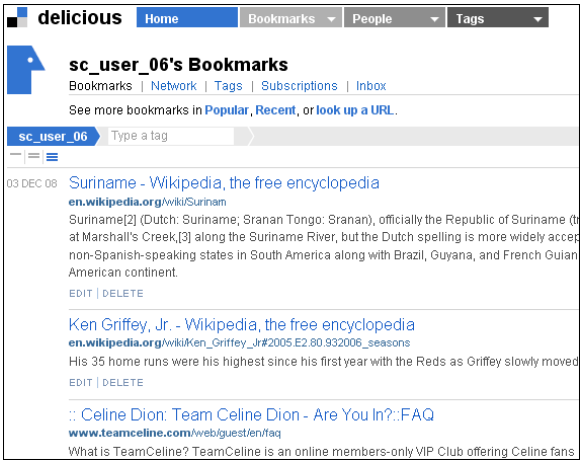


Figure 2.3 Sample of Del.icio.us' bookmarks of one of the participants.

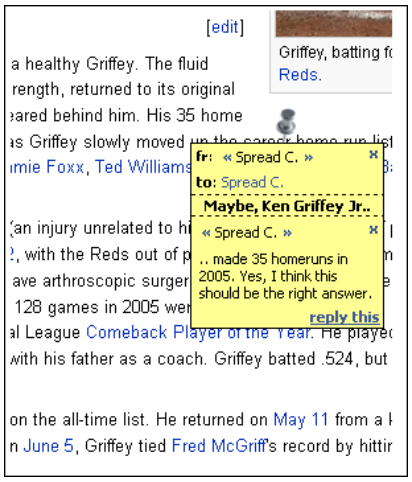


Figure 2.4 A SpreadCrumbs' annotation on Wikipedia of one of the participants.

2.6.2 First Session: Find

As mentioned above, the purpose of this stage was merely to have the participants locate specific pieces of information, which they will be asked to refind in the second phase of the experiments.

Participants and Settings

Following the same pattern of the majority of relevant user studies, we designed our experiments to be comprised of two sessions; the first one includes all finding tasks, leaving all refinding tasks to the second session. It should be stressed here that the second session took place five months after the first one, thus allowing for the examination of long-term refinding behavior. The participants' pool consisted of 34 people in total (24 males and 10 females), who are aged 27 years old on average. Additionally, it is worth mentioning that all of them are quite proficient in on-line search and web technologies in general, due to their education background (the majority being graduate and PhD students in the field of Computer Science). In the context of the first session, we randomly divided the participants into two equivalent groups, with each one designated to a specific tool: Del.icio.us or SpreadCrumbs. To ensure familiarity with application at hand, each individual was initially presented with a short tutorial of its features and functionalities.

Procedure

During the experiments, the participants were asked to locate the correct answers to 10 questions. All of them were specific information finding tasks that could be solved

by a brief internet search with any popular search engine. They were in fact chosen at random from a set of 16 questions that we had carefully prepared for the experiments. We ensured that the questions were sufficiently obscure, to minimize the chance of participants knowing the answers themselves; most of the answers were numerical in nature - an example question is: 'How many homeruns did Ken Griffey Jr. hit in 2005 playing for Cincinnati?'. After finding the information, the participant was asked to annotate the web page that contained the answer for later reference (as seen in Figure 2.3 and Figure 2.4). It should be clarified here, that there was no predefined set of acceptable web pages, as subjects were free to mark any web resource they wanted. The first task was presented as 'just' an exercise in order to get used to the system. In reality it was a preparational step for the second round of the experiment.

Primary Observations

In the subsection below we discuss some observations from the first session. Upon completion of the first task, we noticed that, far from exhibiting homogeneity, each participant followed her own approach in creating annotations: some included the answers in the annotation text; others added the questions, while the rest of them used just some keywords. In addition, they followed different strategies for positioning the annotations, as not every participant was concerned with placing them in a useful location. Although most of individuals carefully posted them near the text, table, or paragraph containing the answer to a question, few of them just added the annotation on the top of the page or over the margins. The latter are thus expected to experience some overhead in the course of the refinding task, especially in the cases of answers residing within web pages with a great deal of information and unstructured content in general.

2.6.3 Second Session: Refind

During the second session, five months after the initial session of the experiments, the participants were asked to relocate the answers they had previously found during the 'preparational' task of the first round. This long time interval ensures that the participants remembered neither the answers they had provided nor the resources they had used in the course of the first session.

Participants and Settings

In total, 30 participants (21 males and 9 females) were involved in this phase of our study, out of the initial 34. They were asked to repeat the same tasks as the first time; in other words, each individual was given the initial set of 10 questions and had to relocate the answers she had given in the first round. The participants were divided into three equivalent groups of 10 people, each one corresponding to a

specific refinding methodology and tool. The first group corresponds to the search engine approach and its members were allowed to employ solely search engines in their efforts to carry out their tasks.

This group was formed by randomly choosing 5 participants from the bookmark group together with another 5 from the annotation group of the first session. They were not allowed to use the bookmarks or annotations they had already created, which implies that they could refind information only by searching and browsing from scratch. This group served as the baseline group during our analysis.

The second group represents the bookmarks approach and consisted of those subjects that used Del.icio.us both in the finding and the refinding sessions. The members of this group had the URLs of the visited resources at their disposal, saving in this way the burden of repeating the procedure of the first session. Additionally, some of them had added comments to their bookmarks, which invariably provided them with valuable clues for quickly relocating the answers. Finally, the third group corresponds to the in-context annotation approach and was comprised of those participants that used SpreadCrumbs in both sessions. The URLs of the initially visited resources were thus available to them, similar to the bookmark group. Further, they were also assisted in their task by the annotations that they had composed during the first round of the experiments. However, as mentioned before, there was a great diversity not only in the content of these annotations, but also in their positions. We expect these two factors to influence the performance of the participants. We anticipate, however, that their performance will be mostly affected by the dynamic nature of the Web. In other words, due to the long period of time over which the experiments stretch, there will inevitably be a considerable number of misplaced annotations that may cause inconvenience and delay to the subjects. This is actually the factor that has been mainly ignored in all previous studies, since there are rarely any modifications in the content or the layout of a page causing annotation misplacement over a short period of time.

Procedure

During the second phase, the participants were presented with one question at a time, chosen randomly, so that the order of questions is different from the one used in the first round. In this way, even the participants of the last two groups that were assisted by an application, had to devote some time to pinpoint the appropriate bookmark or annotation in their collection. After the appropriate web resource was found, thus completing the searching stage, the participant had to locate the answer in the page and highlight it using the mouse (browsing stage). There were no instructions or restrictions as on how to proceed with this stage; the participants were allowed to perform this task the way they would in a non-controlled environment. It turned out that the vast majority of the participants took advantage of the browser's 'find' functionality, which rapidly locates and highlights the given words in the page in

view. This functionality was not only used in conjunction with some keywords taken from the question, but sometimes also with the whole answer, since some subjects had it included in their bookmark comments. In this context, and especially in the latter case, browsing time is minimized, particularly for the users of Del.icio.us and to a lesser extent for those of the search engines.

Once the desired piece of information was highlighted, the participant was given the next question. Upon completion of all tasks, the subjects were asked to answer two questionnaires, one regarding the information refinding experience and another one investigating their opinion on the tool they used. The necessary data for estimating and evaluating the average and overall browsing time per individual were collected with the help of screen capture and data-logging software that recorded all participants' actions.

2.6.4 Results

In this section we discuss the outcomes of our experiments, which are mainly concerned with browsing time - the time participants spent in the browsing phase while carrying out their task. In other words, our analysis focuses on the period of time that starts as soon as the page of interest finishes loading and ends the moment that the participant finds the required information. We begin with the analysis of the time measurements that were derived from the 297, in total, refinding activities. The corresponding tasks are evenly shared among the three groups mentioned above the Search Engine, the Bookmark and the Annotation groups. That means that the performance of each group is represented by 99 time intervals expressing the duration of the tasks involved.

Browsing Time Measurements

The most appropriate metric for expressing the overall performance of each group is arguably the average time taken to complete the browsing phase therewith ignoring the time it took participants to locate the page in the searching phase. In our case, the available sample of 99 browsing times produces the following mean values: 46s for Search Engine, 38s for Bookmark, and 21s for Annotation. With an average mean of 21 seconds, the annotation group was significantly faster than the bookmarking group (38 seconds; $t(98)=3.88$, $p<0.01$, $r=.36$) and the search engine group (46 seconds; $t(98)=4.07$, $p<0.01$, $r=.38$). The differences between the two latter groups were found to be non-significant. It turns out, therefore, that the performance of Annotation is substantially better, corresponding to a time that is almost the half of the other two groups. This suggests that in-context annotation boosts refinding to a great extent. By contrast, when comparing the performance of the first two groups, the outcome does not match our initial expectation that Bookmark would outperform Search Engine due to the wealth of cues associated with them, i.e. the comments

that were attached to bookmarks as well as the keywords of the tags that were drawn from the questions or even the answers. This can partially be explained by the theory of context-dependency [17], arguing that all context knowledge acquired in the refinding process serves as relevant cues for refinding information. , This includes even the non-semantically related elements located within the target information that search engine users acquired while searching and browsing the search results. Bookmark users, on the other hand, had to acquire the context during the browsing stage itself. It should also be stressed that the performance of the search group would have been significantly worse if we also took into account the searching stage, which is minimized for bookmarks' users.

CTRL+F

Thus far we have focused on the effect of the diverse tools on users' efficiency, thus ignoring another important factor: the use of browsers' 'find' functionality. This functionality plays a major role in relocating a specific piece of information within a web page. In order to quantify its degree of use, we measured the percentage of each group's tasks that were carried out with its help: 53.5% (Search Engine Group), 62.3% (Bookmark Group) and 17.2% (Annotation Group). Apparently, CTRL+F has been extensively used by the subjects of Search Engine and Bookmark, whereas participants of Annotation resort to it less frequently. They actually use it solely in the cases of modified web pages that result in misplaced or orphan annotations; in these cases annotations are anyway of little help and the user has to resort to other means for pinpointing the desired information.

Judging from the wide use of the CTRL+F strategy, it is reasonable to assume that 'find' helps participants to perform better in refinding information. To verify this assumption, we estimated the average browsing time that corresponds to subjects using it and compared against the browsing time of those that did not use it. This comparisons were made in the context of all three groups and their outcomes, presented in Figure 2.5, suggest the opposite: participants that took advantage of this functionality needed significantly more time in completing their tasks than those that did not. Hence, although this functionality is supposed to constitute a quite handy tool for locating information, in practice there is no evidence supporting its beneficial contribution to re-visitation efficiency.

Page Content Size

As mentioned in the previous subsection, there is likely to be a correlation between the length of web pages and the usage of CTRL+F as a means for refinding. To investigate this correlation, we divided the participants of each group into two subgroups; one that used CTRL+F, and one that did not. For each of these subgroups we estimated the average size of the accessed web pages in terms of number of words and calculated

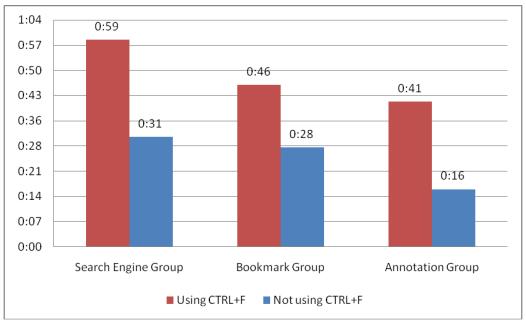


Figure 2.5 Average times of each group distinguishing tasks where the browser's 'find' functionality was used.

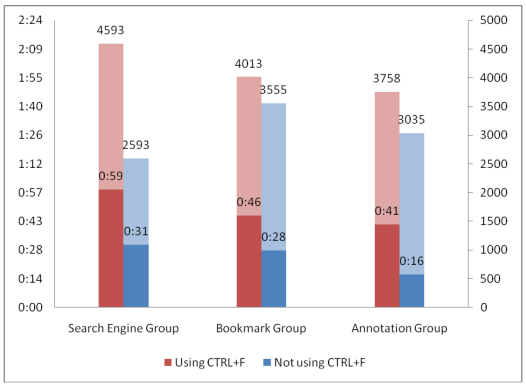


Figure 2.6 Average times of each group and average Web page sizes (number of words) distinguishing tasks where the browser's 'find' functionality was used.

the average browsing time - see Figure 2.6. The figure shows that, in particular in the Bookmark and Annotation groups, there is no interdependency between page size and the usage of CTRL+F. However, it shows that the browsing time is significantly higher in the CTRL+F condition, which suggests that the find functionality does not sufficiently leverage the detrimental effect of long and possibly unstructured pages.

Discussion

In this section, we evaluated three different approaches of refinding information, namely web search engines, online bookmarks and online in-context annotations. The main focus of our study was on the reading and browsing phase that follows the searching stage of this process. The outcomes of our experiments suggest that bookmarks and annotations clearly outperform search in terms of performance. Moreover, we observed a clear benefit of in-context annotation compared to bookmarks in terms of content recognition.

We also investigated the correlation of the browser's 'find' functionality (CTRL+F) with refinding efficiency and observed that it does not actually account for any improvement in the browsing time. The questions for which 'find' was used typically involved larger and unorganized pages; however, the expected benefits in terms of saving time could not be observed. For this reason, it would be beneficial if annotation tools could reduce this burden by minimizing users' cognitive load, interactions and wasted time. Moreover, the structure of web pages was proven to have a substantial impact on the efficiency of the refinding process, with well designed web pages ameliorating it. On the other hand, web pages with practically no structure apart from lengthy texts impede users to a great extent in their finding and refinding efforts.

From the background research it has become clear that the act of annotating supports the learning process in paper-based situation. However, when it comes to online learning, annotation becomes an additional cognitive burden, due to the lack of suitable tools and intrinsic problems related to reading from a screen and interacting via keyboard and mouse. From the comparison of online annotation with paper-based annotation it becomes clear that there is a difference between both types. Online annotations were typically short and had a certain purpose in terms of re-finding, sharing or commenting. The high amount of highlighting in paper-based annotations has an intrinsic value. Based on the results we conclude that emphasis in the development of annotation tools should be put on added value by better exploiting the annotations (for example for enhanced re-finding tools, visual overviews, grouping, sharing, collaborating) rather than to try and mimic the 'old-fashioned' paper-based annotation. At the same time, writing an annotation should cost as little effort as possible, as otherwise people will inevitably resort to other ways of getting their things done.

2.7 Annotations in Real Scenarios

In this section, we apply the experience and knowledge acquired in the field of Web annotations to deploy an annotation tool in a real scenario. Briefly, we developed an annotation tool that allows users to highlight excerpts of learning objects in online courses. Additionally, we present quantitative results of usage from a community of over 750 learners subscribed to the courses where the tool was available, together with qualitative results collected from a questionnaire.

2.7.1 Online Courses

The distance course 'Technology applied in Education' is designed for postgraduate students who wish to achieve literacy not only in information technology, but also to deepen the knowledge of it in the classroom. The course is aimed to those who are teachers in the educational public network in Brazil and aims to generate knowledge, promote teacher development and educational reform.

The first two editions of this course were held in 2006-2007 and 2009-2010 and resulted in the specialization of over 6000 teachers, distributed throughout Brazil. Although the main structure of the course is kept the same, each version of the course incorporates new tools and means of communication available, in order to suit current needs and prepare teachers to use and create new learning situations in their future lectures.

The 'Technology applied in Education' course is available all over Brazil and in the current version has over 750 subscribed students. Along with, the course has over 50 tutors that are responsible for monitoring, evaluating and teaching through

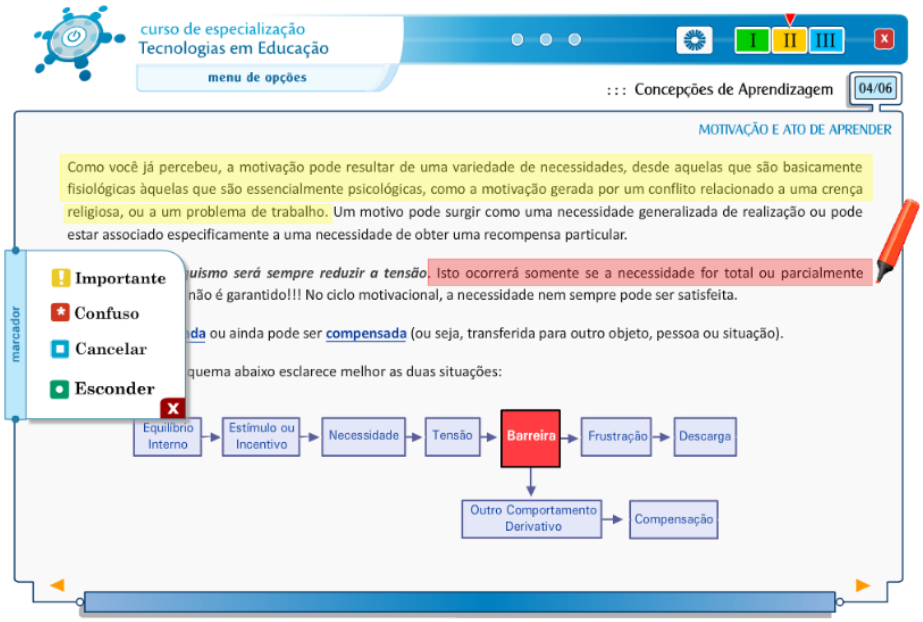


Figure 2.7 Examples Student Module. Example of annotations realized by a student. On top, an annotation marked as important(yellow) followed by an annotation marked as confusing(red).

our Learning Management System (LMS). Each tutor has a group of maximum 30 students. The course is delivered through online lectures, discussion forums, Web seminars and practical projects that support learning by doing.

Accordingly, we have deployed, in each edition of this specific course, several tools that can help students and teachers in the learning process. In this manner, here, we introduce the 'highlight tool', a simple yet powerful annotation tool.

2.7.2 The Highlight Tool

The highlight tool consists of two main modules, Student Module and Staff Module. The Student Module is responsible for recording all the annotations done by a student. The process is triggered at the moment the student selects one of the available highlight pens (*confusing* or *important*).

For matters of simplicity and usability, we adopted only the two semantic-annotations types mentioned before. However, the tool can be customized to use different colors and semantic-annotations types. Furthermore, before start the use of these tools, we introduced a brief description of its usage in order to ensure their understanding about each semantic-annotation type. Once the annotation is done, the annotated area is recorded in the LMS database. Figure 2.7 depicts the tool in use.

2.7.3 Evaluation

In order to evaluate the Highlight tool, we collected quantitative usage statistics and qualitative feedback from the learners. To assess qualitative feedback we set up an online questionnaire with 17 questions. We distributed the questionnaire to all students subscribed in the online courses that had access to the highlight tool. The questionnaire was not mandatory and was completely anonymous. We divide the questionnaire in five different tiers of questions, namely: usage, satisfaction, application, collaboration and future use.

- **Usage.** The questions regarding usage collect feedback regarding the students' access frequency to the online courses, the usage of the highlight tool and revisitation to the annotations.
- **Satisfaction.** Satisfaction covers the students' personal feeling regarding the tool concerning utility of the tool, if it supported their studies and easiness of use.
- **Application.** The questions of the application tier collect feedback considering possible applications and activities to be done on top of the annotations. For example, 'Is it beneficial to provide extra material for the annotations that are marked as important/confusing?'.
- **Collaboration.** In terms of collaboration, although the tool did not provide any means for the students to exchange annotations, the questionnaire inquire them about their will to share and collaborate with other students. As one example, among the collaboration-related questions, we asked if they would like to have access to other students' annotations.
- **Future Use.** Finally the questions regarding future use address general opinions and inquire the students about their desires and plans to use the highlight tool in future courses.

For each question, the participants had to choose their agreement on a 5-point Likert scale.

2.7.4 Results

We collected the data from the first two courses where the tool was available. Each course consists of the main document - the one that can be annotated - together with other activities previously described, and has a time-span of approximately eight weeks. We gathered the students interaction during these 8 weeks. The first course consisted of a document containing 43 pages, while the second had 65 pages. In total we collected 279 annotations where 88% were marked as important. In Figure 2.8

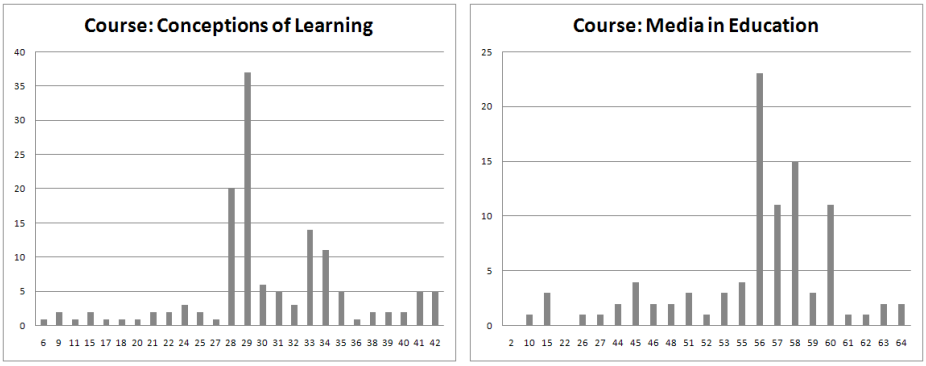


Figure 2.8 Distribution of the annotations marked as important (y-axis) by page (x-axis) in the content of two distinct courses.

we discriminate the annotations (important) by course and by page in each learning object. We did not find any correlation between the number of annotations marked as confusing and important. A thorough analysis of the portion of the most important-annotated pages, revealed us that the contents mainly contain definitions of concepts significant to the respective course.

In total, 132 students answered the questionnaire. In Table 2.4 we compile the answers distinguishing them by tear and agreement. Over 75% of the students that answered the questionnaire stated that they often (or very often) accessed the online content, however only 25% stated to use the highlight tool with the same frequency.

From the Satisfaction tier, over 77% agreed or strongly agreed that the highlight tool contributed to their learning process. Also, over 75% of the participants considered the tool straightforward to use.

Regarding the annotations and further activities that should be provided to the students, over 50% of the questionnaire participants agreed (or strongly agreed) that it is important to have further materials, discussion forums and other extra activities on the annotated topics. Peculiarly, the students considered on the same degree of agreement (without significant difference), that these activities would be helpful for both types of annotations, confusing or important.

Although the first goal of highlight tool is to provide students an individual method to support active-reading and refinding information, collaboration and communication also plays a major role in the learning process. Over 63% of the participants strongly agreed or agreed that collaborative features, as for example, sharing annotations and accessing other students' annotations, would definitely be beneficial during the learning process. By sharing annotations, or merely visualizing colleagues' highlights, students can have a better overview on the importance of some portions of the learning objects, and also on the portions raised more questions among her learning group. Shared annotations improve the individual learning and boost the online group discussion as well.

Table 2.4 Results of the user experience questionnaire.

	Favor	Neutral	Against
Usage	54.23%	23.81%	21.96%
Utility/Satisfaction	66.14%	20.11%	13.76%
Application	51.72%	25.79%	22.49%
Collaboration	63.49%	23.81%	12.70%
Future Use	53.57%	25.40%	21.03%

Finally, over 53% of the participants would recommend the tool for colleagues and are also willing to use the tool in the next courses.

2.7.5 Conclusions

The tool was deployed in an e-learning course with over 750 students that actively used it. Through the use of the tool, the tutors could create new discussion topics to handle some students questions or to extend topics that was marked as interesting.

The annotations also contributed to improve the content available to the students. The contents' author reviewed the passages of the text that were very often marked as confusing or important. The texts that were marked as confusing are being reformulated. The texts that were marked as important are being expanded and in the next version of the course a complementary material will be available for the students.

Finally, in the point of view of the course coordination or even of the institution, the tool is important to give feedback about the student needs, content quality and the continuity of the course. Through the use of this tool, the teachers can go beyond the group needs but also address individual needs of each student. Further, through the feedback collected in our user evaluation, we conclude that the annotations had a positive impact in the user satisfaction.

2.8 Chapter Summary

In this chapter, we presented an in deep study of Web annotation, its characteristics, advantages, drawbacks and some design guidelines to make the best out of it. The studies performed with proposed annotation tool, SpreadCrumbs, emphasize the importance of contextual information during the annotation process.

Here, context is understood as a spatial attribute within a Web resource and a personal reflection from the annotator from the time the annotation was performed.

We proved that contextualized Web annotations are significantly better to support users in the task of refinding information. Additionally, in the scope of online education, we demonstrated the applicability of a contextualized Web annotation tool that regards all lessons learned in our preceding research. The annotation tool was

introduced to a real education scenario and has shown great value to educators and students.

3 Contextualized Profiles

In this chapter, we will present studies done in order to achieve the assembling of contextualized profiles. First, the study presents the development of a social tagging system that supports additional tagging facets. On top of that, we propose a contextualized folksonomy model and strategies to build contextualized resource profiles. Finally, we evaluate the proposed method and strategies in different systems, validating their benefits for information retrieval and recommendation tasks.

3.1 Introduction

Tagging systems like Flickr or Delicious enable people to organize and search large item collections by utilizing the Web 2.0 mechanisms: Users attach tags to resources and thereby create so-called tag assignments which are valuable metadata. However, imprecise or ambiguous tag assignments can decrease the performance of tagging systems regarding search and retrieval of relevant items.

For example a tag assignment, alloted to an image may only describe a small part of an image and hence cannot be used to derive the overall topic of the image correctly. Some tag assignments are valid for a user-specific point of view, e.g., a tourist would tag an image of a landmark in a different way than a geologist. And finally tag assignments suffer from ambiguity in natural languages.

For disambiguation, approaches like MOAT [PL08] exist, which support users to attach URIs describing the meaning of a tag to a particular tag assignment analogously to semantic tagging in Faviki[1]. A more sophisticated approach, which exploits Wikipedia and WordNet[2] to detect the meaning of tags, is presented in [MTR+07].

In this chapter, we extend the common folksonomy model by flexible, contextual tagging facets. We present the TagMe! system that introduces novel tagging facets:

[1] http://faviki.com
[2] http://wordnet.princeton.edu/

Tag assignments are enriched with a DBpedia URI [ABK+07] to disambiguate the meaning of a tag. So-called *area tags* enable users to tag a specific part of an image (spatial tagging). Furthermore, a *category* dimension is offered to categorize tag assignments.

Finaly, we formalize the notion of tag-based and context-based resource profiles and introduce a generic strategy for building such profiles by incorporating available context information from all parts involved in a tag assignment. Our method takes into account not only the contextual information attached to the tag, the user and the resource, but also the metadata attached to the tag assignment itself. We demonstrate and evaluate our approach on two different social tagging systems and analyze the impact of several context-based resource modeling strategies within the scope of tag recommendations. The outcomes of our study suggest a significant improvement over other methods typically employed for this task.

In this light, the main research questions we address in this chapter are:

- Can context be exploited in tagging scenarios to improve retrieval of relevant items?
- How to include contextualized information in the profile of resources in a folksonomy?

In the reminder of this chapter we answer these questions and provide the following contributions:

- We extend the traditional folksonomy of tagging systems to a faceted folksonomy model.
- We build an in-context tagging system, TagMe!.
- We propose and evaluate strategies that exploit the contextualized folksonomies in order to improve profiling and retrieval of resources.

3.2 TagMe!

The studies presented in this chapter are associated with the social tagging system TagMe!, developed specifically to validate our ideas and assumptions, thus, an introduction to the system is appropriate. TagMe! [AKKS09] is an online image tagging system where users can assign tags to pictures available in Flickr. Figure 3.1 outlines the conceptual architecture of TagMe!, which can basically be considered as an advanced tagging and search interface on top of Flickr. Users can directly import pictures from their own Flickr account or utilize the search interface to retrieve Flickr pictures. If users tag their own images in TagMe! then the tags are propagated

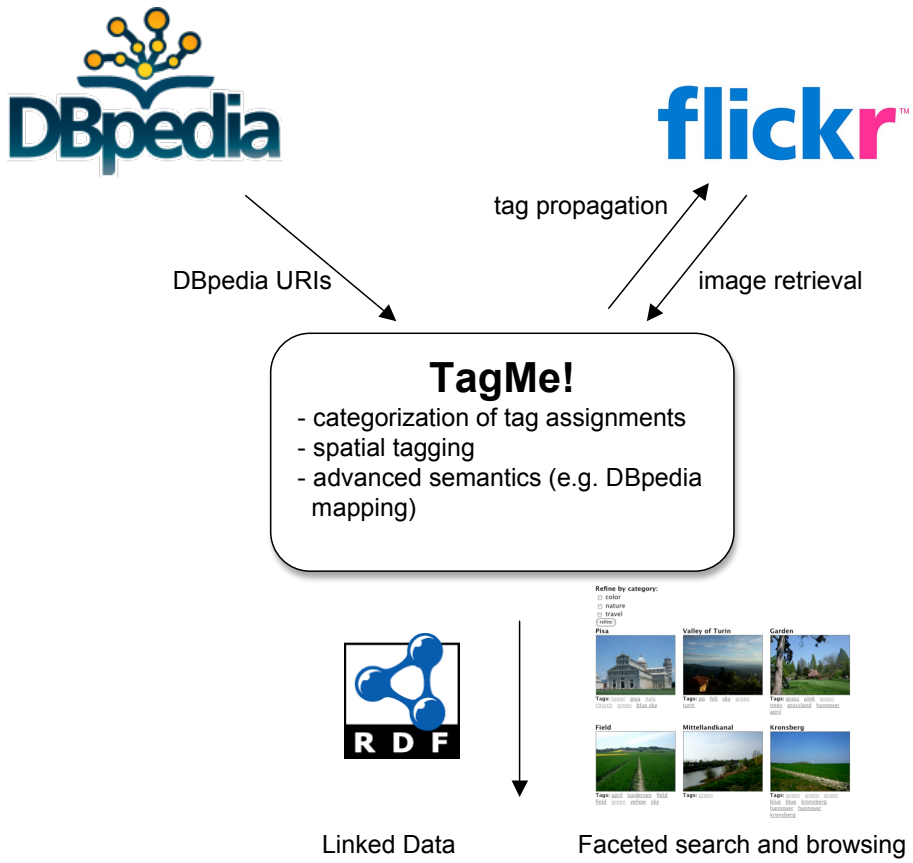


Figure 3.1 Conceptual architecture of TagMe!

to Flickr as well. Moreover, TagMe! maps DBpedia URIs to tag and category assignments by exploiting the DBpedia lookup service[3]. Hence, all tags and categories have well-defined semantics so that applications, which operate on TagMe! data, can clearly understand the meaning of the tag and category assignments. The (meta)data created in TagMe! is made available according to the principles of Linked Data [BL07] using the MOAT ontology[4] and Tag ontology[5] as primary schemata.

TagMe! extends the Flickr tagging functionality in two further facets, specifically *categories* and *area tags*. For each tag assignment the user can enter one or more categories that classify the annotation. While typing in a category, the users get auto-completion suggestions from the pre-existing categories of the user community (see bottom in Figure 3.2). TagMe! users can immediately benefit from the categories as TagMe! provides a faceted search interface that allows to refine tag-based search activities by category (and vice versa). Additionally, users are enabled to perform *spatial tag assignments*, i.e. use tags to annotate a specific areas of an image, which

[3]http://lookup.dbpedia.org
[4]http://moat-project.org/ns
[5]http://www.holygoat.co.uk/projects/tags

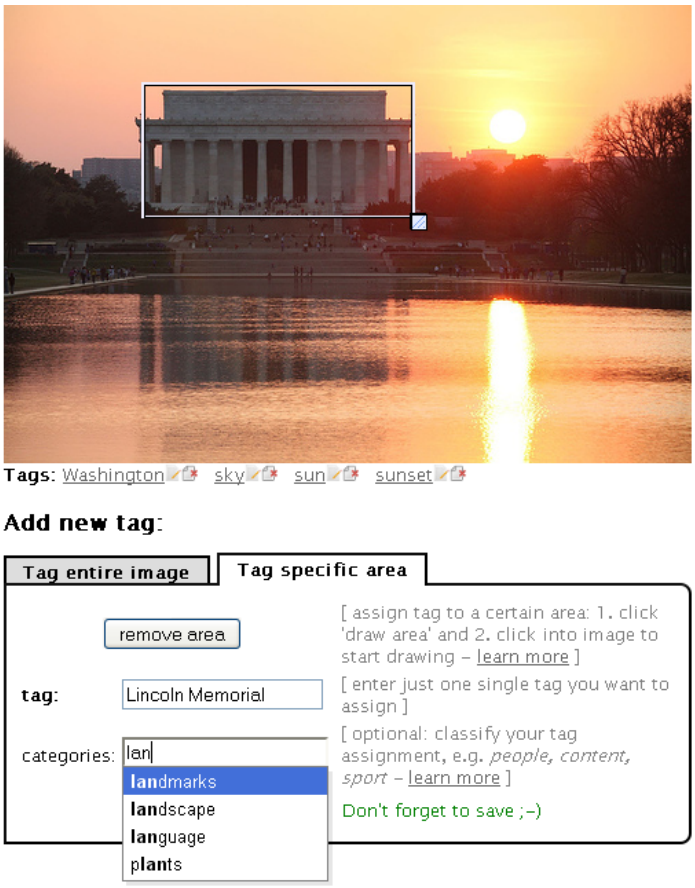


Figure 3.2 User tags an area within an image and categorizes the tag assignment with support of the TagMe! system.

they can draw within the picture (see rectangle within the photo in Figure 3.2) similarly to *notes* in Flickr or annotations in LabelMe [RTMF08]. When tagging, people usually only tag the main content of the picture, giving less or almost no importance to supplementary scenery images.

Area tags motivate the users to do so adding significant semantic value to each annotated image. Moreover, each spatial tag assignment has a globally unique URI and is therewith linkable, which allows users to share the link with others so that they can point their friends and other users directly to a specific part of an image. For example, if users follow the link of the spatial tag assignment "opera"[6], shown in Figure 3.2 then they are directed to a page where the corresponding area is highlighted, which might be especially useful in situation where users discuss about specific things within a picture. While the area tags add an enjoyable visible feature for highlighting specific areas of an image and sharing the link to such areas with friends, we consider them as highly valuable to improve search by detecting tag correlations or to enhance

[6] http://tagme.groupme.org/TagMe/resource/403/tas/1439

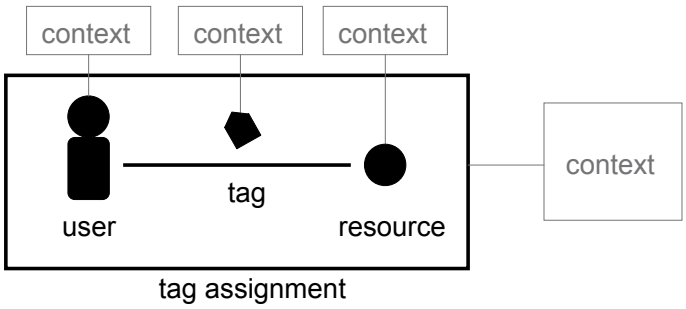


Figure 3.3 Contextual information of social annotations can refer to the *user* that performed the tag assignment, to the *tag* that was designated by the user, to the *resource* that was annotated, or to the entire *tag assignment* itself.

the identification of similar tags.

3.2.1 Faceted Tagging

To express the introduced enhancements of the TagMe! tagging system in a formal way, current folksonomy models need to be extended. Formal folksonomy models are e.g. presented in [HRS07, Mik05]. They are based on bindings between users, tags, and resources. According to [HJSS06a] a folksonomy is defined as follows:

Definition 2 (Folksonomy) *A* folksonomy *is a quadruple* $\mathbb{F} := (U, T, R, Y)$*, where* U*,* T*,* R *are finite sets of instances of* users*,* tags*, and* resources*, respectively.* Y *defines the* tag assignment*, which is a relation between these sets (i.e.,* $Y \subseteq U \times T \times R$*) that is potentially enriched with a timestamp indicating* when *it was performed.*

The above definition abstracts from the tagging activities and does not incorporate contextual information. The latter refers either to the entities involved in a tag assignment (i.e., the user, the tag, and the resource), or to the tag assignment itself. This is clearly illustrated in Figure 3.3.

To allow users to create these different facets of a tag assignment, we extend the given folksonomy in order to consider all possible dimensions of contextual information: the meta-data attached to the tags, to the resources and to the users, as well as the usage context attached to tag assignments, as a whole. To cover the last case, we need to accommodate the attachment of any kind of context to a tag assignment. We employ an extension of Definition 2, namely the *context folksonomy model* [AHKK10].

Definition 3 (Context Folksonomy) *A* context folksonomy *is a tuple* $\mathbb{F} := (U, T, R, Y, C, Z)$*, where:*

- U*,* T*,* R*,* C *are finite sets of instances of* users*,* tags*,* resources*, and* context information*, respectively,*

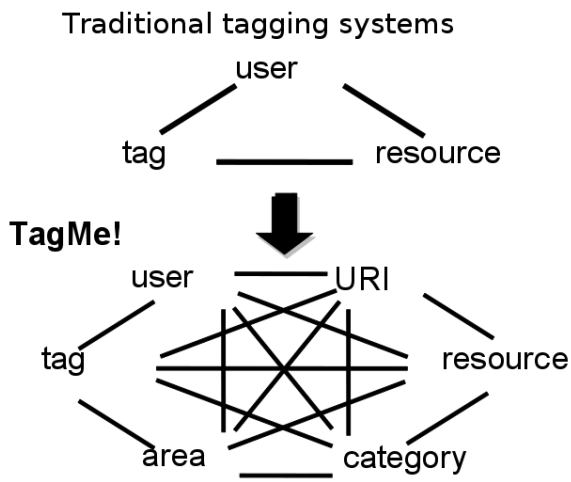


Figure 3.4 The Faceted Folksonomy in the TagMe! system

- *Y defines the* tag assignment*, which is a relation between U, T, and R (i.e., $Y \subseteq U \times T \times R$), and*
- *Z defines the* context assignment*, which is a relation between Y and C (i.e., $Z \subseteq Y \times C$).*

In the TagMe! system, the *context information* can be a) an area, b) a DBpedia URI or c) a category. All context information are assigned to a tag assignment by a relation Z.

By utilizing the additional information, tag assignments become more connected to each other (see Figure 3.4). For example, two tags assigned to the same area within an image or having the same DBpedia concept can be considered as synonyms, while two tags that are assigned to different areas in an image are possibly not that strongly related to each other.

3.3 Related Work

The analyses in the previous section revealed several technical advantages of the tagging facets available in the TagMe! system. In this section we compare the tagging and tag-based exploration features of TagMe! from the perspective of the end-users with other tagging systems: Flickr, Delicious, Faviki [Mil08] and LabelMe [RTMF08]. Our comparison among the systems is partially based on the dimensions of the *tagging system design taxonomy* proposed by Marlow et al. [MNBD06a]. For example, we compare the (i) "Tagging rights", (ii) "Tagging support" and (iii) "Aggregation model" of those systems. These characteristics define respectively (i) who can tag, (ii) if the user gets assistance from the system during the tagging process and (iii)

whether the system allows users to assign the same tag more than once to a particular resource (aggregation model = bag) or not (aggregation model = set).

In TagMe!, we extend the tagging design taxonomy with the following additional dimensions related to tagging.

Semantic tagging We consider tagging as semantic tagging whenever the meaning of a tag is clearly defined, for example, by attaching a URI explaining the meaning of the tag [PL08].

Spatial tagging The practice of annotating a specific piece of a resource, e.g., parts of an image or paragraphs in a text.

Tag categorization A method enabling users to categorize or classify the tags and tag assignments.

Further, we introduce two dimensions that characterize to which degree users can exploit the tags to retrieve resources within the system.

Tag-based navigation Not all systems that provide tagging functionality also allow their users to explore and browse content based on tags, e.g. initiating search by clicking on a tag.

Faceted navigation By faceted navigation, we mean the feature of filtering resources based on the different dimensions of a tag assignment, i.e. by user, tag, or resource, category, or area (cf. Folksonomy model, Section 3.2.1). For example, in Delicious people can navigate through bookmarks annotated with specific tags (tag dimension) by a specific user (user dimension).

TagMe! provides two tagging features that are currently not sufficiently implemented in other systems: spatial tagging and tag categorization. Flickr and also MediaWiki[7] platforms enable users to add notes or comments to specific areas within pictures. However, similarly to LabelMe, which allows users to attach keywords to arbitrarily formed shapes within an image, these systems do not provide means for tag-based navigation based on such spatial annotations, i.e. users cannot click on a spatial tag assignment to navigate to other resources that are related to the corresponding tag (and possibly to the area). TagMe! offers tag-based navigation, which is common in tagging systems such as Flickr and Delicious, also for spatial tag assignments. A further innovation of TagMe! is the tag categorization that is performed on the level of tag assignments (*tas categorization*) and can therewith be used to disambiguate the meaning of a particular tag assignment. Delicious, on the contrary, only supports grouping of tags in so-called *tag bundles*. These tag bundles enable users to organize tags but do not help them to disambiguate specific tag assignments. They

[7] http://www.mediawiki.org

are moreover seldomly used: Tonkin reports that approx. 10% of Delicious users have more than five tag bundles [Ton06].

The structure of folksonomies can be exploited to navigate through the resource corpus of a tagging system with respect to different facets. For example, when clicking on a tag in Flickr to explore related pictures, users can filter the results to narrow down the results to pictures of a specific *user* or pictures that occur in a specific *group* of pictures. In addition to the feature of browsing resources in context of specific users (as possible in Flickr, Delicious, and Faviki), TagMe! allows such tag-based faceted navigation by applying the categories of tag assignments as filters.

3.4 Introducing Resource Profiles

Tag assignments are typically marked with subjectivity: different authors can interpret the same tag in different ways. Although this conveys significant benefits in the case of personalized tags, it poses a significant obstacle to the usefulness of the collective ones: the purpose of a tag assignment is not always clear to users other than its creator. For example, a tag associated with an image may describe it with respect to different aspects: the place and the persons depicted, the owner, an opinion or even its usage context (i.e., associated task). Thus, tags can be valid solely from a user-specific point of view [GH06]. This also explains why not all tags are suitable for search [BFNP08a]; even those tags that mainly aim at describing the content of an item might characterize just a small part of the resource, without being representative of the entire resource. Some systems like LabelMe [RTMF08] and, our proposed system, TagMe![8] [AHKK10] offer solutions to this problem by providing tags of finer granularity to their users.

In addition, tag assignments suffer from the ambiguity, inherent in any natural language: multiple meanings can be associated with the same tag (*polysemy*), while a specific tag can have multiple interpretations (*synonymy*). As previously explained, in TagMe! users can enrich tag assignments with additional facets: semantic categories, URIs and spatial information. These facets represent contextual information that contribute to the disambiguation of the tag assignments, thus facilitating the search and the recommendation of tags or resources to a great extent.

We argue that both the aforementioned shortcomings of social annotations can be ameliorated by considering their context. Abel et al. have already demonstrated the benefits of context for recommendation strategies [AHK08]. However, the methods presented there were tailored to a particular system and, thus, were not generalizable to other social tagging systems. Instead, here we introduce a general, versatile modeling approach for building comprehensive resource profiles that can be easily adapted to any folksonomy. It exploits the contextual information that is available in tagging systems rich in metadata, which are usually neglected.

[8] See `http://tagme.groupme.org`.

At the core of our approach lies the idea of encapsulating not only the information that exclusively pertain to tags, but also additional contextual facets that refer to the other components of a tag assignment: the user, the resource and the tag assignment itself. Merging these facets appropriately, we can derive weighted tag lists that form comprehensive contextual profiles, which are compatible and easily combined with typical tag-based profiles. These profiles can be employed in a diversity of common tasks that rely on tags, such as personalization, search and tag recommendation. We further describe how context-based profiles can be transformed into semantic URI-based profiles. We also put our generic resource modeling approaches into practice, demonstrating its applicability in two different social tagging systems: TagMe! and BibSonomy[9]. In both cases, we evaluate the impact of context-based profiles on the task of tag recommendations. The outcomes of our experimental study verify our premise that contextual profiles convey significant improvements in the performance of a social tagging system.

On the whole, the main contributions can be summarized as follows:

- we introduce the notion of tag-based and context-based resource profiles and present a generic context model for social tagging systems,
- we propose a generic strategy for exploiting context information embodied in social annotations, exemplifying it with a variety of resource modeling strategies, and
- we evaluate our strategies in two different tagging systems, verifying that the incorporation of contextual information clearly outperforms typical methods for generating resource profiles.

3.4.1 Tag-based Profiles

At the core of this work lies the notion of folksonomy structures from the perspective of resources. Similar to a *personomy* (i.e., the user-specific part of a folksonomy, coined by Hotho et al. in [HJSS06b]), we formally define the resource-specific fraction of a context folksonomy - called *personomy of a resource* from now on - as follows:

Definition 4 (Resource Personomy) *The personomy $\mathbb{P}_r = (U_r, T_r, Y_r, C_r, Z_r)$ of a given resource $r \in R$ is the restriction of $\mathbb{F}$ to r, where U_r and T_r are the finite sets of* users *and* tags*, respectively, which are referenced by the tag assignments Y_r that are attached to r. C_r comprises the contextual information that are associated with the tag assignments in Y_r, and Z_r are the corresponding context assignments.*

In essence, a resource personomy encompasses the tag assignments that refer to a specific item along with their context. In a more abstract level, the *tag-based resource profile* $P(r)$ represents a resource as a set of weighted tags.

[9] See `http://www.bibsonomy.org`.

Definition 5 (Tag-based Resource Profile) *The* tag-based profile $P(r)$ *of a resource* $r \in R$ *is a set of weighted tags, where the weight of a tag* t *is computed by a certain strategy* w *with respect to the given resource* r*:*

$$P(r) = \{(t, w(r,t)) | t \in T, r \in R\},$$

where $w(r,t)$ *is the weight that is associated with tag* t *for a given resource* r*.*

$P(r)@k$ denotes the subset of a tag-based profile $P(r)$ that contains the k tag-weight pairs with the highest weights. $\bar{P}(r)$ represents a tag-based profile whose weights are normalized such that their sum is equal to 1, while $|P(r)|$ expresses the number of distinct tags contained in $P(r)$. It is worth clarifying at this point that the tags contained in $P(r)$ are not restricted to the tags that are explicitly associated with r (i.e., the tags included in the resource's personomy $\mathbb{P}_r$). Instead, $P(r)$ may also specify the weight for a tag t_i that is associated to the resource r indirectly, through another element of its context.

In line with Definition 5, tag-based profiles can be built for a given user $u \in U$ and for a particular context $c \in C$, as well. For instance, tag-based user profiles (i.e., $P(u)$) have been studied by Firan et al. [FNP07] and Michlmayr and Cayzer [MC07].

A straightforward approach to create a **tag-based context profile** $P(c)$ is to consider the tag assignments that pertain to c and to weight each of them according to the number of annotations that are contextualized with c and mention it. More formally: $w(c,t) = |\{(u,t,r) \in Y : (c,(u,t,r) \in Z)\}|$ (cf. Definition 3). In Section 3.4.3, we introduce more advanced strategies that exploit the characteristics of the respective type of context and show how these context profiles can be employed to enhance tag-based resource profiles.

3.4.2 Baseline Strategies for Tag-based Resource Profiles

The main challenge in generating tag-based profiles for resources is the definition of a strategy w that appropriately assigns weights to the involved tags. In the following, we present two weighting approaches that are typically used in the literature, but do not exploit all aspects of the context of tag assignments.

Tag Frequency. The rationale behind this approach is the assumption that the more users annotate a resource r with a tag t, the more salient is t for the description of r. Given the personomy of a resource $\mathbb{P}_r$, the corresponding tag-based resource profile $P(r)$ can be formed by counting the number of distinct users that assigned at least one tag $t \in T_r$ to the resource r. Hence, the weight $w(r,t)$ attached to a specific tag t in $P(r)$ is equal to: $w(r,t) = |\{u \in U_r : (u,t,r) \in Y_r\}|$. This approach was essentially employed by Cai and Li in [CL10] with the aim of improving tag-based personalized search.

Tag-based Co-Occurrence. In tagging systems like Flickr, resources are typically annotated with a limited number of distinct tags [SvZ08]. For this reason, Sigurbjörnsson and Zwol suggested in [SvZ08] to enrich the profile of a resource r with those tags that frequently co-occur with the tags assigned to r (i.e., T_r). The weight of those additional tags is equal to the frequency of their co-occurrence in the folksonomy:

$$w(r,t) = |\{(u, t_i, r_j) \in Y : \exists t_i \in T_r \wedge t \in T_{r_j}\}|.$$

The second method is typically employed in the context of tag recommendation techniques, which rely on association rules to capture the co-occurrence patterns (see, for instance, a recent, state-of-the-art method, introduced by Heymann et al. in [HRGM08]). For this reason, we employ it as the baseline method in our experimental study that examines the applicability of our algorithms in the tag recommendation task.

3.4.3 Context-based Strategies for Tag-based Resource Profiles

As their name suggests, context-based strategies rely on the contextual information available in folksonomies, and in resource personomies in particular: they build the profile of a resource r by merging (some of) the tag-based context profiles $P(c)$ associated with r. Moreover, one can also consider contextual information attached to the tag assignments which refer to r (cf. Figure 3.3). The process of generating context-based resource profiles is outlined in the form of a generic approach in Definition 6.

Definition 6 (Context-based Resource Profile) *Given a tag-based profile $P(r)$ of a resource r and the set of tag-based context profiles $P(c_1)$,..., $P(c_n)$, where $c_1, .., c_n \in C_r$ form the context information available in the resource personomy $\mathbb{P}_r$, the* context-based resource profile *$P_c(r)$ is computed by aggregating the tag-weight pairs (t_j, w_j) of the given profiles according to the following algorithm. Note that the parameter α_i allows for (de-)emphasizing the weights originating from profile $P(c_i)$.*

Input:
$P(r), ContextProfiles = \{(P(c_1), \alpha_1), ..., (P(c_n), \alpha_n)\}$
Initialize: $P_c(r) = P(r)$
for $(P(c_i), \alpha_i) \in ContextProfiles$:
  $P(c_i) = \bar{P}(c_i)$
  for $(t_j, w_j) \in P(c_i)$:
    if $(t_j, w_{P_c(r)}) \in P_c(r)$:
      replace $(t_j, w_{P_c(r)})$ *in* $P_c(r)$ *with* $(t_j, w_{P_c(r)} + \alpha_i \cdot w_j)$
    else:

```
      add (t_j, α_i · w_j) to P_c(r)
    end
  end
end
Output: P̄_c(r)
```

The above algorithm is independent from the type of context information that is exploited to construct the context-based profiles and is, thus, generalizable to any tagging system. The construction of context-based resource profiles $P_c(r)$ depends, however, on the type of context that is considered. In the following, we present several weighting strategies for building them in systems rich in metadata, like TagMe! and BibSonomy.

First, we describe the strategies used to build contexts for resources in TagMe! This system offers spatial tag assignments, enabling users to draw a rectangular area that specifies the part of the image that is relevant to the corresponding tag. These rectangular areas carry implicit information, which add more value to a tag assignment. Consider, for instance, the size and the distance of the tag's area from the center of the resource; the former represents the portion of the visual space that is covered by the tag, with larger areas denoting tags that are more representative of the whole resource (i.e., tags with small area pertain to a particular object depicted in the picture, whereas large areas correspond to tags describing the picture in its entirety) [AHKK10]. Similarly, the latter expresses the relevance of tag assignments to the resource, with tags closer to its center being more important than tags placed at the margin of a resource [AHKK10].

In addition to this spatial facet, TagMe! provides another two dimensions that are suitable for building context-based resource profiles: the categories and the semantic-meaning of tags. Categories can be freely entered by users via the tagging interface, in order to provide a more general description that disambiguates and describes tags more clearly. For instance, the tag "Deutscher Bundestag" can be assigned to the category "Building". In addition, TagMe! automatically enriches tags and categories assignments with DBpedia URIs to further disambiguate the meaning of a tag. In the following, we introduce strategies for building context-based profiles with the help of the tagging facets of TagMe!.

User-based Co-Occurrence. The rationale behind this weighting method is that an individual typically annotates similar resources, thus employing relevant tags in her tag assignments. This strategy considers, therefore, all users that assigned a tag to a given resource r and aggregates all the tags that they used (even for annotating other resources) into the context-based resource profile $P(r)$. The weight $w(r,t)$ is calculated by accumulating the weights $w(u,t)$ of the tags available in the tag-based profiles of these users:
$w(r,t) = \sum_{u \in U_r} |\{r_k \in R : (u,t,r_k) \in Y, r_k \neq r\}|$.

Semantic Category Frequency. This strategy considers as evidence for the significance of a tag, the popularity of the category(ies) associated with the respective tag assignment(s). The premise here is that a tag associated with a category is more important and, thus, more relevant to the annotated resource than a tag without a category. In fact, the more frequent its category, the more relevant it is. The weight of each tag is, therefore, equal to the frequency of its category. In case a tag is associated with multiple categories, its weight amounts to the sum of the respective frequencies: $w(r,t) = \sum_i |\{(c_i,(u_j,t_k,r_l)) \in Z : \exists(c_i,(u,t,r)) \in Z_r\}|$.

Co-occurring Semantic Category Frequency. The motivation for this strategy is the idea that tags described by the same categories are semantically relevant with each other. Consequently, when one of them is assigned to a particular resource r, the rest are also representative of r. Thus, given a resource r, this weighting method retrieves all categories associated with r and places all tags associated with them (even through another resource) in the profile of r, $P(r)$. In line with the previous strategy, the value of each tag is set equal to the (sum of) frequency(ies) of the related category(ies): $w(r,t) = \sum_i |\{(c_i,(u_i,t_j,r_k)) \in Z : c_i \in C_r \wedge \exists(c_i,(u,t,r)) \in Z_r\}|$.

Semantic Meaning. The rationale behind this approach is the assumption that semantically annotated tags constitute the more carefully selected annotations of a resource, thus being more representative of it and the basis for a more comprehensive description. This strategy defines, therefore, two levels of importance for tags, depending on whether they have been linked to a URI that uniquely identifies their meaning. In other words, it assigns a binary value to each tag, with those tags that satisfy this condition receiving the value of 1, while the rest take the value of 0. More formally:
$w(r,t) = 1$ if $\exists(URI,(u,r,t)) \in Z_r$.

Co-occurring Semantic Meaning. At the core of this strategy lies the idea that tags that are semantically equal to, but more popular than the tags directly associated with r, are more representative of its content. Thus, given a resource r, this strategy aggregates all the URIs involved in the tag assignments of r and builds the resource profile $P(r)$ by aggregating all tags that were associated with these URIs, independently of the respective resource. Tags are weighted according to the frequency(ies) of the URI(s) assigned to them: $w(r,t) = \sum_{URI_i \in C_r} |\{(URI_i,(u_j,r_k,t_l)) \in Z : \exists(URI_i,(u,r,t)) \in Z_r\}|$.

Area Size. The intuition behind this method is that the importance of tags is proportional to their size: the larger the area occupied by a tag, the more relevant the tag is to the annotated resource. On the other hand, tags that have been associated with a particular part of a resource, are considered more specific, and thus less significant. Thus, this strategy assigns to each tag a weight proportional

to its area. More formally: $w(r,t) = |x_1 - x_2| \cdot |y_1 - y_2|$, where (x_1, y_1) and (x_2, y_2) are the Cartesian coordinates of the lower left and the upper right edge of the tag's rectangle ($x_1, x_2, y_1, y_2 \in [0,1]$).

Distance From Center. This strategy is based on the assumption that the closer a tag is to the center of a resource (e.g., image), the more relevant it is. Hence, it weights tags according to their distance from the resource's central point, with smaller distances corresponding to higher values.

Expressed mathematically, we have: $w(r,t) = \frac{1}{\sqrt{(x_{t_c} - x_{r_c})^2 + (y_{t_c} - y_{r_c})^2}}$, where (x_{r_c}, y_{r_c}) and (x_{t_c}, y_{t_c}) are the coordinates of the center of the resource and the center of the tag, respectively ($x_{r_c}, x_{t_c}, y_{r_c}, y_{t_c} \in [0,1]$). Note that, with respect to the annotations of the above strategy, we have $x_{t_c} = \frac{x_1 + x_2}{2}$ and $y_{t_c} = \frac{y_1 + y_2}{2}$.

It should be stressed at this point that the above strategies rely on different facets of the context folksonomy of TagMe!. Thus, instead of being competitive to each other, they are complementary and can be arbitrarily combined. In total, we can have $(2^7 - 1 =)127$ distinct strategies, either *atomic* (i.e., just an individual weighting method) or *composite* ones (i.e., derived from the combination of multiple weighting techniques).

To demonstrate the adaptability and generality of our approach to building context-based resource profiles, we also propose concrete context modeling strategies for the folksonomy of BibSonomy.

Co-occurring Journal Frequency. BibSonomy resources (i.e., publications) are typically associated with the journals or conferences, where they were published. This strategy exploits these metadata information, assuming that each specific journal is focused on a particular subject that represents the aggregation of similar resources. Thus, its publications are highly relevant to each other, and the tags assigned to one of them are probably applicable to the rest, as well. Given a resource r, this weighting method retrieves the *Journal* metadata associated with r and aggregates in $P(r)$ the tags of all the resources that were published by the same journal. The value of each tag is equal to its frequency: $w(r,t) = |\{(c_j, (u_j, t, r_l)) \in Z : \exists (c_j, (u, t, r)) \in Z_r\}|$, where c_j stands for the journal metadata of the given resource r.

Co-occurring Journal-Year Frequency. The rationale behind this strategy is the assumption that the topics of the papers published in a specific journal drift with the passage of time. As a result, the papers published in the same journal in a particular year are more relevant in with each other than with the papers published at a different point in time. In this context, this weighting method retrieves for every resource r the *Journal* and *Year* metadata associated with it; then, it generates a list of the tags of all resources that were also published within

the same journal in the same year. Tag weights are set equal to the frequency of the tags:
$w(r,t) = |\{(c_{j,y},(u_j,t,r_l)) \in Z : \exists(c_{j,y},(u,t,r)) \in Z_r\}|$, where $c_{j,y}$ stands for the journal and year metadata of the given resource r.

3.4.4 Transforming Tag-based Profiles into Semantic Profiles

The aforementioned context-based modeling strategies can form the basis for the creation of **semantic profiles**; that is, profiles that explicitly specify the semantics of a tag by means of URIs. For social tagging systems that assign meaningful URIs to tag assignments (e.g., TagMe!) or systems that make use of the MOAT framework [PL08] (e.g., LODr [Pas08]), we propose the transformation of tag-based profiles into semantic profiles that, instead of a list of tags, consist of a weighted list of URIs.

It is worth noting at this point that the semantic meaning of tags depends on the context of their use. For example, the tag "paris" most likely refers to the city, but for some tag assignments it could also refer to a person. It is not possible, therefore, to have a global mapping of tags to URIs; instead, it is necessary to map each individual tag assignment to a particular URI. Thus, we propose to transform the personomy $\mathbb{P}_r$ (see Definition 4) and its tag assignments as follows:

Definition 7 (URI-based Resource Personomy) *Given the tag-based personomy $\mathbb{P}_r = (U_r, T_r, Y_r, C_r, Z_r)$ of a specific resource r and its URI assignments $Z_{r,uri} \subseteq Y \times C_{uri} \subseteq Z_r$, where C_{uri} is the set of URIs, the URI-based resource personomy, $\mathbb{P}_{r,uri} = (U_r, T_{r,uri}, Y_{r,uri}, C_r, Z_r)$, can be constructed by iterating over the tag assignments and replacing the tags with URIs of the corresponding URI assignments according to the following algorithm:*

$T_{r,uri} = T_r \cup C_{uri}$
$Y_{r,uri} = \{\}$
for $(u,t,r) \in Y_r$*:*
  for $((u,t,r), uri) \in Z_{r,uri}$*:*
    $Y_{r,uri} = Y_{r,uri} \cup (u, uri, r)$
  end
end
$\mathbb{P}_{r,uri} = (U_r, T_{r,uri}, Y_{r,uri}, C_r, Z_r)$

Given the URI-based Resource Personomy and a URI-based Context Folksonomy (which can be constructed in a similar manner as the semantic personomy), we can apply the resource modeling strategies presented in Sections 3.4.2 and 3.4.3 in order to generate semantic resource profiles. In this way, the resource modeling framework presented above supports tag-based tasks in both the social tagging and Semantic Web systems.

Table 3.1 Technical characteristics of the TagMe! data set.

Tag Assignments (TAs)	1,288
TAs with Spatial Information	671
TAs with Category Information	917
TAs with URI Information	1,050
TAs with all information	432

3.5 Experimental Setup

To measure the quality of the above, context-based resource modeling strategies, we apply them to the *tag recommendation* task: given a set of resources annotated with tags and metadata, the goal is to predict other tags that are also relevant to a specific resource, but have not yet been assigned to it. In the subsequent paragraphs, we describe the setup of the thorough, experimental evaluation we conducted in this context.

3.5.1 Social tagging data sets

In the course of our experiments, we employed two real-world data sets that stem from the aforementioned social tagging applications: TagMe! and BibSonomy. A detailed description of the technical characteristics of the data sets is presented below.

TagMe! This web application constitutes a multifaceted social tagging system that allows users to associate their annotations with a variety of (optional) metadata, suitable for building context-based resource profiles. The data we collected comprise the whole activity of the first three weeks after the launch of the system in June, 2009. In total, its user base comprises 30 users; half of them had a Flickr account and, thus, were able to tag their own pictures, while the rest assigned tags to random pictures and pictures of their own interest. A summary of the technical characteristics of this data set is presented in Table 3.1.

BibSonomy. BibSonomy [HJSS06a] is a social bookmarking and publication-sharing system that has been running for over four years. The resources in Bibsonomy are publications, stored in BibTeX format. Each resource has several additional metadata, such as the corresponding journal, volume, year, as well as the author names. We employed Bibsonomy's public data set that is available on-line from the 1st July 2010. It consists of 566,939 resources, described and annotated by 6,569 users. In total, there are 2,622,423 tag assignments and 189,664 unique tags. For our experimental study, we considered those resources that had the *journal* information and were tagged with at least five distinct tags. We randomly selected 500 of those resources and derived their context-based profiles from the whole data set.

3.5.2 Procedure: leave-one-out cross-validation

To evaluate the effect of context-based resource profiles on tag recommendations, we employed the leave-one-out cross-validation methodology in the following way: at each step, we hid one of the tag assignments and, then, we built the profile of the corresponding resource according to the selected strategy, based on the remaining assignments. The resulting profile encompasses a ranked list of tags, whose value is estimated according to the facets of the folksonomy that the current strategy considers. The goal is to predict the hidden tag by placing it in the top positions of the ranking. Hence, to estimate the performance of the algorithms, we considered the following *metrics*:

Success Rate at 1 (S@1), which denotes the percentage of tag predictions that had the missing tag at the first position of the ranking. It takes values in the interval $[0, 1]$, with higher values corresponding to higher performance.

Success Rate at 10 (S@10), which stands for the percentage of tag predictions that had the missing tag in one of the top 10 positions of the ranking. Similar to $S@1$, it takes values in the interval $[0, 1]$, and the higher the value, the better the performance of the corresponding method.

As *baseline* strategies, we consider the approaches that exclusively rely on the information encapsulated in tag assignments (i.e., user, tag, and resource) that are described in Section 3.4.2. The *tag frequency* strategy adds to a resource profile $P(r)$ only tags that have already been assigned to the resource. Consequently, it cannot be applied to the tag prediction problem without any further extension. Thus, we employ *tag-based co-occurrence* as the main baseline strategy and compare it to the context-based strategies of Section 3.4.3. These strategies enrich the traditional tag frequency with context-based profiles, with the process described in Definition 6.

3.6 Results

3.6.1 TagMe!

As mentioned above, the large number of facets of the TagMe! data leads to a total of 127 distinct context-based strategies. For the sake of readability and due to space limitations, we provide the results only for the atomic ones (see Definition 6) together with the best performing composite methods. It is worth noting at this point that our methods are employed as extensions to the baseline one, merging them with a weight α as described in Definition 6.

A summary of the performance of the baseline method and the atomic weighting strategies is presented in Table 3.2. It is evident that all context-based methods

Table 3.2 S@1 and S@10 results for the atomic context-based strategies combined with the baseline in TagMe! data set.

Context ID	Context	S@1	S@10	Context Weight
0	Baseline	0.076	0.331	-
1	User-based Co-Ocurrence	0.087	0.407	0.8
2	Spatial Annotation Size	0.089	0.408	0.4
3	Spatial Annotation Distance	0.094	0.377	0.9
4	Categorized Tag Frequency	0.085	0.352	0.5
5	Category-based Co-Occurrence	0.102	0.401	0.7
6	Semantic Description	0.078	0.407	0.8
7	Semantic-based Co-Ocurrence	0.083	0.406	0.1

improve on the baseline, to a varying, but statistically significant extent ($p < 0.01$). Semantic Description brings about a minor increase in $S@1$ of 2.6%, whereas Spatial Annotation Distance and Category-based Co-Occurrence account for an improvement well above 30%. Equally significant is the improvement with respect to the S@10 metric that varies from 6.3% for Categorized Tag Frequency context up to 23.3% for Spatial Annotation Distance.

The fourth column of Table 3.2 contains the optimal value of the weight used to merge the corresponding individual strategy with the baseline method. This value was determined through an exhaustive search of all values in the interval [0,2] with a step of 0.1. The actual effect of this parameter is demonstrated in Figure 3.5, where

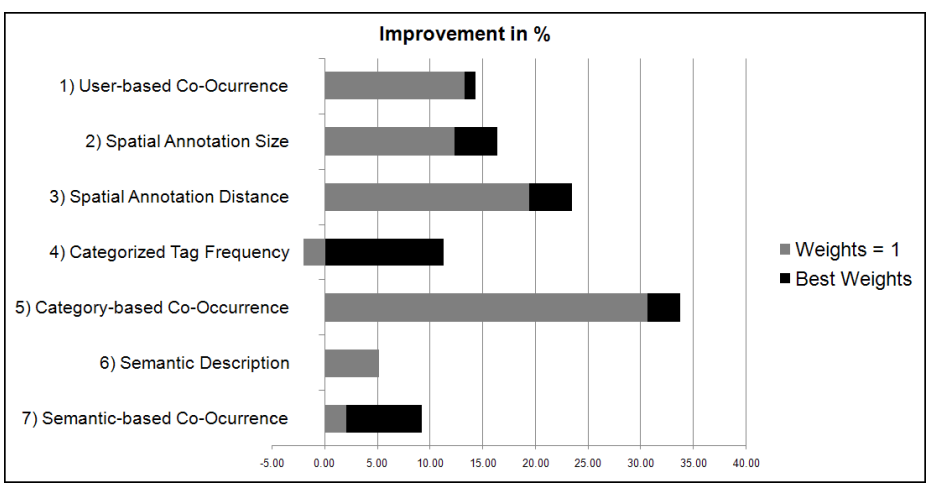


Figure 3.5 The improvement (in percentage) of each context over the baseline in the TagMe! data set. Gray bars show the results when the Context-Weight is set to 1, while black bars correspond to the performance of the best performing Context-Weight of each context.

Table 3.3 S@1 results for the composite context-based methods that have the optimal performance on the TagMe! data set. ContextIDs refer to the methods of Table 3.2.

ContextID (Context-Weight)	S@1	Improvement(%)
2(0.4) & 5(0.7) & 7(0.1)	0.106	38.8
2(0.4) & 3(0.1) & 5(0.7) & 7(0.1)	0.105	37.7
2(0.4) & 5(0.7)	0.105	37.7
5(0.7) & 7(0.1)	0.104	36.7

the performance for weight 1 (i.e., merging the baseline and the contextual strategy on an equal basis) is compared with the best performing weight. With the exception of Semantic Description, we can notice that the calibration of this parameter conveys significant improvement, ranging from 2% for User-based Co-Occurrence to 12% for Categorized Tag Frequency.

Additionally, we experimented with all possible composite strategies (i.e., combinations of the atomic ones), employing again a variety of context-weights for each of them (i.e., $w_i \in [0, 2]$ for each method i). The best performing ones are presented in Table 3.3, along with the respective weight and the improvement they convey with respect to $S@1$. We can see that all of them perform significantly better than the individual methods comprising them. Note, though, that they all involve the atomic strategy with the highest value for $S@1$ (i.e., Category-based Co-Occurrence) and assign to it the highest weight. However, they improve its performance by just 2%. This clearly means that merging different contexts does not result in a cumulative improvement, because their combination leads to noise in the form of contradictory evidence: a tag rated high by a specific weighting strategy is rated lower by another one.

On the whole, we can conclude that the contextualized strategies that rely on the spatial features, the categories and the semantics, produce the best results. They perform individually well enough and can be slightly improved when combined with the appropriate weights.

3.6.2 BibSonomy

The use case of BibSonomy demonstrates how our model of context-based resource profiles can be beneficially applied to any folksonomy, and how we can derive contextual information from the relations between the tag assignments. The outcomes of our evaluation are summarized in Table 3.4. We can see that both context-based methods substantially improve over the baseline, with the Co-occurring Journal-Year Frequency doubling its precision. Nevertheless, the overall success rate remains very low ($\sim$ 1%) in all cases. Note that the combination of the contextual weighting strategies with the baseline was done on an equal basis ($Context - Weight = 1$).

Table 3.4 S@1 and S@10 results for the baseline and the contextualized strategies (strategy-weight one to one) on the Bibsonomy data set.

Context	S@1	Improvement(%)
Baseline	0.00712	-
Co-occurring Journal Frequency	0.00991	39.02
Co-occurring Journal-Year Frequency	0.01425	100.00
Context	**S@10**	**Improvement(%)**
Baseline	0.0701	-
Co-occurring Journal Frequency	0.0770	10.42
Co-occurring Journal-Year Frequency	0.1045	49.13

3.7 Chapter Summary

In this chapter, we presented the research done in order to improve social annotation systems and folksonomies by exploiting additional contextualized information.

Here, context is understood as an additional piece of information (metadata) that can be grasped from tag assignments or from resources.

First, we developed TagMe!, a tagging and exploration front-end for Flickr, that allows for enriching tag assignments with spatial information, categories and DBpedia URIs. With this extended folsonomy system in hand, we introduced the notion of tag-based and context-based resource profiles and present a generic context model for social tagging systems. Additionally, we proposed a generic strategy for exploiting context information embodied in social annotations, exemplifying it with a variety of resource modeling strategies.

Additionally, we proposed novel approaches to exploiting the multiple types of context information available in social tagging systems in order to generate and enrich resource profiles. We demonstrated that context can be derived from almost any metadata of the components of a tag assignment (i.e., user, tag, and resource) as well as from the tag assignment as a whole. Furthermore, we formalized the approach for modeling context-based profiles and described various versatile strategies for combining them. To verify the benefits of context-based resource profiles, we considered the task of tag recommendation, which typically relies on naive resource profiles derived from tag co-occurrences.

We applied our strategies on two real-world datasets, with the outcomes indicating a considerable improvement over the baseline recommendation method. This verifies our premise that items sharing similar metadata (with respect to parts of their tag assignments) are highly likely to be described by similar tags. Additionally, contextual information pertaining to entire tag assignments provides significant evidence for modeling the resource profiles, too. This was proven to be particularly true for the cases where tag assignments are categorized, and spatially or semantically anno-

tated. Finally, we have validated that merging different contexts does not result in a culmulative gain, since the arbitrary combination of them may lead to muddled results. In the end, we verified that the incorporation of contextual information clearly outperforms typical methods for generating resource profiles.

4 Browsing Context

In this chapter, we present research done in order to improve Web navigation through the prediction and recommendation of webpage visits. We propose a contextual *revisit* prediction method that encompasses ranking and propagation methods. Apart from that, we also present the *Web History Repository* initiative, an effort that helped us gather client-side navigational information to validate our proposed methods and that we made publicly available. In addition to that, we present an effective sensemaking classification of users' behavior on the Web that helps to identify users' interest according to their browsing.

4.1 Introduction

The World Wide Web has become an important part of our lives. Search engines, travel planners, dictionaries and other online services have become essential for dealing with numerous task. News sites, portals, online games as well as streaming video are popular resources for information and entertainment. We communicate with our friends via email, social networking, forums, blogs and chat.

Many of these online activities are carried out on a hourly, daily, weekly or monthly basis. To facilitate them, we typically rely on known, trusted Websites that we have visited before. Web browsers support revisitation of pages and sites through mechanisms such as URL auto-completion, the forward and back buttons, bookmarks and the history sidebar. However, this support is found to be suboptimal and skewed toward a small set of frequently visited resources [OWHM07].

For this reason, analysis and prediction of online browsing behavior and revisitation patterns have received much attention, not only from the research community, but also from the industry [ATD08, TT10, KT10, CKT10, PKBGM10]. Academic research delivered several alternative history mechanisms, including gesture navigation [CM01], a SmartBack button that recognizes waypoints [MFJR+04], a browsable

SearchBar organized around a hierarchy of past queries [MMV08] and many types of history visualizations: lists, hierarchies, trees, graphs, 2d and 3d stacks, footprints (see [May09] for an overview). Browser add-ons that support users in revisiting pages and sites include Del.icio.us (social bookmarking), Infoaxe[1] and Hooeey (full-text history search), WebMynd[2] (history sidebar for search) and ThumbStrips (history visualization).

Here, we study how we can improve the state-of-art solutions for Web revisitation prediction task. In this chapter we propose, analyze and compare several revisitation contexts which aim at better supporting revisitation prediction. We achieve this by introducing SUPRA, a generic library for real-time, contextual prediction of navigational activity that encompasses a set of methods aligned in two tiers. The first tier ranks resources according to their likelihood of being used in the immediate future, as it is derived from their recency and/or frequency of use. The second tier, complements the ranking methods with propagation methods that identify resources that are commonly visited within the current user context.

The contextual prediction library is used as a basis for the PivotBar (See Section 4.5 for more details), a dynamic browser toolbar that recommends visited pages, relevant to the currently viewed page. The toolbar bears similarities to the concept of dynamic bookmarks [TW98, NS05, GMP07]. In contrast to them, however, the recommendations of the PivotBar are contextualized and reflect the dynamics of user behavior, as they are encapsulated by the ranking methods.

We evaluated the prediction performance of the generic surfing prediction library with two datasets: one consisting of a detailed client-side log of 25 users, gathered over a period of six months, and another, more extensive one that contains anonymized usage logs that were recently collected through the Web History Repository project. The results of the experiments indicate that taking the user context into account (i.e., combining ranking methods with propagation methods) drastically improves prediction performance. Moreover, the outcomes verify that predicting sites instead of individual pages is an easier task, thus exhibiting higher performance. The actual usage and appreciation of these recommendations has been evaluated in two user studies with the PivotBar browser add-on. The log data shows that a significant amount of revisits has taken place via the PivotBar.

In this light, the main research questions we address in this chapter are:

- To which extent user behavior can be predicted in Web revisitation activities?
- How to exploit browsing context to improve revisitation prediction?
- Can we make sense of browsing context?

[1] See `http://infoaxe.com`.
[2] See `http://www.webmynd.com`.

In the reminder of this chapter we answer these questions and provide the following contributions:

- We propose and evaluate several contextual prediction methods that exploit the user browsing context.
- We develop a tool to collect contributions of browsing user data and publicly provide the dataset for future research.
- We evaluate the proposed methods on real usage data to nominate the best performing ones.
- We make sense of users browsing behavior by clustering activities and uncovering related tags.

4.2 Related Work

In the first part of this section, we summarize the findings from several studies on how and why users revisit pages. In the second part, we discuss common approaches for predicting revisit patterns on the Web.

4.2.1 Studies on Web Usage and Revisitation

One of the first studies on Web usage behavior was carried out by Tauscher and Greenberg [TG97] in 1995. They recognized the fact that Web users often carried out recurrent tasks on the Web. Their empirical results confirmed Catledge and Pitkow's [CP95] finding that the links and the back button were the most frequently used methods for accessing a Web page; bookmarks and the temporally ordered history list were rarely used. They defined the *recurrence rate* to be the probability that any page visit is a repeat of a previous visit, expressing it as a percentage. An average recurrence rate of 58% was estimated for their participants; reanalysis of the data from the Catledge and Pitkow study yielded a recurrence rate or 61%.

The authors made some further characterizations of page revisits. It was found that the relation between the number of page requests and the number of unique pages visited thus far is roughly linear; the *URL vocabulary* grows linear with the number of page requests. Two important characteristics of revisited pages were described: first, most page revisits pertain to pages visited very *recently*; the probability for a page to be revisited decreases steeply with the number of page visits since the last visit. Second, there is a small number of highly *popular* pages that are visited very frequently; the probability for a page to be revisited decreases steeply with its popularity ranking.

Another long-term click-through study was carried out by Cockburn and McKenzie [CM01]. They observed that browsing is a *rapidly interactive* activity; the most frequently occurring time gap between subsequent page visits was around 1 second and gaps of more than 10 seconds are relatively rare. Analysis from the bookmark files revealed that most users have or will have problems with the size and the organization of their bookmark collections.

More recently, Weinreich et al. [WOHM06] carried out a long-term study in which they analyzed the interactions of 25 users with the Web browser during a period of four months and compared the results with the studies discussed above. They showed that the introduction of new browser features - such as tabbed browsing - and the change of the Web from a rather static hypermedia document repository to an interaction and transaction oriented platform, has a dramatic impact on the way users navigate the Web. Tabbed browsing has been established as a useful alternative for hub-and-spoke navigation that replaces backtracking to a significant extent.

Based on user action logs and interviews, Obendorf et al. [OWHM07] distinguished *short-term revisits* (backtrack or undo) from *medium-term* (re-utilize or observe) and *long-term revisits* (rediscover). For short-term revisits, the back button was found to be the most commonly used tool. For medium-term revisits, users normally type the page address directly into the address bar, making use of the URL completion. However, after a certain period the page is removed from the URL completion list. In these situations, if a user does not remember the exact address and if the address has not been bookmarked, she needs to rely on *waypoints*, from which a trail to the desired page can be followed. Further, the results showed that different categories of sites invite different revisit behavior: search engines and other portal sites typically have one page that users frequently return to, whereas institutional and project-related sites also comprise a long tail of pages visited several times.

Adar et al. [ATD08] further investigated revisitation behavior, making use of a large user base collected via the Windows Live Toolbar. They found out that short-term revisits involve hub-and-spoke navigation, visiting shopping or reference sites or pages on which information was monitored. Medium-term revisits involve popular home pages, Web mail, forums, educational pages and the browser homepages. Long-term revisits involve the use of search engines for revisitation, as well as weekend activities, such as going to the cinema. A subsequent study was carried out [TT10], based on a merged dataset of search engine logs, Web browser logs and a large-scale Web crawl, comprising several millions of users. The results confirmed earlier findings: within-session refinding mainly involves continuing work on a task or a routine behavior, whereas cross-session revisits mainly involves re-evaluation (e.g., "Did I remember the information correctly?", "Did something change?" or "Has something new been added?").

The above observations were confirmed by Kumar et al. [KT10], who compared pageview categories for 'regular' revisits and long-term revisits, based on a random sample of users drawn from Yahoo! toolbar logs. The main finding of this study was

that half of all pageviews are content (news, portals, games, multimedia), one-third are communication (email, social networking, forums, blog, chat) and the remaining one-sixth are search (including item search and multimedia search). Portal pages receive the largest percentage of revisits, which can be attributed to the promotion and use of homepages of - among others - Yahoo! and MSN as "entry points".

4.2.2 Prediction of Revisits

The problem of the next-page prediction has been extensively studied in the literature. The method that has prevailed in this field, at least in terms of popularity, is Association Rules Mining. *Association rules* (**AR**) constitute a well-established method for effectively identifying related resources without taking into account their order of appearance (e.g., pages that are typically visited together, in the same session, but not necessarily in the same order) [AIS93, AS95]. Numerous works have investigated the performance of different variations of AR [AT01, MCS00, YP02, FBH00, SMB07]. A recent work by Kazienko [Kaz09] explores indirect AR for web recommendations, involving resources that are not 'hardly' connected, as in typical AR.

However, AR suffer from a variety of drawbacks: first, they rely on the most frequent patterns identified in the training set, thus misclassifying new patterns that are not included in it. Second, they fail to recommend rarely visited, and, thus, non-obvious and serendipitous items, since such resources never reach the minimum support limit. Third, disregarding the order of itemsets invariably leads to loss of information about the frequency of different patterns that involve the same resources (e.g., all six permutations of the itemset $I_1 = \{1, 2, 3\}$ are treated equally).

To overcome this last problem, *sequential patterns* have been employed in the context of prediction methods as well. Among them, state-based models like the Markov one, are particularly popular [ZAN99, AZN99, YSW09, DK04, ESRR04, SHB05, AKT08]. More recently, Chierichetti et al. [CKT10] introduced a hybrid of a Markov process capturing the graph of web pages together with a branching process that captures the creation, splitting and closing of tabs. This model was then used to compare tabbed browsing with the simple PageRank model [BP98].

Slightly different from these models are sequence mining techniques that do not take into account the strict order between items [AS95, PHW02, PKBGM10]. A comparison of such techniques with AR was conducted by Géry and Haddad [GH03]. The authors evaluated AR against *Frequent Sequences* (which can be considered equivalent to association rule mining over temporal data sets) and *Frequent Generalized Sequences* (which constitute a more flexible form of the previous technique, involving wildcards [GST01]).

With the aim of introducing a prediction method that is equally effective with unseen data, Awad et al. [AKT08] combined Markov Models with Support Vector Machines (SVM) under Dempster's rule. They compared experimentally their hybrid model with the individual methods comprising it, as well as with AR. The outcomes

demonstrate the superiority of their model (especially when domain knowledge is incorporated into it). Although this is a considerable step toward a method with better generalization capabilities, it is far from being practical: it requires a different SVM classifier for each one of the available resources and a considerably high training time (in fact, their experimental study involved 5,430 classifiers and 26.3 hours of training for a single dataset).

In a more recent work by Parameswaran et al. [PKBGM10], the authors coin *precedence mining* and build a suite of recommendation algorithms based on it. They model a users' history as a set of items having co-occurred in the past (without considering their order of appearance), and predict the set of items most likely to follow in no particular order and not necessarily in the next action of the user. Though quite interesting, their approach is not crafted to deal with the next-page prediction problem, as they explicitly point out.

4.3 Preliminary Analysis of Users' Browsing Behavior

In this section, we briefly report preliminary analysis that lead us to understand users' browsing behavior. We illuminate the most important aspects of users' revisit behavior - general characteristics as well as individual differences - which are used as a basis for the predictive methods that are evaluated in the rest of this chapter. We based this analysis on the history logs of a internet browser described in the next subsection.

4.3.1 Dataset

The participant pool of our data set consists of 25 participants, 19 male and 6 female. Their average age is 30.5, ranging from 24 to 52 years. The participants were logged for some period between August, 2004 and March, 2005. The average time span of the actual logging periods was 104 days, with a minimum of 51 days and a maximum of 195 days. Participants were logged in their usual contexts - 17 at their workplace, 4 both at home and at work, and 4 just at home. During the logging period, 152,737 page requests were recorded. 10.1% of them were removed, as they were artifacts (advertisements, reloads, redirects, frame sets). Hence, in total we have 137,737 page requests available for analysis.

4.3.2 Revisitation Statistics

We recorded an average revisitation rate of 45.6%. Note that this number is lower than in earlier studies, due to the fact that we took into account both GET and POST

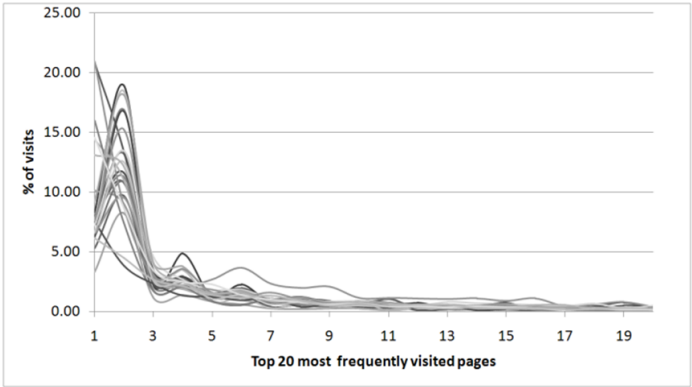


Figure 4.1 Distribution of most frequently visited pages for each user.

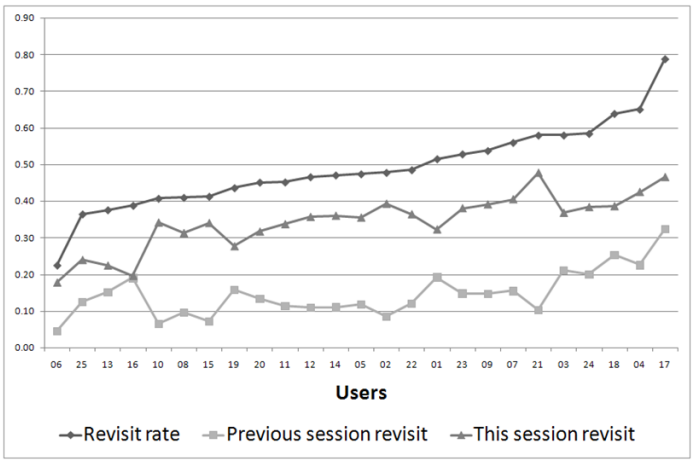


Figure 4.2 Backtracking and routine behavior plotted against the revisit rate (order by revisit rate).

para-meters. The wide variability of individual revisitation range (between 17.4% and 61.4%) suggests that revisitation behavior is heavily influenced by personal habits, private interests and the sites visited [OWHM07]. In this subsection, we concentrate on individual differences between users in their revisitation profile.

As previosly discussed, several studies have identified regularities in revisitation behavior. Users typically have a small set of frequently visited pages, including for example the browser's home page, search engines, favorite news sites, and social networking sites. As can be observed in Figure 4.1, the distribution of most frequently used pages clearly follows a power law for most of the users, but not for all - some have a large number of pages in their browsing routine.

The distribution of revisits to pages based on the number of pages between the last visit and the current visit does follow a power law distribution for all users. Consequently, the backtracking activities (revisits to pages in the current session) and routine behavior (revisits to pages in previous sessions) grow roughly linear with

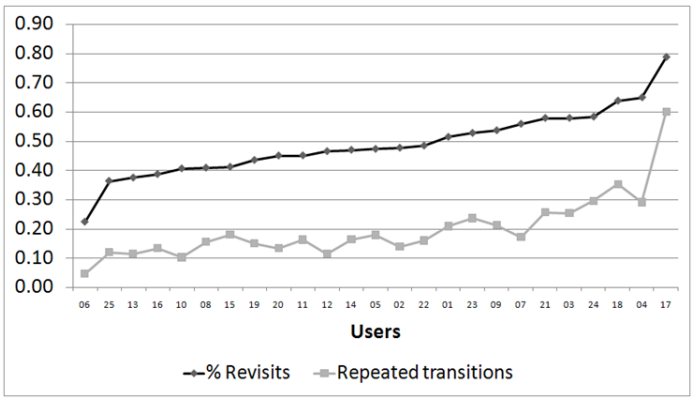


Figure 4.3 Repetitive behavior (% repeated actions) plotted against the revisit rate.

the revisit rate. This illustrated in Figure 4.2 - note that despite the correlation there are still users that can be identified as predominantly backtrackers or predominantly routine revisitors. The average percentage of backtracking actions among revisits is 75%, with a minimum of 51% and a maximum of 84%.

Predictive models of Web navigation, such as Markov models, typically assume that users exhibit a rather large percentage of repetitive behavior, including sequences of pages that are regularly visited in the same order. In Figure4.3 we plot the users' repetitive behavior (based on the ratio between the number of unique pairs of pages that a user visited consecutively and the total number of transitions). The average percentage of repetitive transitions is 20%, with a minimum 5% and a maximum of 60%.

Based on these statistics, it becomes clear that Web page revisitation behavior follows sufficient regularities to be exploited for enhanced revisitation support - in a similar manner as is already common in recommender systems based on Web usage mining and collaborative filtering. Earlier work on revisitation confirms this observation, but only to a limited extent. In this following, we investigate, compare and combine the performance of several predictive methods for page revisitation. Our analysis attempts a general comparison of several prediction mechanisms, with the aim of identifying the best performing one, knowing though that their performance depends heavily on the regularities in the individual user's revisitation activities.

4.4 Contextual Revisit Prediction

In this section, we explain the methods and algorithms used for generating contextual predictions of revisits on the Web. The prediction task can be more formally defined as follows:

Problem 1 *[Page Revisition Prediction] Given a collection of Web Pages,* $P_u = \{p_1, p_2, ...\}$*, that have been visited by a user,* u*, during her past* n *page requests,* $R_u = \{r_1, r_2, , r_n\}$*, rank them so that the ranking position of the* ***page*** *revisited in the next,* $n + 1$*, transaction is the highest possible.*

We developed a generic framework that consists of *two tiers of methods*. The first tier involves usage-based *ranking methods*, which estimate for each web page the likelihood that it will be accessed in the next request. The methods derive their estimate from evidence drawn from the surfing history of a web site or user, such as the recency and/or the frequency of accesses to each page. The second layer covers *propagation methods*; these are techniques that capture repetitiveness in the navigational activity of a Web user and identify groups of pages that are typically visited together (in the same session, but not necessarily in a specific order). Depending on the degree of connectivity between the associated Web pages, their values (assigned by the ranking methods) are then propagated to each other.

The implementation of the framework, *SUPRA*[3], is freely available at SourceForge[4] In this way, we encourage other researchers to experiment with them and to extend the library with new ranking and propagation methods. Special care has been taken to make this a straightforward procedure: any implementation complying with Definitions 9 and 10, which specify the minimal requirements for a ranking and a propagation method respectively, can be easily integrated into the library.

In the next subsections, we discuss the ranking and propagation methods we considered, and how they are combined. We conclude this section with the results of an experimental evaluation of the framework.

4.4.1 Ranking Methods

The aim of ranking methods is to provide for each item a numerical estimation of the likelihood that it will be accessed in the next transaction. After each page request, the selected ranking method goes through all visited items of interest (either pages or sites), estimates their value and sorts them in descending order of their expected value. The estimation is based on the access history of each item, represented by the indices of the related requests:

Definition 8 *[Item Request Indices] Given the page requests* R_u *of a user* u*, the request indices of an item* m_i*,* I_{m_i}*, is the set of the serial numbers of those requests in* R_u *that pertain to* m_i*. The serial number of the chronologically first request is 1 and is incremented by 1 for each of the subsequent page visits.*

Given this definition, a ranking method is defined as follows:

[3]SUPRA stands for SUrfing PRediction frAmework.

[4]See `http://sourceforge.net/projects/supraproject`.

Definition 9 *[Ranking Method] A ranking method is a function that takes as input the request indices of an item m, $I_{m_i} = \{i_1, i_2, \ldots, i_k\}$ together with the index of the latest request, i_n, of the respective user, and produces as output a value $v_{m_i} \in [0, 1]$ that is proportional to the likelihood of m_i being accessed at the next page request, r_{n+1} (i.e., the closer v_{p_i} is to 1, the higher this likelihood).*

In this work, we consider the following ranking methods (modified appropriately to be consistent with Definition 9):

1. Least Recently Used (**LRU**),

$$LRU(m_i, I_{m_i}, i_n) = \frac{1}{i_n - i_k + 1},$$

2. Most Frequently Used (**MFU**),

$$MFU(m_i, I_{m_i}, i_n) = \frac{|I_{m_i}|}{i_n},$$

3. Polynomial Decay (**PD**),

$$DEC(m_i, I_{m_i}, i_n) = \sum_{j=1}^{|I_{m_i}|} \frac{1}{1 + (i_n - i_j)^{\alpha}}, \ \alpha > 0$$

where i_k is the index of the chronologically last transaction in I_{m_i}, i_n is the index of the latest request of the system or user, and $|I_{m_i}|$ is the cardinality of I_{m_i}.

The first two methods, LRU and MFU, constitute well-established caching algorithms that are typically employed in prediction tasks. LRU is based on the assumption that the longer in the past a page was accessed, the less likely it is be accessed in the future. Similarly, MFU assumes that the more frequently a page is accessed, the more likely it is to be accessed in the next request. Thus, the former orders items according to the recency of their last request, whereas the latter sorts them in descending order of their popularity. PD, on the other hand, is based on the *decay ranking model* introduced by Papadakis et al. [PNN10]. It incorporates recency and degree of usage into a single, comprehensive method, balancing them harmonically through the smooth decay of the contribution of each request to the total value of an item. Factor a is available for tuning this equilibrium, by defining the intensity of the decay: values larger than 1 convey a steeper decay, which puts more emphasis on recency, while values close to 0 promote frequency of usage. In general, the best value for a depends on the application at hand, but, as verified in [PNN10], values between 1.0 and 2.0 provide performance close to the optimum, outperforming both LRU and MFU.

4.4.2 Propagation Methods

The purpose of propagation methods is to capture contextual information through the detection of patterns in the surfing activity of users. They identify those items that are commonly visited within *the same session* and associate them with each other. The 'links' created by these methods are used to propagate between the associated pages the values assigned to them by the ranking methods. In this way, the highest the value of a web page, the more the pages associated with it are boosted and the more their ranking is upgraded.

Sessions are transparently defined by browsers, and typically include all pages visited within the same tab of the browser for up to a specific time period. The temporal limit, though, can vary from browser to browser, and, thus, we do not provide a formal definition of a session. Instead, we consider a **session** S to be a bag of visited items, defined by the browser, that are placed in chronological order from the earlier to the latest: $S = \{m_1, m_2, \ldots, m_k\}$.

Based on the above, propagation methods can be formally defined as follows:

Definition 10 *[Propagation Method] A propagation method is a function that takes as input the latest requested item, m_i, within a session S, and defines appropriately the degree of connection between m_i and all other items visited during S. Hence, given two items, X and Y, it returns a value $v_{XY} \in [0, 1]$ that is proportional to the likelihood of Y being accessed immediately after X (i.e., the closer v_{XY} is to 1, the more likely this transition is).*

We distinguish between two families of propagation methods: *order-neutral* methods, which disregard the order of transactions within a session and *order-preserving* methods, which take this order into account. For the former case, we examine association matrices. For the latter case, we consider transition matrices.

Order-Neutral Propagation Methods. Order-neutral methods are based on the idea that pages visited in the course of the same session should be equally connected with each other, regardless of their order and the number of transitions that intervene between them. The rationale behind this idea is that users may visit a group of pages X, Y, Z on a regular basis, but not necessarily in that order.

We employed association matrices (**AM**) for order-neutral propagation. An AM is a matrix, whose rows and columns are the given set of web pages P. The AM is built by associating all pages visited in a single session with each other (i.e., each web page is connected not only with the pages preceding it, but also with those following it). Thus, an AM is always a symmetrical matrix ($\forall x \;\; AM(x, x) = 0$) and each cell $AM(x, y)$ expresses the number of sessions that involve both items x and y.

Order-Preserving Propagation Methods. This category of propagation methods relies on the idea that pages are typically accessed in the same or similar order. Order-preserving methods build the associations between pages according to this

ordering: each page is connected only with pages preceding it. To capture these transitions that form chronological patterns in the navigational activities of systems and users, we employ transition matrices.

In short, a transition matrix (**TM**) is a matrix with its rows and columns representing all pages visited by the user. Each cell $TM(x, y)$ expresses the number of times that a user visited item y *directly after* x. Given that a transition matrix respects the order of accesses within a session, it is not symmetrical: $\exists x, y : TM(x, y) \neq TM(y, x)$. Moreover, its diagonal cells are all equal to 0: $\forall x \ TM(x, x) = 0$.

We conducted a series of experiments to identify which propagation method produces the best results for our problem [KPH10]. Together with the order-neutral AM, we evaluated four kinds of order-preserving propagation methods. *Simple Connectivity TM* (**STM**): after each transition $x \rightarrow y$, only the value of the cell $TM(x, y)$ is incremented by one, thus functioning exactly like a first-order Markov model. *Continuous Connectivity TM* (**CTM**): each web page visited within the current session is associated with all previously accessed pages. *Decreasing Continuous Connectivity TM*(**DTM**): this strategy lies in the middle of STM and CTM; the cell values are determined based on a decay parameter representing the distance between page visits. *Increasing Continuous Connectivity TM*(**ITM**): this strategy increases the value added to $TM(x, y)$ in proportion to the distance between pages visits.

Of the above methods, the *Simple Connectivity Transition Matrix* (**STM**) produced the best results, which provides support to the assumption that page revisits tend to occur in the same strict order. It is worth noting at this point that STM was also employed in Awad et al. [AKT08], but its frequencies were merely used as features for a classification algorithm. It was also employed in [RFST10] as a means to model the behaviour of individual users and to recommend relevant items to users by combining their matrices.

4.4.3 Combining Ranking with Propagation

To combine a raking method with one of the propagation techniques, we employ a simple, linear scheme: following the i_n-th page request, the value of all items is (re)computed, according to the selected ranking method. Then, for each non-zero cell of the TM (or AM) at hand, $TM(x, y)$ (or $AM(x, y)$), we increment the value assigned to page y by the ranking method, v_y, as follows:

$$v_y += p(x \rightarrow y) \cdot v_x, \text{ where}$$

- $p(x \rightarrow y)$ is the transition probability from item x to item y, estimated by $p(x \rightarrow y) = \frac{TM(x,y)}{\sum_i^{i_n} TM(x,i)}$ (or $p(x \rightarrow y) = \frac{AM(x,y)}{\sum_i^{i_n} AM(x,i)}$), and
- v_x is the value of x estimated by the ranking method.

Table 4.1 Summary of Experimental Results

Method	ARP	S@1	S@10
MFU	307 (σ=178)	12.7 (σ=3.8)	32.2 (σ=5.4)
LRU	65 (σ=30)	19.3 (σ=3.8)	71.2 (σ=4.3)
PD	60 (σ=27)	19.3 (σ=3.8)	71.7 (σ=4.2)
MFU+STM	288 (σ=168)	12.6 (σ=3.8)	32.1 (σ=5.4)
LRU+STM	32 (σ=14)	23.8 (σ=3.7)	81.5 (σ=2.8)
PD+STM	31 (σ=14)	22.7 (σ=3.3)	81.8 (σ=2.8)

4.4.4 Experimental Study on Page Prediction

Setup. To evaluate our framework, we conducted an experimental study using data from a client-side Web usage log of 25 users with a total of 137,737 page requests, gathered in the course of 6 months [5]. The participant pool of the data set consists of 25 participants, 19 male and 6 female. Their average age is 30.5, ranging from 24 to 52 years (the same dataset used in Section 4.3).

We simulated the navigational activity of each user independently of the others. After each page request, the ranking of all visited pages was updated, and, in case the next access was a revisitation, the position of the corresponding web resource was recorded. Having all these ranking positions for all prediction methods, we derived the following metrics to evaluate their performance:

1. *Success at 1* (**S@1**). It denotes the portion of revisitation requests that involved the page placed at the first ranking position by the prediction method. The higher its value, the better the performance of the method. S@1 is interesting as it provides evidence for the accuracy of a prediction method in identifying the next revisited page.

2. *Success at 10* (**S@10**). It expresses the portion of revisits placed in one of the first 10 places. The higher its value, the better the performance of the method. S@10 expresses the actual usability of the prediction method, as users typically have a look only at the first 10 pages presented to them (just like they do with web search engine results [HCBG01]).

3. *Average Ranking Position* (**ARP**). It represents the average position of a revisited page in the ranking list that the prediction method produces. ARP provides, thus, an estimation of the overall performance of a prediction method, as it considers the performance over all the revisits in the navigational history of a system or user, and not only the top ranked ones. The lower its value, the better the performance of the prediction algorithm.

Results. We compared the performance of the ranking methods LRU, MFU and PD by simulating these methods on the dataset described earlier in this section. Similarly, we evaluated the three ranking methods when combined with the propagation

[5] This is the data set that was used in [OWHM07, WOHM06]

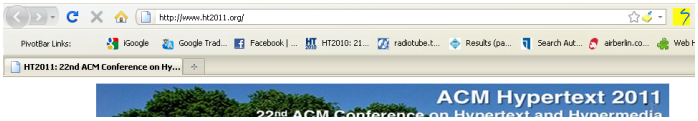


Figure 4.4 PivotBar recommendations

method STM[6]. We configured the steepness of the PD decay model with $\alpha = 1.5$. The results are summarized in table 4.1.

Of the ranking methods, MFU performs much worse than LRU and PD with respect to all metrics. This indicates that backtracking is more common than revisiting popular sites; moreover, frequent revisits to popular sites are largely covered by the list of recently used pages. LRU and PD have similar performance in terms of S@1 and S@10, but PD has a slightly better ARP, due to the incorporation of the frequency of usage. Their combination with the propagation method STM takes into account the current user context, as well, thus improving significantly the performance of LRU and PD (t(24)=10.1, p¡0.01); combining MFU with STM hardly causes any change. This can be explained by the interaction between the recency effect and the current user context. The S@10 of the combined ranking and propagation methods performs up to 81.8%.

4.5 User Evaluation of Contextual Recommendations

To explore the actual usage and appreciation of our prediction framework, we developed the *PivotBar*, a browser toolbar that looks quite similar to the bookmark toolbar, containing favicons and links to already visited pages (see Figure 4.4). In contrast to the bookmark toolbar, however, PivotBar is dynamic, providing contextual recommendations; after each navigation action or tab change, the list of pages in the bar changes, containing the most relevant visited pages to the current one.

The design of the toolbar is kept minimalistic, in order to avoid occupying a large part of the browser's interface. By placing it right under the URL field, we ensure that the dynamic character of the list catches user's attention only in the periphery and just for a short time period - unless the user chooses to follow a recommendation.

For the first implementation of the PivotBar, we chose Mozilla Firefox as the host browser, since it constitutes a freely available and platform-independent browser that provides developers with clear-cut documentation and transparent access to client data. The PivotBar Add-On makes use of the existing user history in the browser database and all computations take place on the client-side.

[6]The results using the other propagation methods were lower than the STM results, therefore we left them out of the discussion.

It is worth clarifying at this point that PivotBar is not designed for extensive search into the history - an activity that users hardly undertake anyway. Instead, it exclusively aims at reminding users of past visits that are judged relevant to the currently viewed page. For example, when planning a train-ride, the user will be prompted to visit his favorite hotel booking site, if she had done so in a similar situation in the past.

4.5.1 Diversity of Recommendations

At the core of our toolbar lies a composite prediction method that employs PD for ranking web pages and STM for propagating their values (see Section 4.4.1). The reason for this choice is twofold: first, these methods have exhibited the highest performance in their category, not only individually, but also in combination. Second, PD (and subsequently the propagation method on top of it) provides the best trade-off between the diversity and the relevance of the recommendation sites.

To verify the latter claim, we compared the average entropy of the top-10 recommendations, as generated by the ranking methods of Section 4.4.1, making use of a dataset consisting of 116 users with an average of 960 revisitations per person (see Section 4.6.2). The average entropy was estimated to be 4.2 for MFU, 7.9 for PD and 8.8 for LRU. This means that these methods recommend, on average, 18 (MFU), 240 (PD) and 445 (LRU) distinct pages. In contrast to the rather static nature MFU, LRU provides more diverse recommendations - but these pages are already accessible through the back button. In the middle of these two extremes lie the recommendations of PD.

4.5.2 Study Setup

The goal of our user study was to get an answer to the following questions: first, will users actually click on recommendations? In other words, will the toolbar be used? Second, what would be the user's appreciation of a dynamic toolbar? Third, which could be the directions for further improvement of the recommendations?

To this end, we asked 11 participants, aged 28 on average, to install the toolbar, either on their business computer or on their private one. Eight opted for the former choice, and the remaining three for the latter. Users were then provided with a brief introduction to the tool and some instructions for the experiment[7]. The participants were asked to keep the tool installed for a period of five working days. With the passage of this period, we collected the the quantitative results through the click-data of each participant, while qualitative feedback was elicited via an open-ended interview.

[7]The exact instructions given to the participants were: "PivotBar automatically generates suggestions based on the current page you are accessing. You can use them simply by clicking on a link to be redirected to the target page. Feel free to use them or not."

Table 4.2 Click data during the evaluation period.

User	Total Visits	Revisits	PivotBar	Percent (%)
1	541	264	104	39.4
2	596	248	38	15.3
3	352	147	49	33.3
4	828	424	49	11.6
5	321	63	10	15.9
6	567	283	39	13.8
7	259	137	20	14.6
8	179	102	40	39.2
9	183	75	19	25.3
10	312	149	14	9.4
11	423	145	46	31.7

4.5.3 Results

All participants claimed to use the computer for about 6 to 8 hours per day. They all indicated that they typically use the auto-completion feature for revisitation, while half of them actively uses bookmarks, as well. Further, they acknowledged that they often use search engines to refind a known page. The recurrence rate during the evaluation period reached an average of 44.2% (σ=10.4), lying at the same levels indicated by previous studies [CM01, OWHM07].

Table 4.2 summarizes the usage of the PivotBar for each participant. The second column indicates the total number of pages visited. The third column represents the number of revisits among the page requests (including requests for pages visited before the start of the evaluation period). The fourth column corresponds to the number of revisits that were initiated through the PivotBar. The fifth column shows the percentage of revisits covered by the PivotBar.

The average percentage of revisits through the PivotBar was 22.7% (σ=11.4), reaching a peak of 39.3% for participant 1. This number is surprisingly high - even if one takes the novelty effect into account. As a comparison, [OWHM07] observed that the back button covered 31% of all revisits, while bookmarks, the history list and the homepage button together were responsible for a mere 13.2% of all revisits.

Quite interesting was the qualitative feedback that we received via the open-ended interviews. When asked about the usage of the toolbar, one of the participants explicitly commented: "*I actually scan the shortcuts automatically when they change. The movement attracts my attention, without being distractive*". Another participant said: "*It's nice that I can see the pages that I usually access*". At the same time, though, he admitted that his routine behavior was hard to change: he still tended to automatically open a new tab and directly type the address of a page using auto-complete. This explains why for some users the usage percentage of PivotBar remains low, around 10%.

The participants also provided suggestions for further improvements. Some of them proposed to further reduce the influence of recency on the recommendations,

favoring more serendipitous ones. Others thought that recommendations should be based on the currently visited site instead of the page (*site-level recommendations*). Finally, quite a few participants argued that the toolbar should recommend (portal pages of) sites instead of (specific) pages.

The comments about the preference for site-level recommendations can be explained by the growing importance of revisits to service-oriented sites and the monitoring of news sites [ATD08]. However, site-level recommendations would ignore the informational value of specific news articles, blogs and other listings. News portals, on the other hand, continuously add new articles, which cannot be covered by a revisitation prediction method.

4.6 Recommending Pages vs Sites

The results of the user evaluation in the previous section show that, when users are provided with relevant suggestions for page revisits, they will click on these suggestions. In the evaluation, the PivotBar recommended pages based on the currently visited page. Participant feedback suggested that it might be even more beneficial to provide suggestions for (portal pages of) Web sites instead of individual Web pages - or to use the currently visited site (not the specific page) as a basis for the prediction.

We address these suggestions with a second experiment and user evaluation in the following sections. In this section, we formalize the site prediction task and introduce the dataset used for the second experiment.

4.6.1 Site Revisitation Prediction

For clarity, we start the discussion of our experiments with a couple of definitions. We consider as a *Web Site* a domain that comprises a set of *Web Pages*. For instance, `http://www.ht2011.org/tracks.html` is a page under the site `http://www.ht2011.org`. In the following, we consider each page to contain in its description, the URL of the corresponding Web Site.

Similar to Problem 1, the task of recommending sites is defined as:

Problem 2 *[Site Revisition Prediction] Given a collection of Web Pages, $P_u = \{p_1, p_2, ...\}$, that have been visited by a user, u, during her past n page requests, $R_u = \{r_1, r_2, , r_n\}$, rank them so that the ranking position of the* ***site*** *revisited in the next, $n+1$, transaction is the highest possible.*

In the following sections, we present two approaches to this problem, together with a new one for Problem 1.

4.6.2 Dataset

In order to verify the efficiency and performance of the new predicting methods, we started an effort to gather users' navigational data through the Web History Repository[8]
(**WHR**).

Web History Repository

The Web History Repository Project aims to build a public repository of web usage data, which can be used by researchers to gain new insights in online browsing behavior. Using a Mozilla Firefox add-on[9], users can upload their anonymized usage data to the server. These data include the list of visited pages together with the timestamp and browser session of each request. A separate table stores for each visited page its (encrypted) URL and host, the total number of visits, the frecency and the last visit.

The Web History Repository was promoted through several targeted mailinglists, Facebook, Blogspot and Twitter. One month after the release of the add-on, more than 100 anonymous volunteers contributed over 1 million entries from their browser history. At this point, we considered the dataset large enough to give us significant results for our experiments. In contrast to the dataset of Section 4.4.4, the data of WHR are totally anonymized, and, thus, we do not have at our disposal any demographic information about the users.

Characteristics and Analysis

The dataset we used contained the navigational history of 116 users with a total of 1,006,941 page visits. The user with the largest history contributed exactly 6 months of data with 77,398 page visits. The average time period of the history for all users was 56 days. We pruned the data to remove users with less than one thousand visits. The remaining dataset consisted of 61 users and a total of 951,995 page visits, still representing 94.5% of the dataset.

For each of the selected users, the average number of page visits is 15,606 (σ=18,893), in a period of 87 days (σ=82). This corresponds to an average of 179 pageviews per day. The average recurrence rate is 34% (σ=14), slightly lower than the recurrence rate in other studies. By contrast, the recurrence rate per host (the relative number of visits to a site that constitute a revisit to this site) is astonishingly high at 92% (σ=5, min=69%, max=99,9%). This implies that only 8 out of every 100 pages we visit on the Web belong to new, unseen domains; in other words, Web use is mainly restricted to a more or less fixed set of sites that provide the services or information that the user needs. Figure 4.5 illustrates the linear growth of page visits, unique

[8]See `http://webhistoryproject.blogspot.com`.
[9]https://addons.mozilla.org/en-US/firefox/addon/226419

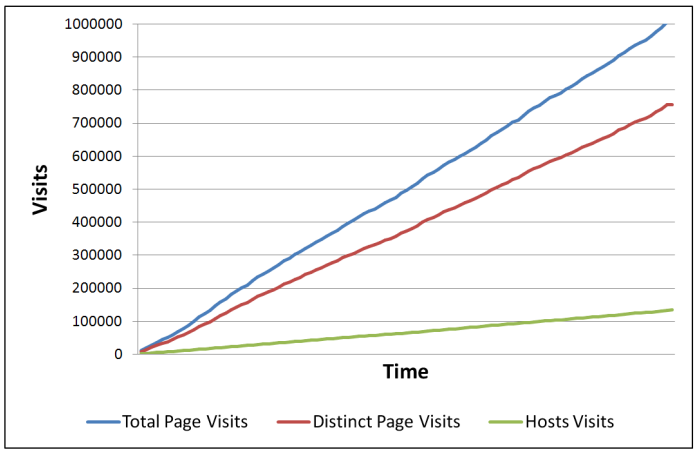


Figure 4.5 Growth of page visits over time.

Table 4.3 Comparison of the datasets with previous studies.

	Catledge & Pitkow	Tauscher & Greenberg	Dataset 1 [OWHM07, WOHM06]	Dataset 2 (WHR)
Time of study	1994	1995-1996	2004-2005	2010
No. of users	107	23	25	61
Length (days)	21	35-42	52-195	1-385
No. of visits	31,134	84,841	137,272	951,995
Recurrence	61%	58%	45.6%	35.9%
Back	35.7%	31.7%	14.3%	∼7.5%

pages and hosts (domains) in the dataset. Table 4.3 provides a comparison with the datasets used in previous studies.

4.7 Experimental Study on Site Prediction

To re-evaluate our method, we ran a second experiment following the same procedure as in the first experiment, which we described in Section 4.4.4. We used the pruned dataset of the Web History Repository. For this experiment we did not vary the prediction method, but employed the first experiment's best-performing method: PD+STM. Instead, we varied the basis for the contextual prediction (page or site) and the type of suggestions (page or site). The following four strategies were considered:

- Page to Page recommendation (as in Experiment 1)
- Page to Site recommendation
- Site to Page recommendation
- Site to Site recommendation

Table 4.4 Summary of Experimental Results

Method	ARP	S@1	S@10
Site-to-Page	285 (σ=166)	5.0 (σ=2.1)	46.2 (σ=8.0)
Page-to-Page	168 (σ=80)	15.3 (σ=6.9)	61.6 (σ=5.9)
Site-to-Site	22 (σ=12)	20.9 (σ=4.0)	78.0 (σ=4.1)
Page-to-Site	23 (σ=12)	33.9 (σ=5.6)	79.4 (σ=4.2)

4.7.1 Evaluation Measures

Similar to the first experiment, the evaluation measures used are S@1, S@10 and the Average Ranking Position. Further, in order to investigate differences in prediction performance between users, we used a number of measures to characterize their individual behavior. The first measure is the (page) recurrence rate [TG97]:

$$R = (1 - \frac{individual\,pages\,visited}{total\,page\,visits}) \times 100\%$$

The site recurrence rate is defined analogously to the page recurrence rate. A further measure we used was the *page* and *site entropy*, which characterizes the variance (or disorder) in the user's log:

$$E = \sum_i (p_i \times log_2(p_i)),$$

where p_i is the probability of page/site i estimated as $p_i = (|I_{p_i}| - 1)/(\sum_i (|I_{p_i}| - 1))$.

The other measures we used are fairly straightforward: the average number of pages visited per site, per day and per session.

4.7.2 Results

The results of the four prediction strategies are summarized in Table 4.4. A first observation is that the page-to-page prediction results for this dataset are considerably lower (S@10=61.6) than for the dataset used in the first experiment (S@10=81.8). We attribute this to the larger variance in user behavior due to the way the dataset was created. Further, the S@k measures suggest that site predictions are more successful than page predictions. In addition, looking at the ARP, the average ranking position is much lower for sites than for pages. This effect can be explained by the fact that there are far less candidate sites to predict than candidate pages, which makes site prediction a safe fallback alternative for page prediction. It is also clear that page and site predictions alike perform better if they are based on the current page that the user visits instead of the current site. Finally, the differences in performance of the four strategies between individual users are highly correlated with $p<0.01$ (Pearson, 2-tailed), which implies that for users for whom one strategy performs well, other strategies will perform well too.

Individual differences. It is a likely assumption that individual differences in prediction performance are caused by differences in the user's online browsing behav-

ior. For each user, we captured the browsing behavior in the measures introduced earlier in this section. We carried out stepwise linear regression to find out which aspects perform best in predicting the performance of page-to-page and page-to-site recommendation (in terms of S@10). The results indicate that the site entropy is the most important predictor, accounting for 22% of the variation in page prediction and 53% of the variation in site prediction; the page entropy explains another 9% of the variation in site prediction. Surprisingly, the page and site recurrence rates (which indicate to what extent a user revisits pages) are only weakly correlated to the prediction performance as well as to the entropy measures. In summary, the results indicate that it is not the amount of revisits, but the variance in revisit behavior that directly impacts the performance of any prediction algorithm.

4.8 Second User Evaluation

To evaluate the new methods with respect to real users, we carried out a second user evaluation with the PivotBar, as introduced in section 4.5. We modified the underlying methods so that users get a combined set of recommendations of pages and sites. For this, the following heuristic was used: if a recommended page has been visited less than 10 times before, the recommendation is replaced by the portal page of the recommended page's site, on the condition that this portal page has been visited before. The threshold of 10 is derived from the average distribution of page visits, which approximately defines the end of the head. In addition, following the suggestions of participants from the first evaluation, we added a new feature to the PivotBar that allows users to permanently hide a recommendation by adding a page or a site to a blacklist.

As in the first evaluation, our goal is to check the usability of the tool and whether the recommendations have an impact on users' navigational behavior. We evaluated this with respect to two evaluation measures: first, we observed the number of revisits triggered by clicks on the PivotBar. Second, we estimated the number of "*blind hits*"; that is, revisits that were not triggered by the PivotBar, but that were in the list of recommendations displayed in the toolbar.

4.8.1 Evaluation Setup

The setup for this evaluation was similar to the evaluation presented in Section 4.5. This time we had a total of 13 participants, aged 29 on average. Eight participants had the PivotBar installed at their work computers, the other five at their private computers. The instructions for using the tool were the same as before, with additional details about the functionality of the blacklist The participants were asked to keep the tool installed at least for a period of ten days. After this time period, we collected the click-data of each participant for the quantitative results; qualitative

feedback was elicited through open-ended interviews.

4.8.2 Results

Table 4.5 summarizes the usage of the PivotBar for each participant. The second column indicates the total number of pages visited during the evaluation period. The third column represents the recurrence rate among the page requests (including revisits to pages visited before the evaluation). The fourth column shows the percentage of revisits triggered by the PivotBar and finally, the fifth column shows the percentage of blind hits.

On average, 12.1% (σ=7.3) of all revisits resulted from a click on the PivotBar, reaching a peak of 30.8% for participant 1. The average percentage of blind hits was 18.1% (σ=12.0), meaning that these revisits were suggested in the PivotBar but not triggered by it. The strong correlation between the PivotBar clicks and blind hits ($r=0.92$, $p<0.01$) suggest a direct connection between the quality of recommendations and the take-up of the tool.

A further indicator of engagement is provided by the usage of the blacklist. The average number of removed pages per user was 7.2 (σ=14.3). Participant 10 had a total of 52 pages and hosts in her black list, while 3 other participants had an empty blacklist. However, the usage rate of the PivotBar for participant 4 (15.9%), who had an empty blacklist, was above the average and much higher than the engaged Participant 10 who pruned her results.

During the open interviews, all participants stated that the PivotBar was indeed useful, with few complains about visual issues due to compatibility with a specific operational system. When asked for what reasons they considered PivotBar to be useful, answers included the following: "*Because for the pages that were good suggestions, I didn't need to start typing the URL*", "*It was faster for reaching the pages I wanted*" and "*It was easier to remember pages that I have visited*".

None of the participants noticed that sometimes a specific page was recommended and sometimes the website. At the same time, we also did not receive any remarks that recommendations for very specific pages could better point to the associated site's portal page (which was one of the main remarks during the first experiment). We consider this lack of remarks as positive feedback.

4.9 Making Sense of Browsing Context

Many of the places that we visit on the Web are places that we visited before. The majority of revisits is covered by a small number of popular places- such as the user's favorite search engine, online retailers, social networking and news sites - and places visited in the very recent past [CP95]. The same power law distribution can be observed in the overall popularity of Web sites, friend connections in social networking

Table 4.5 Click data during the evaluation period.

User	Total Visits	Revisit (%)	PivotBar (%)	BlindHits (%)
1	603	50.1	30.8	22.8
2	535	45.0	19.5	51.0
3	445	39.6	15.9	8.5
4	578	51.2	15.9	15.9
5	1,111	36.1	13.0	20.7
6	716	45.5	12.3	28.8
7	1,219	49.1	8.8	18.0
8	899	41.7	8.8	8.5
9	379	56.2	7.0	11.7
10	1,047	39.6	5.8	16.1
11	1089	43.3	4.7	7.6
12	674	29.4	11.1	6.6
13	896	34.6	3.9	19.0

sites and tag collections [BFNP08b]. Search engines and recommender systems have exploited these regularities for several decades [BP98].

In addition, as we have already seen, the *folksonomy* of tags given by users to various resources [HJSS06b] is successfully exploited to build tag-based profiles of both users and resources. These profiles are used for personalization [CCC+08], recommendation and improvement of search [BFNP08b].

The starting point of the research discussed in this section is the observation that recurrent activities on the Web represent recurrent user interests, tasks and goals; many of the revisited resources have been annotated with tags by various users, and these tags represent the 'wisdom of the crowd' on what these resources are used for. This public folksonomy is assumed to be more representative than the user's individual tags - which are often subjective [vSBvV+06] and low in number - or the keywords in the page title.

We aim to identify and explain 'canonical' patterns of reoccurring activities based on tag occurrences in the users' online lives. We do this by relating client-side Web usage logs with the tags that describe the resources in these logs. We developed a classification of the most common patterns of user interest by clustering keywords by their appearance on the users' timelines. These patterns vary from one-time interest to repetitive peaks and constant interest. An analysis of the top keywords related to these patterns shows that these patterns differ from one another in terms of user interests, tasks and goals. To proof that it is feasible to use the classification for automatically recognizing these patterns, we implemented and evaluated a simple rule-based heuristic classifier.

4.9.1 Generating a Virtual Folksonomy

A traditional *folksonomy* is a quadruple $\mathbb{F} := (U, T, R, Y)$, where U, T, R are finite sets of instances of *users*, *tags*, and *resources*, respectively. Y defines a relation, the *tag assignment*, between these sets, that is, $Y \subseteq U \times T \times R$, possibly enriched with

a timestamp that indicates *when* it was performed [HJSS06b].

We created a *virtual folksonomy* by enriching the a client-side Web usage log - which contains Web pages (R) that are visited by users (U) - with tags (T), making use of the social bookmarking system Delicious. We call the folksonomy 'virtual' because of the indirect manner in which tags are associated with the individuals' Web histories.

Once again we used the Web History Repository (see Section 4.6.2). The data includes the list of visited pages, including timestamp and browser session. For each visited page, the (encrypted) url and host, the total number of visits, the frecency and the last visit is listed in a separate table. At the time of this research, the repository has grown, containing data of 201 users, with a total of 1,324,041 visits to 857,271 unique URLs.

We crawled the online bookmarking system Delicious to retrieve the user-provided tags associated with each URL, thus enriching the web usage log into a virtual folksonomy. In total we found 10,696 unique URLs that have been tagged with 331,699 tags, summing up to a total of 64,179 unique tags. As expected, Delicious contained tags for only a subset of the pages in the Web usage logs. Still, these pages accounted for 7% of the total number of page visits and thus sufficiently covers the long tails in the user's logs.

As analyzed by [AHH+12], the combination of user-specific usage data with popular tags is an effective mechanism for improving the performance of tag recommendations, in particular during the cold-start period, when little or nothing is known about the user. Further, apart from enhancing incomplete profiles, it is a method for diversifying the profile semantics by combining a user's specific behavior with the wisdom of the crowd.

4.9.2 Tag-based Interest Patterns

In this section, we focus on the identification of tag-based user interest patterns. As a first step, we clustered the tag revisitation curves based on the similarities with respect to time; we use the most common keywords associated with each cluster to explain its meaning. Second, based on the general shapes of the clusters, we developed a rule-based classifier that maps each keyword to the groups derived from the cluster. At the end of the section, we discuss the characteristics of the interest patterns found.

Clustering Interests

In order to identify 'canonical' patterns of recurrent user interests, we followed the clustering and classification approach introduced by Adar et al. [ATD08], who used it for evaluating revisitation behavior for different URLs. The clustering output is a normalized revisitation curve that identifies different types of revisitation. In our case, the URLs are translated into the tags that are associated with the URL in

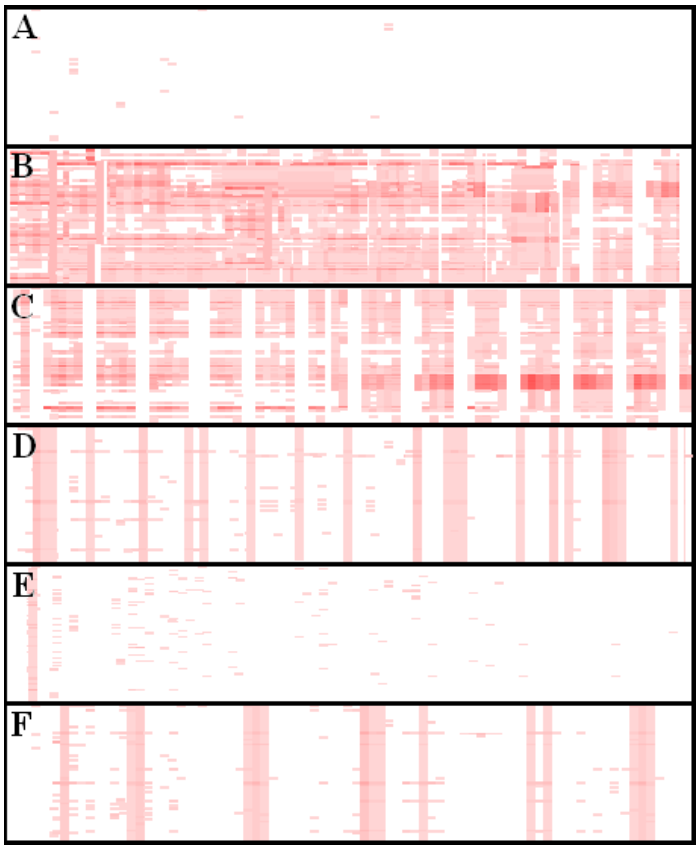


Figure 4.6 The different clusters plotted by Cluto. Each row represents a cluster; darker colors represent a higher number of occurences

the virtual folksonomy. Due to the overlap of keywords between URLs, the curves represent generic user interests rather than reuse of specific pages or sites. As we are interested in longer-term patterns, we group the keywords of interest in buckets of each one day. To align differences in starting point and time span covered by the logs between users, we employed several normalization strategies, as used by Yang and Leskovec [YL11].

All data was aligned by shifting all first keyword appearances to a 'point zero', all further appearances of this keyword were shifted to the corresponding distances from this point zero. To observe weekly routines, we preserved the weekday information during the shifting process: for example, a curve of interest that started on a Tuesday in the second month of a user's history is shifted to start on Tuesday in 'week zero'. We did not normalize the time span of the keyword life times, as techniques such as Dynamic Time Warping would introduce artificial patterns due to the stretching.

With the aligned data, we used repeated-bisection clustering with a cosine similarity metric [RK04] [10]. Varying the similarity metrics and the number of clusters, we found six well defined clustered behaviors, as depicted in Figure 4.6. We manually analyzed these clusters, named them based on general trends and summarized these trends with descriptions and example keywords, as depicted in Table 4.6. It is worth noting that the descriptions are derived from our qualitative analysis of the most representative tag-examples found in each cluster.

Classifying Interests

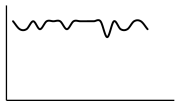
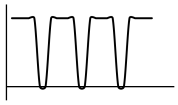
Following the clustering process, we implemented a rule-based heuristic classifier that assigns a keyword to one of the groups that correspond to the identified clusters, based on the keyword's occurrence pattern. With the classifier, we aim to verify the usefulness of the classification derived from the clusters, in terms of discriminative power. Further, the distribution of keywords in the groups is expected to provide insight in temporal dynamics of user interest.

During the classification process we recognized a missing pattern that was not clearly identified during the data clustering, due to few occurrences and similarities with other clusters. The 'missing' canonical curve represents interests that happened during a continuous period of time in the users' history but that never pops up again (C7), as depicted on the top right of Figure 4.7.

Once the seven classifications were defined, we implemented a mutually exclusive classifier based on a set of rules that identify the canonical curves. In other words, each user's interest belongs to one, and only one group. The classifier incrementally iterates over the whole array of occurrences of a tag and, for each iteration, it assigns the possible group to the tag. This implementation allows us to incrementally verify the classification changes of each tag over time and also supports streaming data (as

[10] http://glaros.dtc.umn.edu/gkhome/views/cluto/

Table 4.6 Summarization of Web users' interest.

Group	Shape	Description
C1 - One-time interest.		This group represents an interest that happens a single time during the user's history. This includes spam, involuntary access and typos.
C2 - Constant interest.		This group shows high a constant interest of the user in a topic or service: Search engines, news papers, webmail.
C3 - Constant interest with repetitive drops.		This group represents an constant interest with repetitive drops, mostly caused by weekend breaks. Similar to constant interest in a working environment.
C4 - Repetitive peaks.		This group represents regular, repetitive peaks of interest, mostly caused by exclusive weekend accesses and weekly routines. Websites of games, sports, TV shows, regular meetings.
C5 - Sporadic standalone peaks.		This group contains interests that return on an irregular basis and do not last longer than a day. This includes finance, specialized reference sites, restaurant finding.
C6 - Sporadic connected peaks.		This group shows interests that return on an irregular basis and that typically last longer than a day, such as online shopping, travel planning and research activities.

is the case of browsing history in real use). The rules can be summarized as follows:

- C1 - a keyword is used on one single day
- C2 - a keyword is used during a longer consecutive period, containing only a few days on which the keyword is not used
- C3 - a keyword is used during the whole logging period, containing several days on which the keyword is not used; gaps between these days are evenly distributed

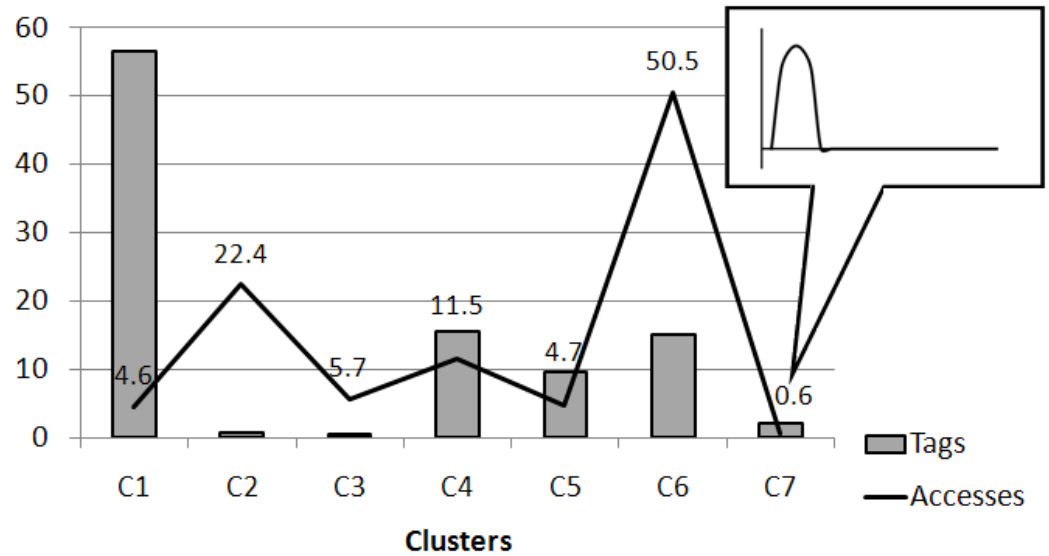


Figure 4.7 Y-Axis shows the distribution of the tags in clusters (bars) and the number of page visits covered by the group (line)

- C4 - a keyword is used on a regular basis, low deviation of gaps between appearance
- C5 - a keyword is used on a regular basis, high deviation of gap length between appearances, peaks last only a day
- C6 - similar to C5, but the peaks of keyword appearance last longer than a day
- C7 - a keyword is used in one single period of more than one consecutive days

We evaluated the classifier on the virtual folksonomy, as described in Section 4.9.1. To avoid bias introduced by popular URLs with many tags (such as the Google portal page), we only considered the top-10 tags per URL. Second, since we are interested in modeling user interests based on a long-term history, we ignored all users that had less than 28 days of history. The resulting dataset consisted of 71 users and 8095 tags representing these users' interests.

Classification Results

The distribution of the users' interests is exposed in Figure 4.7. The size of the bars represents the number of keywords assigned to a group; the line indicates the number of page accesses related to the keywords in this group.

A first observation is that the majority of keywords (around 55%) is used only once during a single day (C1). However, the page accesses related to these keywords covers less than 5% of the users' access logs. By contrast, less than 1% of the users'

keywords is used on a daily or very constant basis (C2 and C3), but these groups cover with 28% a large portion of the users' accesses. In other words, C2 and C3 represent the head of the power law distribution of keyword usage, C1 the very end of the tail.

Groups C4, C5 and C6 represent the middle part of distribution, covering 40% of the users' interests and 65% of the page accesses. These groups are associated with the users' repetitive and sporadic interests. Group C4 confirms the existence and importance of routine weekly interests. Still, most page visits concern group C6 - the sporadic peaks of interest that last more than one day. This implies that irregularly returning tasks - such as online shopping, travel booking and background research - result in far more revisits than daily or constant interests (C2, C3) and that these returning tasks typically last longer than one day.

Finally, the 'unidentified' cluster C7 (one-time interests that cover several days) contains very few keywords and page accesses. We assume that this indicates that the lifetime one-time interests is typically short (C1); if the interest remains longer than one day (C7), most likely it will happen again (C4, C5 or C6).

4.10 Chapter Summary

In this chapter, we introduced a generic framework for contextual prediction of revisits, evaluating its results experimentaly and with user studies. Additionally, we provided a sense making classification of Web users' interests based on their context.

Here, context is understood as the 'surrounding' Web pages accessed by a user in a specific time.

This information pertaining in these types of contexts is of utmost importance to understand the users' current tasks. As we demonstrated in this chapter, by choosing the appropriate methods to exploit this information, much of the users behavior can be perceived and thus predicted. Our proposed framework builds on top of this principle.

The framework consists of two tiers of methods: ranking methods, which rank resources based on the recency and/or frequency of access to this resource, and propagation methods, which detect items that are commonly visited together with the currently visited resource.

Experimental evaluation shows that combining ranking methods with propagation ones drastically improves performance. In a second experiment, we found that site prediction is simpler than page prediction, and that the performance of a prediction strategy mainly depends on variance in the users' online behavior (in particular, the page and site entropy). The best-performing prediction strategy has been put into practice in the context of a dynamic browser toolbar, the PivotBar. Two user studies with the PivotBar confirm that users appreciate and use the contextual recommen-

dations provided by the toolbar. In addition, the log data shows that a significant amount of revisits has taken place via the PivotBar.

Both the contextual prediction framework, SUPRA, and the dataset that we used in the second experiment, the Web History Repository, have been made available to the community for further experimentation.

Additionaly, in order to make sense of the users' context, we analyzed patterns of returning user interests, making use of a *virtual folksonomy*, composed of the client-side web logs enriched with social bookmarking tags. Using clustering techniques, we identified seven canonical patterns. We developed a rule-based heuristic classifier and evaluated the results of the classification. The results indicate that the greater part of user interests involves tasks that turn up on a more or less regular basis and typically involve longer-lasting activities. If an interest remains longer than one day, it is likely to return at a later stage.

The dominance of the middle part of the power law distribution of keywords is yet another plea for reducing the dominance of the most frequent items and focus on the (start of the) tail instead. In the context of Web browsing, this middle part is mainly formed by interests that return on a more or less regular basis. These patterns of temporal variation can be exploited to better relate keywords, tags or other items in a user profile. Many applications can be thought of in the context of personalization and recommender systems, such as repetitive-interest based collaborative filtering or product recommendations.

5 Conclusions and Outlook

The Web has evolved to a point where most the aspects of our lives are somehow represented or stored online: our letters, documents and pictures, our work, studies and our pleasures, our social connections and our money. More importantly, the Web became part of people's lives. It is very unlikely that new generations and the ones to come will not have an online digital identity.

In the Web, as well as in real life, we naturally separate our tasks and goals. In time, we learn where to go (which pages to access) to read our letters (mailbox), to see how a friend is doing (social networks), to buy groceries (online shopping) or to check our bank balance (online banking). As these separations occur, we implicitly create digital contexts (Recall Definition 1: 'The circumstances that form the setting for an event, statement, or idea...') in terms of spatial attributes, interactions and time.

However, many digital contexts are yet to be exploited. In Chapter 2, we demonstrated how it is possible to exploit spatial contexts on digital resources. Refinding information is significantly improved when spatial context information is attached to digital annotations. We demonstrated this by providing a thorough understanding of annotations in paper-based and webbased scenarios, and by developing a solution to support in-context Web annotations. The same has been verified in learning scenarios where we developed a solution to support in-context annotations in an educational setup (Chapter 2.7). In summary, the chapter demonstrates our proposals to contextualized annotation systems that improve the user experience for the tasks of refinding and also sharing information.

The implications of the findings in this chapter can be directly applied in several existing browsing supporting tools and learning oriented tools, especially at the interface level. Interfaces that allow users to define their own spatial screen context, improves their performance in refinding tasks to a great extent.

Moving on to a broader understanding of context - a piece of information extracted from users' interaction - in Chapter 3, we demonstrated that exploiting contextualized

profiles in fact improves item recommendation tasks. There, we proposed a model that incorporates context in folsonomies. The idea is built upon the concept of an extended folksonomy that includes additional facets (i.e. contexts) regarding a particular system. On top of that, we proposed strategies that exploit the contextualized folksonomies in order to improve profiling of users and resources.

The implication of this chapter is that context information not only support users in their tasks, but additionally provides information for backend prediction methods and recommender systems. In our case, our extended proposed model might be applied to any existing system to improve their recommendation results, consequently improving user experience and satisfaction.

Finally, in Chapter 4, we explore yet another understanding of context, which regards the surrounding events of a web page access. We proposed an efficient contextual revisit prediction method that encompasses ranking and propagation methods. Our focus on revisits is supported by numerous previous works that have exposed the high recurrence rates of page visits. In fact, based on our own data, we discovered that recurrence of domain visits can be as frequent as 92%. In the chapter, we also proposed a categorization of users' interest on the web regarding recurrence that encompasses seven distinct behaviors. In addition, the chapter is complemented with some tangible contributions: the dataset collected with the Web History Repository (Chapter 4.6.2) and the PivotBar (Chapter 4.5).

The implications of the research presented in Chapter 4 are numerous for improving user browsing experience. For example, server-side and client-side caching algorithms can build upon our findings to improve loading time. Additionally, server-side or client-side tools can be further proposed to support users during their daily Web activities. The present PivotBar is our proof of concept of these implications.

In addition to the several aforementioned contributions and implications of this thesis, we hope the reader has also gained a new perspective on the understanding of the so called 'context' and the important role it plays for several applications. Understood as *'spatial attribute'*, *'an additional piece of information'* or the *'surrounding tasks'*, context has demonstrated to be of high effective affix in memory adding and prediction tasks. Throughout this work, we exposed several ways to gather user context by explicitly adding interface features (Chapter 2) or by modeling implicit information (Chapter 3). In all cases, the information conveyed by in the contexts contributed and improved the user Web experience by helping then to refind information, providing better recommendations and finally, supporting them in browsing activities.

5.1 Future Directions

There are several possible continuations of the work presented in this thesis. For the work presented in Chapter 2, we identified that there is a lack of annotation standard

in the Web. We have shown the benefits of annotations and the functionalities that tools and standards should have. Annotations are extremely important for personal impressions and for collaboration, however, there is not yet final understanding of its format (W3C Open Annotation Data Model [1]) nor has it been actually put in practice.

Another direction for future work is regarding the contextual profiles. In this thesis, we explored the dimension of resource profiles. However, our methodology might be modeled to generate a profile of any of the actors pertained in a folksonomy infrastructure. In the work presented, we based our experiments on top of contextualized resource profiles. We believe that, having the user as the pivot actor and applying same contextualization strategies, user profiles shall be improved.

Additionally, we see room for improving the next page prediction task. With disclosed information regarding the users' past action (e.g. queries, page titles, tags, keywords), it is possible to develop prediction methods that achieve higher performances (in this thesis we had only anonymized data). Further, experimentation on the balance between recommendations for pages and sites may lead to better heuristics. Furthermore, a large scale implemantion and experimentation may provide enough data to collaborative contextualization strategies where contexts can be learnt from one user to another user.

Finally, we envisage that context will gradually be introduced and exploited in our every day software in order to improve our experience. While some online service providers in the market are already aware of the advantages of exploiting context information, clearly some others are totally unaware this fact. Still nowadays, several applications and websites do not remember a single user preference. For those software developers willing to provide competitive solutions; implementing their own ways of capturing, understanding and exploiting user context will play a paramount role in their own accomplishment.

[1] http://www.openannotation.org/spec/core/

Bibliography

[ABB+10] Fabian Abel, Matteo Baldoni, Cristina Baroglio, Nicola Henze, Ricardo Kawase, Daniel Krause, and Viviana Patti. Leveraging search and content exploration by exploiting context in folksonomy systems. *The New Review of Hypermedia and Multimedia*, 16(1&2):33–70, 2010.

[ABK+07] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A Nucleus for a Web of Open Data. In Aberer et al., editor, *The Semantic Web, 6th International Semantic Web Conference (ISWC), 2nd Asian Semantic Web Conference (ASWC)*, pages 715–728, November 2007.

[AD72] M. J. Adler and C. Van Doren. *How to Read a Book*. Revised edition, Simon and Schuster, New York, 1972.

[AGH+98] Annette Adler, Anuj Gujar, Beverly L. Harrison, Kenton O'Hara, and Abigail Sellen. A diary study of work-related reading: design implications for digital reading devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '98, pages 241–248, New York, NY, USA, 1998. ACM Press/Addison-Wesley Publishing Co.

[AHH+12] Fabian Abel, Eelco Herder, Geert-Jan Houben, Nicola Henze, and Daniel Krause. Cross-system user modeling and personalization on the social web. *User Modeling and User-Adapted Interaction*, pages 1–41, 2012.

[AHK08] Fabian Abel, Nicola Henze, and Daniel Krause. Exploiting additional Context for Graph-based Tag Recommendations in Folksonomy Systems. In *WI 2008*, pages 148–154. IEEE, 2008.

[AHK+10] Fabian Abel, Nicola Henze, Ricardo Kawase, Daniel Krause, and Patrick Siehndel. Tagme!: Enhancing social tagging with spatial context. In *WEBIST (Selected Papers)*, pages 114–128, 2010.

[AHKK10] Fabian Abel, Nicola Henze, Ricardo Kawase, and Daniel Krause. The impact of multifaceted tagging on learning tag relations and search. In *ESWC (2)*, pages 90–105, 2010.

[AIS93] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, SIGMOD '93, pages 207–216, New York, NY, USA, 1993. ACM.

[AKK10a] Fabian Abel, Ricardo Kawase, and Daniel Krause. Leveraging multi-faceted tagging to improve search in folksonomy systems. In *HT*, pages 299–300, 2010.

[AKK+10b] Fabian Abel, Ricardo Kawase, Daniel Krause, Patrick Siehndel, and Nicole Ullmann. The art of multi-faceted tagging - interweaving spatial annotations, categories, meaningful uris and tags. In *WEBIST (2)*, pages 21–28, 2010.

[AKKS09] Fabian Abel, Ricardo Kawase, Daniel Krause, and Patrick Siehndel. Multi-faceted Tagging in TagMe! In *8th International Semantic Web Conference (ISWC2009)*, October 2009.

[AKT08] Mamoun Awad, Latifur Khan, and Bhavani Thuraisingham. Predicting www surfing using multiple evidence combination. *The VLDB Journal*, 17(3):401–417, May 2008.

[Alt88] Gabriel Altmann. Verteilungen der satzlngen (distribution of sentence lengths). *K.-P. Schulz, editor, Glottometrika 9. Brockmeyer*, 1988.

[AS95] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, ICDE '95, pages 3–14, Washington, DC, USA, 1995. IEEE Computer Society.

[AT01] Gediminas Adomavicius and Alexander Tuzhilin. Using data mining methods to build customer profiles. *Computer*, 34(2):74–82, February 2001.

[ATD08] Eytan Adar, Jaime Teevan, and Susan T. Dumais. Large scale analysis of web revisitation patterns. In *CHI*, pages 1197–1206, 2008.

[AZN99] David W. Albrecht, Ingrid Zukerman, and Ann E. Nicholson. Presending documents on the www: A comparative study. In *IJCAI*, pages 1274–1279, 1999.

[BFNP08a] Kerstin Bischoff, Claudiu S. Firan, Wolfgang Nejdl, and Raluca Paiu. Can all tags be used for search? In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 193–202, New York, NY, USA, 2008. ACM.

[BFNP08b] Kerstin Bischoff, Claudiu S. Firan, Wolfgang Nejdl, and Raluca Paiu. Can all tags be used for search? In *CIKM*, pages 193–202. ACM, 2008.

[BL07] Tim Berners-Lee. Linked Data - design issues. Technical report, W3C, May 2007. http://www.w3.org/DesignIssues/LinkedData.html.

[Bou99] Niels Olof Bouvin. Unifying strategies for web augmentation. In *Proceedings of the tenth ACM Conference on Hypertext and hypermedia : returning to our diverse roots: returning to our diverse roots*, HYPERTEXT '99, pages 91–100, New York, NY, USA, 1999. ACM.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107–117, April 1998.

[Bru01] Peter Brusilovsky. Adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 11(1-2):87–110, March 2001.

[Bus45] Vannevar Bush. As we may think. *The Atlantic Monthly*, 176(1):101–108, 1945.

[BVKKS08] Michael Bernstein, Max Van Kleek, David Karger, and M. C. Schraefel. Information scraps: How and why information eludes our personal information management tools. *ACM Trans. Inf. Syst.*, 26(4):24:1–24:46, October 2008.

[CCC+08] Francesca Carmagnola, Federica Cena, Luca Console, Omar Cortassa, Cristina Gena, Anna Goy, Ilaria Torre, Andrea Toso, and Fabiana Vernero. Tag-based user modeling for social multi-device adaptive guides. *User Model. User-Adapt. Interact.*, 18(5):497–538, 2008.

[CGO08] Elizabeth Churchill, Elizabeth S. Goodman, and Joseph O'Sullivan. Mapchat: conversing in place. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '08, pages 3165–3170, New York, NY, USA, 2008. ACM.

[CKT10] Flavio Chierichetti, Ravi Kumar, and Andrew Tomkins. Stochastic models for tabbed browsing. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 241–250, New York, NY, USA, 2010. ACM.

[CL10] Yi Cai and Qing Li. Personalized search by tag-based user profile and resource profile in collaborative tagging systems. In *CIKM*, pages 969–978, 2010.

[CM01] Andy Cockburn and Bruce Mckenzie. What do web users do? an empirical analysis of web use. *Int. J. Hum.-Comput. Stud.*, 54(6):903–922, June 2001.

[CP95] Lara D. Catledge and James E. Pitkow. Characterizing browsing strategies in the world-wide web. *Comput. Netw. ISDN Syst.*, 27(6):1065–1073, April 1995.

[CTB+00] Elizabeth F. Churchill, Jonathan Trevor, Sara Bly, Les Nelson, and Davor Cubranic. Anchored conversations: chatting in the context of a document. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, CHI '00, pages 454–461, New York, NY, USA, 2000. ACM.

[CTBN00] Elizabeth Churchill, Jonathan Trevor, Sara Bly, and Les Nelson. Stickychats: remote conversations over digital documents. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, CSCW '00, pages 350–, New York, NY, USA, 2000. ACM.

[DAFK+11] Ernesto Diaz-Aviles, Marco Fisichella, Ricardo Kawase, Wolfgang Nejdl, and Avaré Stewart. Unsupervised auto-tagging for learning object enrichment. In *EC-TEL*, pages 83–96, 2011.

[DAK12] Ernesto Diaz-Aviles and Ricardo Kawase. Exploiting twitter as a social channel for human computation. In *CrowdSearch*, pages 15–19, 2012.

[Dil92] Andrew Dillon. Reading from paper versus screens: a critical review of the empirical literature. *Ergonomics*, 35(10), 1992.

[Dil04] Andrew Dillon. *Designing Usable Electronic Text, Second Edition*. Taylor & Francis, Inc., Bristol, PA, USA, 2004.

[DK04] Mukund Deshpande and George Karypis. Selective markov models for predicting web page accesses. *ACM Trans. Internet Techn.*, 4(2):163–184, 2004.

[ESRR04] Maged El-Sayed, Carolina Ruiz, and Elke A. Rundensteiner. Fs-miner: efficient and incremental mining of frequent sequence patterns in web logs. In *Proceedings of the 6th annual ACM international workshop on Web information and data management*, WIDM '04, pages 128–135, New York, NY, USA, 2004. ACM.

[FB05] Rosta Farzan and Peter Brusilovsky. Social navigation support through annotation-based group modeling. In *In Proceedings of 10th International User Modeling Conference*, pages 463–472, 2005.

[FB08] Rosta Farzan and Peter Brusilovsky. Annotated: A social navigation and annotation service for web-based educational resources. *The New Review of Hypermedia and Multimedia*, 14(1):3–32, 2008.

[FBH00] Xiaobin Fu, Jay Budzik, and Kristian J. Hammond. Mining navigation history for recommendation. In *IUI*, pages 106–112, 2000.

[FCKM12] Marco Fisichella, Juri Luca De Coi, Ricardo Kawase, and Maristella Matera. User profile based activities in flexible processes. In *WIMS*, page 33, 2012.

[FCMS05] Xin Fu, Tom Ciszek, Gary Marchionini, and Paul Solomon. Annotating the Web: An Exploratory Study of Web Users' Needs for Personal Annotation Tools. *Proceedings of the American Society for Information Science and Technology*, 42(1), 2005.

[FNP07] Claudiu S. Firan, Wolfgang Nejdl, and Raluca Paiu. The benefit of using tag-based profiles. In Virgílio A. F. Almeida and Ricardo A. Baeza-Yates, editors, *LA-WEB*, pages 32–41. IEEE Computer Society, 2007.

[FRBP08] Farzan, Rosta, Brusilovsky, and Peter. AnnotatEd: A social navigation and annotation service for web-based educational resources. *New Rev. Hypermedia Multimedia*, 14(1):3–32, January 2008.

[GG99] Carl Gutwin and Saul Greenberg. The effects of workspace awareness support on the usability of real-time distributed groupware. *ACM Trans. Comput.-Hum. Interact.*, 6(3):243–281, September 1999.

[GH03] Mathias Géry and Hatem Haddad. Evaluation of web usage mining approaches for user's next request prediction. In *Proceedings of the 5th ACM international workshop on Web information and data management*, WIDM '03, pages 74–81, New York, NY, USA, 2003. ACM.

[GH06] Scott Golder and Bernardo A. Huberman. The structure of collaborative tagging systems. *Journal of Information Sciences*, 32(2):198–208, April 2006.

[GMP07] José A. Gámez, Juan L. Mateo, and José M. Puerta. Improving revisitation browsers capability by using a dynamic bookmarks personal toolbar. In *Proceedings of the 8th international conference on Web information systems engineering*, WISE'07, pages 643–652, Berlin, Heidelberg, 2007. Springer-Verlag.

[GST01] Wolfgang Gaul and Lars Schmidt-Thieme. Mining generalized association rules for sequential and path data. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, ICDM '01, pages 593–596, Washington, DC, USA, 2001. IEEE Computer Society.

[Gui87] Yves Guiard. Asymmetric division of labor in human skilled bimanual action: The kinematic chain as a model. *Journal of Motor Behavior*, 19(4):486–517, 1987.

[Has96] C. Hass. *Writing technology: Studies on the materiality of literacy.* Erlbaum, Manwah, N.J., 1996.

[HCBG01] David Hawking, Nick Craswell, Peter Bailey, and Kathleen Griffihs. Measuring search engine quality. *Inf. Retr.*, 4(1):33–59, April 2001.

[HJSS06a] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. BibSonomy: A Social Bookmark and Publication Sharing System. In *Proc. First Conceptual Structures Tool Interoperability Workshop*, pages 87–102, Aalborg, 2006.

[HJSS06b] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In *Proc. of the 3rd European Semantic Web Conference*, volume 4011 of *LNCS*, pages 411–426, Budva, Montenegro, June 2006. Springer.

[HRGM08] Paul Heymann, Daniel Ramage, and Hector Garcia-Molina. Social tag prediction. In *SIGIR*, pages 531–538, 2008.

[HRS07] Harry Halpin, Valentin Robu, and Hana Shepherd. The Complex Dynamics of Collaborative Tagging. In *Proc. of 16th Int. World Wide Web Conference (WWW '07)*, pages 211–220, New York, NY, USA, 2007. ACM Press.

[Kaz09] Przemyslaw Kazienko. Mining indirect association rules for web recommendation. *Applied Mathematics and Computer Science*, 19(1):165–186, 2009.

[KFN+13a] Ricardo Kawase, Marco Fisichella, Katja Niemann, Vassilis Pitsilis, Aristides Vidalis, Philipp Holtkamp, and Bernardo Pereira Nunes. Openscout: harvesting business and management learning objects from the web of data. In *WWW (Companion Volume)*, pages 445–450, 2013.

[KFN+13b] Ricardo Kawase, Marco Fisichella, Bernardo Pereira Nunes, Kyung-Hun Ha, and Markus Bick. Automatic classification of documents in cold-start scenarios. In *WIMS*, page 19, 2013.

[KHN09] Ricardo Kawase, Eelco Herder, and Wolfgang Nejdl. A comparison of paper-based and online annotations in the workplace. In *EC-TEL*, pages 240–253, 2009.

[KHN10] Ricardo Kawase, Eelco Herder, and Wolfgang Nejdl. Annotations and hypertrails with spreadcrumbs - an easy way to annotate, refind and share. In *WEBIST (2)*, pages 5–12, 2010.

[KHPN10] Ricardo Kawase, Eelco Herder, George Papadakis, and Wolfgang Nejdl. In-context annotations for refinding and sharing. In *WEBIST (Selected Papers)*, pages 85–100, 2010.

[KMN+09] Ricardo Kawase, Enrico Minack, Wolfgang Nejdl, Samur Araújo, and Daniel Schwabe. Incremental end-user query construction for the semantic desktop. In *WEBIST*, pages 270–275, 2009.

[KN09] Ricardo Kawase and Wolfgang Nejdl. A straightforward approach for online annotations: Spreadcrumbs - enhancing and simplifying online collaboration. In *WEBIST*, pages 407–410, 2009.

[KNH+13] Ricardo Kawase, Bernardo Pereira Nunes, Eelco Herder, Wolfgang Nejdl, and Marco Antonio Casanova. Who wants to get fired? In *WebSci*, pages 191–194, 2013.

[KNS13] Ricardo Kawase, Bernardo Pereira Nunes, and Patrick Siehndel. Content-based movie recommendation within learning contexts. In *ICALT*, pages 171–173, 2013.

[KPA11] Ricardo Kawase, George Papadakis, and Fabian Abel. Generating resource profiles by exploiting the context of social annotations. In *International Semantic Web Conference (1)*, pages 289–304, 2011.

[KPH10] Ricardo Kawase, Georgios Papadakis, and Eelco Herder. How predictable are you? a comparison of prediction algorithms for web page revisitation. In *Proc. of 18th International Workshop on Personalization and Recommendation on the Web and Beyond (ABIS), October 2010, Kassel, Germany.*, 2010.

[KPH11] Ricardo Kawase, George Papadakis, and Eelco Herder. Supporting revisitation with contextual suggestions. In *JCDL*, pages 227–230, 2011.

[KPHN10] Ricardo Kawase, George Papadakis, Eelco Herder, and Wolfgang Nejdl. The impact of bookmarks and annotations on refinding information. In *HT*, pages 29–34, 2010.

[KPHN11] Ricardo Kawase, George Papadakis, Eelco Herder, and Wolfgang Nejdl. Beyond the usual suspects: context-aware revisitation support. In *HT*, pages 27–36, 2011.

[KSH14] Ricardo Kawase, Patrick Siehndel, and Eelco Herder. Haters gonna hate: job-related offenses in twitter. In *WWW (Companion Volume)*, pages 557–558, 2014.

[KSHN12] Ricardo Kawase, Patrick Siehndel, Eelco Herder, and Wolfgang Nejdl. Hyperlink of men. In *LA-WEB*, pages 9–16, 2012.

[KSN+12] Ricardo Kawase, Patrick Siehndel, Bernardo Pereira Nunes, Marco Fisichella, and Wolfgang Nejdl. Towards automatic competence assignment of learning objects. In *EC-TEL*, pages 401–406, 2012.

[KSNF13] Ricardo Kawase, Patrick Siehndel, Bernardo Pereira Nunes, and Marco Fisichella. Automatic competence leveling of learning objects. In *ICALT*, pages 149–153, 2013.

[KT10] Ravi Kumar and Andrew Tomkins. A characterization of online browsing behavior. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 561–570, New York, NY, USA, 2010. ACM.

[LFK88] Mary D. P. Leland, Robert S. Fish, and Robert E. Kraut. Collaborative document production using quilt. In *Proceedings of the 1988 ACM conference on Computer-supported cooperative work*, CSCW '88, pages 206–215, New York, NY, USA, 1988. ACM.

[Mac05] W. John Macmullen. Annotation as process, thing, and knowledge: Multi-domain studies of structured data annotation. Technical report, in ASIST Annual Meeting, (Charlotte, NC, 2005), ASIST, in review, 2005.

[Mar97] Catherine C. Marshall. Annotation: from paper books to the digital library. In *Proceedings of the second ACM international conference on Digital libraries*, DL '97, pages 131–140, New York, NY, USA, 1997. ACM.

[Mar98] Catherine C. Marshall. Toward an ecology of hypertext annotation. In *Proceedings of the ninth ACM conference on Hypertext and hypermedia : links, objects, time and space—structure in hypermedia systems:*

links, objects, time and space—structure in hypermedia systems, HYPERTEXT '98, pages 40–49, New York, NY, USA, 1998. ACM.

[May09] Matthias Mayer. Web history tools and revisitation support: A survey of existing approaches and directions. *Found. Trends Hum.-Comput. Interact.*, 2(3):173–278, March 2009.

[MC07] Elke Michlmayr and Steve Cayzer. Learning user profiles from tagging data and leveraging them for personal(ized) information access. In *Proc. of the Workshop on Tagging and Metadata for Social Information Organization, 16th Int. World Wide Web Conference (WWW '07)*, May 2007.

[MCS00] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. Automatic personalization based on web usage mining. *Communications of the ACM*, 43(8):142–151, 2000.

[MFJR+04] Natasa Milic-Frayling, Rachel Jones, Kerry Rodden, Gavin Smyth, Alan Blackwell, and Ralph Sommerer. Smartback: supporting users in back navigation. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 63–71, New York, NY, USA, 2004. ACM.

[Mik05] Peter Mika. Ontologies Are Us: A unified model of social networks and semantics. In *Proc. Int. Semantic Web Conference (ISWC 2005)*, pages 522–536, November 2005.

[Mil08] Vuk Milicic. Case study: Semantic tags. W3C Semantic Web Case Studies and Use Cases, December 2008. http://www.w3.org/2001/sw/sweo/public/UseCases/Faviki/.

[MMV08] Dan Morris, Meredith Ringel Morris, and Gina Venolia. Searchbar: a search-centric web history for task resumption and information refinding. In Mary Czerwinski, Arnold M. Lund, and Desney S. Tan, editors, *CHI*, pages 1207–1216. ACM, 2008.

[MNBD06a] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. HT06, tagging paper, taxonomy, flickr, academic article, to read. In *Proc. of the 17th Conf. on Hypertext and Hypermedia*, pages 31–40. ACM Press, 2006.

[MNBD06b] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead. In *Collaborative Web Tagging Workshop at WWW '06*, May 2006.

[MR06] David E. Millard and Martin Ross. Web 2.0: hypertext by any other name? In *Proceedings of the seventeenth conference on Hypertext and*

hypermedia, HYPERTEXT '06, pages 27–30, New York, NY, USA, 2006. ACM.

[MTR+07] Andrea Marchetti, Maurizio Tesconi, Francesco Ronzano, Marco Rosella, and Salvatore Minutoli. SemKey: A Semantic Collaborative Tagging System. In *Workshop on Tagging and Metadata for Social Information Organization at WWW '07, May 8-12, 2007, Banff, Canada.*, May 2007.

[MW87] Carol Bergfeld Mills and Linda J. Weldon. Reading text from computer screens. *ACM Comput. Surv.*, 19(4):329–357, December 1987.

[NCCK12a] Bernardo Pereira Nunes, Alexander Arturo Mera Caraballo, Marco A. Casanova, and Ricardo Kawase. Automatically generating multilingual, semantically enhanced, descriptions of digital audio and video objects on the web. In *KES*, pages 575–584, 2012.

[NCCK12b] Bernardo Pereira Nunes, Alexander Arturo Mera Caraballo, Marco A. Casanova, and Ricardo Kawase. Boosting retrieval of digital spoken content. In *KES (Selected Papers)*, pages 153–162, 2012.

[NDC+13] Bernardo Pereira Nunes, Stefan Dietze, Marco Antonio Casanova, Ricardo Kawase, Besnik Fetahu, and Wolfgang Nejdl. Combining a co-occurrence-based and a semantic measure for entity linking. In *ESWC*, pages 548–562, 2013.

[NKD+12] Bernardo Pereira Nunes, Ricardo Kawase, Stefan Dietze, Gilda Helena Bernardino de Campos, and Wolfgang Nejdl. Annotation tool for enhancing e-learning courses. In *ICWL*, pages 51–60, 2012.

[NKKN14] Kaweh Djafari Naini, Ricardo Kawase, Nattiya Kanhabua, and Claudia Niederée. Characterizing high-impact features for content retention in social web applications. In *WWW (Companion Volume)*, pages 559–560, 2014.

[NKS+13] Bernardo Pereira Nunes, Ricardo Kawase, Patrick Siehndel, Marco A. Casanova, and Stefan Dietze. As simple as it gets - a sentence simplifier for different learning levels and contexts. In *ICALT*, pages 128–132, 2013.

[NPK+13] Bernardo Pereira Nunes, Stella Pedrosa, Ricardo Kawase, Mohammad Alrifai, Ivana Marenzi, Stefan Dietze, and Marco Antonio Casanova. Answering confucius: The reason why we complicate. In *EC-TEL*, pages 496–501, 2013.

[NS05] Till Nagel and Ren Sander. Hyperhistory. In Siegfried Reich and Manolis Tzagarakis, editors, *Hypertext*, pages 276–277. ACM, 2005.

[NSC08] Les Nelson, Diana Smetters, and Elizabeth F. Churchill. Keyholes: selective sharing in close collaboration. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '08, pages 2443–2452, New York, NY, USA, 2008. ACM.

[NSK+10] Katja Niemann, Uta Schwertel, Marco Kalz, Alexander Mikroyannidis, Marco Fisichella, Martin Friedrich, Michele Dicerto, Kyung-Hun Ha, Philipp Holtkamp, and Ricardo Kawase. Skill-based scouting of open management content. In *EC-TEL*, pages 632–637, 2010.

[OS97] Kenton O'Hara and Abigail Sellen. A comparison of reading paper and on-line documents. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, CHI '97, pages 335–342, New York, NY, USA, 1997. ACM.

[OWHM07] Hartmut Obendorf, Harald Weinreich, Eelco Herder, and Matthias Mayer. Web page revisitation revisited: implications of a long-term click-stream study of browser usage. In *CHI*, pages 597–606, 2007.

[Pas08] Alexandre Passant. LODr – A Linking Open Data Tagging System. In *First Workshop on Social Data on the Web (SDoW2008)*, volume 405. CEUR Workshop Proceedings, 2008.

[PHW02] Jian Pei, Jiawei Han, and Wei Wang. Mining sequential patterns with constraints in large databases. In *CIKM*, pages 18–25, 2002.

[Pir07] Peter L. T. Pirolli. *Information Foraging Theory: Adaptive Interaction with Information*. Oxford University Press, Inc., New York, NY, USA, 1 edition, 2007.

[PKBGM10] Aditya G. Parameswaran, Georgia Koutrika, Benjamin Bercovitz, and Hector Garcia-Molina. Recsplorer: recommendation algorithms based on precedence mining. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, SIGMOD '10, pages 87–98, New York, NY, USA, 2010. ACM.

[PKH12] George Papadakis, Ricardo Kawase, and Eelco Herder. Client- and server-side revisitation prediction with supra. In *WIMS*, page 14, 2012.

[PKHN11] George Papadakis, Ricardo Kawase, Eelco Herder, and Claudia Niederée. A layered approach to revisitation prediction. In *ICWE*, pages 258–273, 2011.

[PL08] Alexandre Passant and Philippe Laublet. Meaning Of A Tag: A collaborative approach to bridge the gap between tagging and Linked Data. In *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China*, Apr 2008.

[PNN10] George Papadakis, Claudia Niedere, and Wolfgang Nejdl. Decay-based ranking for social application content. In Joaquim Filipe and Jos Cordeiro, editors, *WEBIST (1)*, pages 276–281. INSTICC Press, 2010.

[RFST10] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 811–820, New York, NY, USA, 2010. ACM.

[RK04] Matt Rasmussen and George Karypis. gcluto - an interactive clustering, visualization, and analysis system. Technical report, 2004.

[RTMF08] Bryan C. Russell, Antonio B. Torralba, Kevin P. Murphy, and William T. Freeman. LabelMe: A Database and Web-based tool for Image Annotation. *International Journal of Computer Vision*, 77(1-3):157–173, 2008.

[SH97] Abigail Sellen and Richard Harper. Paper as an analytic resource for the design of new technologies. In *CHI '97: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, New York, NY, USA, 1997. ACM Press.

[SHB05] Guy Shani, David Heckerman, and Ronen I. Brafman. An mdp-based recommender system. *J. Mach. Learn. Res.*, 6:1265–1295, December 2005.

[SK12] Patrick Siehndel and Ricardo Kawase. Twikime! - user profiles that make sense. In *International Semantic Web Conference (Posters & Demos)*, 2012.

[SKHH13] Patrick Siehndel, Ricardo Kawase, Asmelash Teka Hadgu, and Eelco Herder. Finding relevant missing references in learning courses. In *WWW (Companion Volume)*, pages 425–430, 2013.

[SMB07] Jeff J. Sandvig, Bamshad Mobasher, and Robin Burke. Robustness of collaborative recommendation based on association rule mining. In *RecSys*, pages 105–112, 2007.

[SvZ08] B. Sigurbjörnsson and Roelof van Zwol. Flickr tag recommendation based on collective knowledge. In *Proc. of 17th Int. World Wide Web Conference (WWW '08)*, pages 327–336. ACM Press, 2008.

[TG97] Linda Tauscher and Saul Greenberg. How people revisit web pages: empirical findings and implications for the design of history systems. *Int. J. Hum.-Comput. Stud.*, 47(1):97–137, July 1997.

[Ton06] Emma Tonkin. Searching the long tail: Hidden structure in social tagging. volume 17, Austin, Texas, 2006.

[TT10] Sarah K. Tyler and Jaime Teevan. Large scale query log analysis of re-finding. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 191–200, New York, NY, USA, 2010. ACM.

[TW98] Hajime Takano and Terry Winograd. Dynamic bookmarks for the www. In *Proceedings of the ninth ACM conference on Hypertext and hypermedia : links, objects, time and space—structure in hypermedia systems: links, objects, time and space—structure in hypermedia systems*, HYPERTEXT '98, pages 297–298, New York, NY, USA, 1998. ACM.

[UBL+08] Carsten Ullrich, Kerstin Borau, Heng Luo, Xiaohong Tan, Liping Shen, and Ruimin Shen. Why web 2.0 is good for learning and for research: principles and prototypes. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 705–714, New York, NY, USA, 2008. ACM.

[VB99] Fabio Vitali and Michael Bieber. Hypermedia on the web: what will it take? *ACM Comput. Surv.*, 31(4es), December 1999.

[VC78] L. S. Vygotskii and Michael Cole. *Mind in society : the development of higher psychological processes*. Harvard University Press, Cambridge, 1978.

[vSBvV+06] Mark van Setten, Rogier Brussee, Harry van Vliet, Luit Gazendam, Ynze van Houten, and Mettina Veenstra. On the importance of "who tagged what". In *Proceedings of the Workshop on the Social Navigation and Community based Adaptation Technologies at AH 2006*, pages 552–561, Dublin, Ireland, 2006.

[Wal07] Thomas Vander Wal. Folksonomy. *http://vanderwal.net/folksonomy.html*, July 2007.

[WDBG+02] Jim Whitehead, Paul De Bra, Kaj Grønbæk, Deena Larsen, John Leggett, and monica m. c. schraefel. Seven issues, revisited. In *Proceedings of the thirteenth ACM conference on Hypertext and hypermedia*, HYPERTEXT '02, pages 171–171, New York, NY, USA, 2002. ACM.

[WOHM06] Harald Weinreich, Hartmut Obendorf, Eelco Herder, and Matthias Mayer. Off the beaten tracks: exploring three aspects of web navigation. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 133–142, New York, NY, USA, 2006. ACM.

[YL11] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *WSDM*, pages 177–186. ACM, 2011.

[YP02] Hui Yang and Srinivasan Parthasarathy. *On the Use of Temporally Constrained Associations for Web Log Mining*. Springer-Verlag, 2002.

[YSW09] Yao Yao, Lei Shi, and Zhanhong Wang. A markov prediction model based on page hierarchical clustering. *Int. J. Distrib. Sen. Netw.*, 5(1):89–89, January 2009.

[ZAN99] I. Zukerman, D. W. Albrecht, and A. E. Nicholson. Predicting users' requests on the www. In *Proceedings of the seventh international conference on User modeling*, UM '99, pages 275–284, Secaucus, NJ, USA, 1999. Springer-Verlag New York, Inc.

[ZBJM01] Polle T. Zellweger, Niels Olof Bouvin, Henning Jehøj, and Jock D. Mackinlay. Fluid annotations in an open world. In *Proceedings of the 12th ACM conference on Hypertext and Hypermedia*, HYPERTEXT '01, pages 9–18, New York, NY, USA, 2001. ACM.

[ZCM98] Polle T. Zellweger, Bay-Wei Chang, and Jock D. Mackinlay. Fluid links for informed and incremental link transitions. In *Proceedings of the ninth ACM conference on Hypertext and hypermedia : links, objects, time and space—structure in hypermedia systems: links, objects, time and space—structure in hypermedia systems*, HYPERTEXT '98, pages 50–57, New York, NY, USA, 1998. ACM.

[ZRMC00] Polle T. Zellweger, Susan Harkness Regli, Jock D. Mackinlay, and Bay-Wei Chang. The impact of fluid documents on reading and browsing: an observational study. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, CHI '00, pages 249–256, New York, NY, USA, 2000. ACM.