5-2013

Improving Virtual Collaboration: Modeling for Recommendation Systems in a Classroom Wiki Environment

Derrick A. Lam
University of Nebraska-Lincoln, s.dlam@yahoo.com

Lam, Derrick A., "Improving Virtual Collaboration: Modeling for Recommendation Systems in a Classroom Wiki Environment" (2013). *Computer Science and Engineering: Theses, Dissertations, and Student Research*. 55.
http://digitalcommons.unl.edu/computerscidiss/55

IMPROVING VIRTUAL COLLABORATION: MODELING FOR RECOMMENDATION SYSTEMS IN A CLASSROOM WIKI ENVIRONMENT

by

Derrick A. Lam

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Science

Major: Computer Science

Under the Supervision of Professor Leen-Kiat Soh

Lincoln, Nebraska

May, 2013

IMPROVING VIRTUAL COLLABORATION: MODELING FOR RECOMMENDATION SYSTEMS IN A CLASSROOM WIKI ENVIRONMENT

Derrick A. Lam, M.S.

University of Nebraska, 2013

Adviser: Leen-Kiat Soh

Collaboration is of increased importance in today's society, with increased emphasis placed on working jointly with others, whether it is in the classroom, in the lab, in the workplace, or virtually across the world. The wiki is one particular virtual collaboration tool that is gaining particular prominence in recent years, enabling people – either in small project groups or as part of the wiki's entire user base – to socially construct knowledge asynchronously on a wide variety of topics. However, there are few intelligent support tools for wikis available, particularly those providing recommendation-based support to users.

This thesis investigates the topic of user and data modeling for recommendation systems in a wiki environment. In addition to conventional usage data, the proposed model uses new metrics designed for the wiki domain, including: active-passive activity level rating, minimalist-overachiever score, and others. For evaluation, the Biofinity Intelligent Wiki was designed, developed, and deployed to a classroom environment and is used for collaborative writing assignments. Post-hoc analysis on the usage data demonstrates the effects of assignment criteria on student behavior, the value of the new

metrics and their correlation to various student strategies, and the potential for applying the metrics for collaboration-focused recommendation.

This work provides insights and tools that are beneficial to virtual collaboration. For example, the active-passive activity level rating provides a quick overview of a participant’s collaborative activity composition and can be leveraged to alert moderators when participants aren’t meeting expectations. The minimalist-overachiever score strongly correlates to evaluations that participants have received, and with additional tuning, it can be used as an aid in determining performance in future collaborations. The artifacts that a participant has contributed towards are indicative of the participant’s collaborative value in a successful recommendation. These, along with other findings, serve as the foundation for improved virtual collaboration.

Table of Contents

List of Multimedia Objects

Chapter 1: Introduction

Modern society has undergone a paradigm shift. In the past, businesses highly valued the highly-skilled individual, the “rock stars” who could single-handedly carry teams and corporations to greater heights through their ability, talent, knowledge, and insight. But during the time span since then, focus had shifted more towards that of teamwork and cooperation (Limerick and Cunningham 1993). Today’s society now places increased emphasis on working jointly with others, whether it is in the classroom, in the lab, in the workplace, or across the world (Karoly and Panis 2005). These principles of increased collaboration and information sharing are even reflected in web development paradigms, such as the recently-popularized “Web 2.0” coined by O’Reilly (2004). From these developments, it is evident that the technological trend for the near future is to facilitate and support collaboration.

What is collaboration? Mattessich and Monsey (2001) define collaboration as: "…a mutually beneficial and well-defined relationship entered into by two or more organizations to achieve common goals. The relationship includes a commitment to: a definition of mutual relationships and goals; a jointly developed structure and shared responsibility; mutual authority and accountability for success; and sharing of resources and rewards."

At a broad level, collaborations can be considered to be divided into two different types based loosely on medium: traditional collaboration and virtual collaboration. *Traditional collaboration* – or *face-to-face collaboration* – often evokes the image of people gathered in a roundtable discussion, excitedly scribbling on napkins and bouncing ideas off of one another. Of the two, it is by far the older, dating back much further than the technology needed for virtual collaboration. This form of collaboration is typically characterized by relatively close physical proximity between the participants in a face-to-face manner and typically arise offline from specific needs or from casual, unplanned conversation (Kraut et al. 1988). While traditional collaboration could be augmented with long-distance collaboration, e.g. via telephone, it wasn't until relatively recent years that collaborative projects could be carried out primarily, or even completely, without meeting in person.

With the advent of computer and communication technologies, it became possible to carry out long-distance collaboration over the Internet. *Virtual collaboration* over this medium makes use of tools such as audio and video conferencing, e-mail, forums, and instant messaging to communicate. It is typically used by geographically dispersed groups in lieu of physical proximity, though geographically localized groups may and do use the tools to augment traditional collaboration. The new medium offers benefits over traditional collaboration, providing a more-decentralized environment, wider geographical reach, faster dissemination of information, more structure, and asynchronous communication (Warkentin et al. 1997). However, critical weaknesses are introduced as well: communication is not as effortless or high-quality as face-to-face communication, and common barriers to successful collaboration may be magnified (Kraut et al. 1988). Consequently, it may be considerably harder to build trust and rapport among participants in this setting.

1.1 Problem Declaration

The specific problem that we wish to address is the improvement of virtual collaboration for the research and education domains by improving *initiation rate* and *collaboration quality* on wikis. The following paragraphs will describe these two terms in greater detail before delving into the details of our chosen domain.

1.1.1 Initiation Rate

We define *initiation rate* as the likelihood that collaboration participants: 1) make a significant contribution to the collaboration effort, and 2) continue to make contributions throughout the duration of their involvement. While initiation rate can be considered to be similar to "collaboration quantity," we specifically include "significant" to qualify the contributions to distinguish it from "lesser" contributions. That is, we wish to separate "collaboration" from mere "participation," where the former requires a greater degree of commitment and provides further benefit to the group. This distinction is also made by Katz and Martin (1995), who define collaborators as "those who work together on the research project throughout its duration or part of it, or who make frequent or substantial contributions" and exclude "those who make only an occasional or relatively minor contribution to a piece of research."

1.1.2 Collaboration Quality

Extrapolating from the measures of collaboration effectiveness used by Ocker and Yaverbaum (1999) for comparing face-to-face and computer-mediated collaboration in a classroom setting, *collaboration quality* can be defined in terms of participant satisfaction in the collaboration process, participant satisfaction of the end product, and the quality of the end product. Meier et al. (2007) further break down participant satisfaction in the collaboration process into nine dimensions: sustaining mutual understanding, dialogue management, information pooling, reaching consensus, task division, time management, technical coordination, reciprocal interaction, and individual task orientation. Warkentin et al. (1997) also include participant perception of group cohesiveness in their evaluation of collaboration quality. These are typically measured by participant responses to post-collaboration questionnaires and expert evaluation of the product as done by Ocker, Warkentin, and Meier.

1.1.3 Target Domain

The wiki is one particular virtual collaboration tool that is gaining particular prominence in recent years, enabling people – either in small project groups or as part of the wiki's entire user base – to socially construct knowledge asynchronously on a wide variety of topics (Forte and Bruckman 2007). Wikis generally consist of a collection of interlinked pages and include basic functionality such as creating, updating, and deleting pages, managing revisions, and uploading attachments, sometimes providing a medium for discussion. Participation is typically voluntary, and wikis are generally open for any user to edit, although access can be restricted to prevent "vandalism" of pages. We are particularly interested in targeting wiki usage since wikis embody the principles of Vygotsky's social constructivist theory: "the discovery of knowledge stems from its social construction." As stated by Cosley et al. (2007), "social science theory suggests reducing the cost of contribution will increase [users'] motivation to participate." With few intelligent support tools for wikis available, the wiki domain holds many opportunities for improving virtual collaboration.

Scientific research is an area characterized by a high degree of collaborative activity. As phrased by Hara et al. (2003), research involves "large-scale projects dominated by complex problems, rapidly changing technology, dynamic growth of knowledge, and highly specialized expertise.” Additionally, a single person no longer has the time, skills, or knowledge to single-handedly make large contributions outside of a narrow area of research (Hara et al. 2003). This is reflected in the increase of the average number of co-authors per paper during recent decades: Mattessich and Monsey (2001) cite that the average number of co-authors on a paper rose from 3.9 to 8.3 between 1981 and 2001 in the Proceedings of the National Academy of Sciences of the United States of America. This number can only continue to grow as technology improves, enabling improved collaboration on a global scale.

For participants in the scientific research domain, improvements to collaboration initiation rate and quality are both of importance. Regarding initiation rate, researchers are often interested in discovering new, interesting connections and extensions from their current work. As such, they often express great interest to new ideas and collaboration opportunities in both their area of expertise as well as across disciplines. Collaboration quality is also of interest since it could provide reduced costs, improved productivity, and opportunities for discovering new knowledge and exchanging information with experts and peers.

The classroom, i.e. an educational setting, is another area that is characterized by a high degree of collaborative activity. Most, if not all, educational curricula now include collaborative learning opportunities in the form of group projects and papers or some other activities requiring students to learn and solve problems jointly. Recent research shows that in addition to preparing students for the team environment in post-education work, these activities also provide educational benefits including enhanced critical thinking and increased interest and understanding (Gokhale 1995). To further support the trend towards the development of collaboration tools, there is recent interest in *computer-supported collaborative learning* for the educational setting, where the collaborative learning experience is enhanced over an electronic medium (Stahl et al. 2006).

For participants in the educational domain, improvements to collaboration quality are typically more important than initiation rate. Students may not necessarily be concerned with initiation rate since their “participation” in the collaborative learning is often mandated by the requirements of the assignment or activity, and the instructor may assign particular groups for the exercise. However, collaboration quality may be of greater concern to the students. Since their grades are at stake, they are motivated to perform well and seek high-quality collaboration with other similarly-motivated peers. On the other hand, instructors are interested in both improved initiation rate and collaboration quality. In addition to the desire to see the students collaborating frequently with one another, instructors are interested in seeing the students’ learning experiences enriched by the collaborative learning.

1.2 Motivation

Collaboration may be initiated and sought for multiple reasons, including opportunities for reducing cost, improving productivity, discovering new knowledge, and exchanging information with experts and peers within and across multiple disciplines and domains. But on the other hand, poor collaborations often have costly consequences, including inefficient use of resources, decreased productivity, reduced output quality, and dissatisfaction amongst the members involved. In the worst cases, the group may lose members or the project may be canceled altogether.

The use of virtual collaboration to augment cross-disciplinary research is recently emerging as a hot trend, growing in significance as more virtual teams are formed between geographically dispersed participants. It is well-suited for decentralized work where team members can work separately on pieces of the “bigger picture.” Although there may be great physical distances and few dependencies between them, they are still readily accessible through the various tools for both synchronous and asynchronous communication, including audio and video conferencing, instant messaging, e-mail, forums, wikis, etc. The importance of virtual collaboration will grow exponentially as organizations, commercial and non-commercial alike shift more towards teams geographically dispersed around the globe.

While collaboration is the key core of research, the means through which it happens is the dissemination of ideas, data, findings, and the resulting discussion. The internet as a medium enables the faster, more-widespread reach of information. While this exposure can be seen as a benefit in and of itself, the key value is that it enables Vygotsky’s oft-cited social constructivist theory on a larger scale (Anderson et al. 1997). After data is published online, it can then be discovered by other labs, which can augment the data to their own data sets, analyze and compare the data with their own findings, and discuss the results with the originating lab. The process of sharing, discovering, and discussing can lead to “big picture” connections between the seemingly disjoint data sets, further fueling the inspiration for further research.

Although there are concerns about the electronic medium encumbering the communication process, and consequently, the collaboration process, it should be noted that today’s users are more tech- and Internet-savvy than in the past. Herring (2004) states that users – both those growing up in the “Net Generation” and members of the older generations – have gained “extensive familiarity” with computer-mediated communication over the years. Having considerable experience and familiarity with the technology and tools, these users are less hesitant and less reserved about contributing to online discussions, in some cases even preferring it as a primary means of communication. In turn, their apparent comfort on the medium may encourage the more-apprehensive users to participate (Preece and Shneiderman 2009). Leveraging this characteristic will allow for more widespread “buy in” and more effective online collaboration among all parties.

In spite of the differences in medium and specific process details, improving virtual collaboration processes results in the same benefits as improving traditional collaboration: more-effective and more-efficient collaboration. When working in the partnership is motivating, people deliver higher-quality work in less time and at a lower cost, leaving more time and resources for additional pursuits. Other benefits on the interpersonal front are also available: improved networking and goodwill, and potential for further future collaboration. In addition to these, Katz and Martin (1997) also mention the aforementioned sharing and transfer of knowledge, skills, and techniques, potential cross-fertilization of ideas, intellectual companionship, and enhanced visibility are benefits specific to the research domain.

1.3 What is the State of the Art Lacking?

As previously mentioned, wikis have recently gained particular prominence, both as a collaborative tool and as a research topic. The International Symposium on Wikis and Open Collaboration[1] series is one particular avenue dedicated to sharing, discussing, and advancing research and practice in the area. The most popular topics covered in the proceedings are: 1) the evaluation of the effectiveness of using a wiki for web collaboration in a variety of contexts and for supporting traditionally-offline settings (e.g. business or classroom use), 2) the integration of already-researched or trending ideas, such as the inclusion of semantic web concepts to create semantic wikis (e.g. Schaffert 2006) and trust and reputation aspects (e.g. Suh et al. 2008), and 3) the development of new tools to enhance the wiki user experience. However, very few of these solutions are developed to specifically target our goal of improving the wiki collaboration process (e.g. Coslet et al. 2007, Tansey and Stroulia 2010). We thus widen our scope and also include discussions for solutions to improving non-wiki centric collaboration, intelligent interfaces in wikis, and general recommendation algorithms in our investigation of the state of the art. In particular, we place particular emphasis on the user and data models used in each approach and summarize their applicability (or lack thereof) to the wiki setting.

Our work will focus on addressing the holes found in current user and data modeling approaches for the wiki domain. For further detail on the works mentioned in this section, please refer to Chapter 2.

1.3.1 The Annoki Platform

The Annoki platform by Tansey and Stroulia (2010) is a suite of MediaWiki[2] extensions geared towards improving task-based collaboration. The functionality provided over traditional wikis include: 1) namespace-based access control and an easy-

[1] http://www.wikisym.org

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

to-use interface for managing permissions; 2) annotations (i.e. tags and aliases/nicknames) and a simplified template editor; 3) visualizations for overall wiki structure, page content, and user contributions; 4) sentence differencing; and 5) additional features such as calendar extensions and LaTeX export.

One aspect that most pertains to our work is the tracking and the visualization of specific user contributions. In addition to tracking insertions, deletions, and internal and external links made, Annoki uses a *sentence ownership* mechanism to attribute sentences of an article to specific authors. That is, the model of each user can be derived from the actions they perform and their proportion of "ownership" of the entire wiki. Pages can then be visualized in terms of the actions performed by each contributor to the article.

Another aspect that deserves particular mention is the "wiEGO" graphical page structure editor and Annoki-specific page templates used to assist users with structuring ideas and information content to wiki pages. Aside from the apparent benefit of aiding in the translation of information to pages, it also provides a benefit in page modeling by enabling classification via page type. Coupled with Annoki's capability to annotate pages with tags, this enables pages to be modeled based upon structure and associated keywords.

The models leveraged by the Annoki platform, being based in a wiki setting, are completely applicable to our work. However, these models are not leveraged to their fullest extent since the Annoki platform does not provide any intelligent features — for example, their use is limited to displaying information for users to act upon (should they choose to). The helpfulness of this functionality, excepting general usage numbers, was not quantitatively reported. Although Annoki has been deployed and is in use at the time of this writing, no results have been reported on its effectiveness in improving collaboration. We are thus unable to determine how much collaboration initiation rate and quality are improved with this platform.

1.3.2 Socs

Another approach to improving collaboration in this setting is by improving the user interface to highlight information that is not readily accessible in traditional wikis. This is exemplified in the Socs prototype developed by Atzenbeck and Hicks (2008), an application for Mac OS that “serves as a means to express, store, and communicate social information about people.” Socs provides social and group awareness in Wikipedia by 1) providing a visualization of the authors contributing to each wiki page and the social groups to which they belong, 2) linking authors with their other works (i.e. other pages they have contributed towards), 3) enabling the user to flag authors of interest, and 4) integrating with the Apple Address Book. The goal is that this functionality encourages improved communication among wiki page authors, increases understanding of author intentions, and provides “implicit recommendations” of other works by authors flagged as noteworthy by the user.

Since Socs is primarily a social application, its primary modeling occurs on the user side, associating wiki contributors for a particular page with social groups that potentially overlap. A visualization of these group associations can be seen in Figure 2.4. Further, users are also linked to the pages that they’ve contributed towards, enabling easy access to other work performed by particular users.

Unfortunately, the solution relies heavily on manual user action. For instance, users manually flag specific authors as noteworthy and must manually request the system to retrieve other pages that flagged authors have contributed towards. These "implicit recommendations" are neither automated by the system nor ranked by any sort of relevance.

There are currently no evaluation results reported on Socs's performance, so we are unable to determine how well it improves collaboration initiation rate and quality.

1.3.3 Automated Recommendations

Yet another approach to improving collaboration on a wiki is through the use of *automated recommendations*. A wide variety of general recommendation algorithms already exist, and they span a wide variety of targets, such as resources, products, and people that may be of interest to the user. Much research has already been performed in recommending products and information to users for "consumption," with algorithms such as collaborative filtering (e.g. in Konstan et al. 1997 and Linden et al. 2003) and k-nearest neighbors (e.g. Shepitsen et al. 2008). User-to-user recommendations have also been leveraged for locating expertise and procuring help for specific tasks (e.g. Vassileva et al.2003), and research for recommending users to users for social networking purposes (e.g. Chen et al. 2009 and Guy et al. 2009) is also recently gaining traction.

Notable recommender works directed towards a wiki setting (specifically, Wikipedia[3]) have also been developed including expertise location by Demartini (2007) and topic-based recommendation by Sriurai et al. (2009). Demartini's work attempts to locate experts in the Wikipedia user base, and user models are created by processing contributors' revisions to determine their areas of expertise as well as the level of their

[3] http://wikipedia.org

expertise. Sriurai uses topic-based page modeling to recommend Wikipedia pages of interest related to the currently-viewed page.

While each recommender system builds and leverages user models to derive recommendations from, these models are often limited in scope to the bare necessities needed for the algorithm to function. Consequently, currently-popular algorithms may overlook wiki-specific factors (such as frequency and implicit quality of edits) and synergy between factors in separate algorithms when applied to the wiki domain without modifications.

In general, it can be said that these works focus on helping wiki users *locate* expertise or interesting items. While these works can be *leveraged* to improve wiki collaboration, this improvement is not the focus of the works themselves. Thus, they do not directly address our targeted problem of improving collaboration rate and quality.

The SuggestBot developed by Cosley et al. (2007) deserves particular distinction since it is one such recommendation-based tool that suggests wiki pages to contribute towards. However, it is limited by its recommendation scope, i.e. pages that are: 1) not in the top 1% of most frequently-edited articles and 2) are *explicitly* and *manually* flagged by users as stubs, needing improvement, etc. It is also limited by a design that favors ease of implementation over accuracy. The three intelligent algorithms used – based upon text similarity, explicit links, and co-editing patterns –are combined via random selection. Their work reports that 2.5 to 4.3 percent – roughly 30 to 40 out of 1150 – of the recommended pages for each algorithm were followed and edited within two weeks of being presented to the user.

Due to the approach taken, the user and page modeling performed, while applicable to our goal, is relatively simple. As with the previously listed recommendation algorithms, modeling is limited to only include factors relevant to the recommenders used. In the particular case of SuggestBot, user modeling is limited to only user interests. Profiles are represented as a set of article titles and are implicitly determined via the edits to article pages. Edits to non-articles and revision edits are ignored, and each page is counted only once regardless of the number of edits made to it. Page modeling is also limited, only factoring the article title and direct links to and from other pages.

Regarding the problem of initiation rate, i.e. the likelihood of contributing to an article, this tool only offers a modest success rate, though it should be noted that the results reported are not relative to the users' activity levels. The problem of collaboration quality, i.e. the quality of the edits made or of the collaboration between the authoring users, was not addressed by this solution.

1.4 Our Solution

In the thesis, we present our first steps taken towards a solution for making collaboration-centric recommendations in the wiki environment. We first propose a model for users and wiki data that includes the following factors from existing approaches:

- Page tags and keywords
- Page ratings
- Reputation and expertise of authors
- Links between pages (i.e. PageRank)

- Number of page views and edits

We then propose a prototype recommendation algorithm that leverages these factors to recommend wiki pages. Details for both the models and the recommendation algorithm are detailed later in Chapter 3.

Our solution provides contributions on various fronts. First, we designed and implemented a user and data model specific to the wiki domain that leverages factors used across individual, separate approaches not yet combined in this manner. Second, we outline a preliminary recommendation algorithm that leverages these models to suggest pages to the user. Third, we developed an intelligent wiki within the Biofinity Project (Scott et al. 2008) – a software framework that unifies biodiversity and genomic data across multiple, varied sources – to use and gather data for our models. Fourth, our empirical evaluation of the models provides additional data and insights regarding their applicability to actual wiki users.

The rest of the thesis is organized in the following manner. We will first review the current state of the art and related works in Chapter 2 before introducing and detailing our proposed approach in Chapter 3. Chapter 4 then describes our implementation of the Biofinity intelligent wiki and of our approach, and Chapter 5 details our experimental methodology and the results and analysis of the experiments performed, respectively. Finally, Chapter 6 closes out the thesis with the conclusions drawn from the results and possible future directions for the work.

Chapter 2: Related Work and the State of the Art

This chapter delves into further detail on the existing research related to our target problem of improving collaboration in a wiki setting. In particular, we focus our attention on the user modeling, data modeling, and recommendation techniques used in the solutions summarized in Chapter 1.3. We can broadly categorize these solutions into one of three areas:

1) research specifically focusing on the goal of improving wiki collaboration (Chapter 2.1)
2) wiki-related research that can be applied to improving wiki collaboration (Chapter 2.2)
3) popular general recommendation algorithms (Chapter 2.3)

2.1 Improving Collaboration in Wikis

As previously summarized in Chapter 1.3, there are several approaches to solving the problem of improving collaboration in a wiki environment, including: 1) suggesting work for the user to perform (e.g. recommendations and task routing), 2) reducing the cognitive cost of creating and maintaining wiki content (e.g. content management), and 3) facilitating social awareness and communication between contributors (e.g. relationship visualization). We have thus identified three research works, one in each of these approaches, as relevant to our own.

- SuggestBot by Cosley et al. (2007)
- Annoki platform by Tansey and Stroulia (2010)
- Socs prototype by Atzenbeck and Hicks (2008)

They were chosen because they: 1) specifically target the goal of improving collaboration in a wiki setting, and 2) contain some mechanism for users to explore and discover new information (which may or may not encourage collaboration directly).

The SuggestBot for Wikipedia improves wiki collaboration via recommending pages to contribute towards, and it is considered the most relevant to our work of the three since we also take a recommendations-based approach to improving collaboration. It uses a hybrid recommender system (explained later in Chapter 2.3.3) consisting of three intelligent algorithms based on text similarity, links between pages, and co-editing profiles. Consequently, SuggestBot's user and page models only contain factors relevant to generating these recommendations. Due to its targeted deployment to Wikipedia, there are a few design decisions made which are not applicable to its use in a more general wiki. First, one of the intelligent algorithms hinges upon the use of a specific database (i.e. leverages MySQL 4.1's built-in fulltext search). Second, the pool of candidate articles for the algorithms is limited to those manually marked by users as needing work via Wikipedia-specific notation. While this limits the recommendation scope to a more-reasonable size, this is not generalizable to other wikis since they do not follow the same protocols as Wikipedia. Finally, Wikipedia's limited action tracking and community standards for bots limit the SuggestBot to the use of relatively simple algorithms that emphasize performance over accuracy. More details on its recommendation procedure and performance are covered in Chapter 2.1.1.

The Annoki platform built on top of MediaWiki aims to improve task-based collaboration by providing a suite of tools that facilitate information development, management, and visualization. It contains three features that may indirectly promote collaboration: the tag mechanism and corresponding tag cloud visualization, the WikiMap, and the Wiki Contribution Analysis. While the tags and WikiMap facilitate user exploration of the wiki for related elements, this exploration process is entirely manual, and the Annoki implementation of tags offers little over the basic tagging functionality commonly used in Web 2.0 applications (i.e. not automated and without any prioritization of results). The Wiki Contribution Analysis component displays editing and ownership statistics for each article, which could motivate users to periodically review the pages they've contributed towards or encourage others to increase participation. Overall, the platform lacks "intelligent" features that could further benefit collaboration. The mentioned Annoki platform features are described in greater detail in Chapter 2.1.2.

Finally, the Socs prototype attempts to encourage collaboration by improving social and group awareness and facilitating communication in wikis via an application for the Mac OS. Its "social space" and "awareness features" increase the visibility of the contributions of authors that are noteworthy to the user, and integration with the Apple Address Book facilitates contact with them when the need for collaboration arises. As with the Annoki platform, Socs lacks "intelligent" features that could further benefit collaboration. While the application displays all page contributors for a wiki page and highlights the participating acquaintances, it does not actively promote collaboration or prioritize authors to contact. The Socs functionality is covered in greater depth in Chapter 2.1.3.

2.1.1 SuggestBot

As described previously, the SuggestBot developed by Cosley et al. (2007) is a recommendation-based tool that suggests wiki pages to contribute towards. It limits its recommendations to Wikipedia pages *manually marked* by users with the flags in Table 2.1:

Work Type	Description	Count
STUB	Short articles that are missing basic information	355,673
CLEANUP	Articles needing rewriting, formatting, and similar editing	15,370
MERGE	Related articles that may need to be combined	8,553
SOURCE	Articles that need citations to primary sources	7,665
WIKIFY	Articles whose text is not in Wikipedia style	5,954
EXPAND	Articles longer than stubs that still need more information	2,685

Table 2.1: Work types that SuggestBot recommends, along with an approximate count of articles that need each type of work as of May 2006 (Cosley et al. 2007)

These flags constitute the most common types of work needed on the articles. Additionally, the authors exclude pages that are already frequently edited (i.e. in the top 1% of most frequently-edited articles). Jointly, these two limitations narrow the recommendation scope to articles that are *known* to be in need of work. That is, the SuggestBot does not need to algorithmically determine whether a wiki page needs editing.

Three intelligent algorithms were used to generate recommendations: text similarity, links between pages, and co-editing patterns.

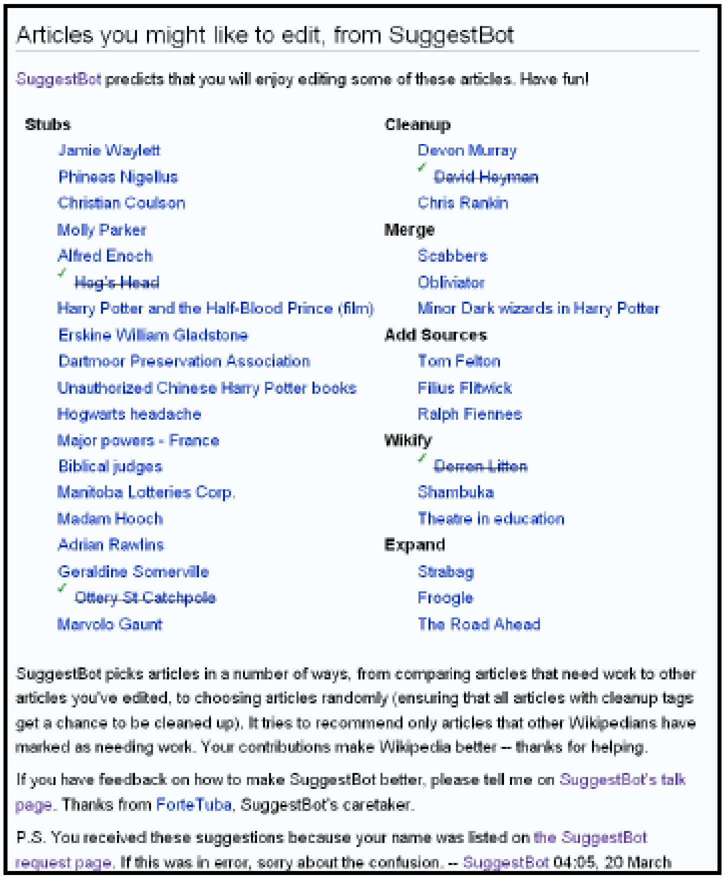
The text similarity-based recommendation operates by: 1) concatenating the titles of articles in the user's editing profile into keywords, and 2) using the keywords in a search against the full text of articles using MySQL 4.1's built-in fulltext search feature[4]. The recommendation set returned is ordered based on the determined relevance from the search algorithm, which uses a modified version of the *term frequency-inverse document frequency* method.

[4] http://dev.mysql.com/doc/refman/4.1/en/fulltext-search.html

The links recommender makes recommendations based on explicit links in articles in the user's editing profile, representing Wikipedia pages as nodes and the links between them as directed edges. The algorithm performs a limited-depth, breadth-first traversal with loops and node revisiting allowed, starting from the articles that the user has edited. Scores are assigned to pages by counting the number of times they have been reached when the algorithm ends, and the recommendation set returned is ordered based on the normalized counts.

The co-edit recommender uses a collaborative filtering (further described in Chapter 2.3.1) variant to recommend pages that authors similar to the user have edited. This version of the algorithm differs from traditional collaborative filtering in a few aspects. First, it uses *editing profiles* rather than *ratings* to calculate similarity between users since Wikipedia does not use a ratings system. Second, an author is considered as a "neighbor" of the user if *any* of the pages in its editing profile is also in the user's profile. Third, the algorithm uses Jaccard similarity instead of the more-common similarity measures, such as cosine similarity and Pearson correlation. The recommendation set returned is then ordered based on the score calculated for each article.

The recommendation set returned to the user consists of 34 article slots: 19 stubs and 3 of each of the remaining five flag types. The articles to place in each slot are chosen in the following manner. First, a recommender is randomly chosen from four approaches: the three intelligent algorithms and random selection. The slot is then filled with the first article that: 1) matches the flag type for the slot, 2)has not already been used in another slot, and 3) is not in the top 1% of frequently edited articles. If the selected engine cannot make a recommendation fulfilling those requirements, another one is randomly chosen. Figure 2.1 illustrates the interface used to display these recommendations to the user.



Articles you might like to edit, from SuggestBot

SuggestBot predicts that you will enjoy editing some of these articles. Have fun!

Stubs

- Jamie Waylett
- Phineas Nigellus
- Christian Coulson
- Molly Parker
- Alfred Enoch
- ✓ ~~Hog's Head~~
- Harry Potter and the Half-Blood Prince (film)
- Erskine William Gladstone
- Dartmoor Preservation Association
- Unauthorized Chinese Harry Potter books
- Hogwarts headache
- Major powers - France
- Biblical judges
- Manitoba Lotteries Corp.
- Madam Hooch
- Adrian Rawlins
- Geraldine Somerville
- ✓ ~~Ottery St Catchpole~~
- Marvolo Gaunt

Cleanup

- Devon Murray
- ✓ ~~David Heyman~~
- Chris Rankin

Merge

- Scabbers
- Obliviator
- Minor Dark wizards in Harry Potter

Add Sources

- Tom Felton
- Filius Flitwick
- Ralph Fiennes

Wikify

- ✓ ~~Derren Litten~~
- Shambuka
- Theatre in education

Expand

- Strabag
- Froogle
- The Road Ahead

SuggestBot picks articles in a number of ways, from comparing articles that need work to other articles you've edited, to choosing articles randomly (ensuring that all articles with cleanup tags get a chance to be cleaned up). It tries to recommend only articles that other Wikipedians have marked as needing work. Your contributions make Wikipedia better -- thanks for helping.

If you have feedback on how to make SuggestBot better, please tell me on SuggestBot's talk page. Thanks from ForteTuba, SuggestBot's caretaker.

P.S. You received these suggestions because your name was listed on the SuggestBot request page. If this was in error, sorry about the confusion. -- SuggestBot 04:05, 20 March

Figure 2.1: Display of SuggestBot recommendations (Cosley et al. 2007)

As briefly mentioned previously, the user and page modeling performed is strongly tied to the recommendation algorithms used. The user model in this particular implementation consists solely of the user's "interests," as implicitly indicated by the user's editing profile. Cosley et al. represent this as the *set* of titles for the articles that the user has edited, ignoring minor revisions (e.g. vandalism reverts), edits to non-article pages, and the number of edits to each page. This user model is leveraged in the text similarity-based and co-edit recommenders. The page model is similarly sparse, containing only the article title, body text, and intra-wiki links to and from the article.

Due to its targeted deployment to Wikipedia, there are a few design decisions made which sacrifice SuggestBot's recommendation accuracy and applicability to a more general wiki for ease of implementation and on-line calculation speed. For instance, the pool of candidate articles for the algorithms is limited to those not in the top 1% of most edited articles and to those manually marked by users as needing work via Wikipedia-specific notation. While this reduces the recommendation scope to a more-reasonable size, it precludes recommendation of potentially "easier" unmarked work that users may be less hesitant to perform, such as tagging articles with the work types in Table 2.1 and providing preliminary stub content for "red links." The Wikipedia community standards for bots also contribute towards SuggestBot's limited accuracy in that they motivate the bot's design focus on simplicity. This focus, coupled with Wikipedia's limited action tracking, leads to the selective exclusion of some tracked features, such as excluding edit counts from users' editing profiles, which in turn may reduce a recommendation's relevance. For this particular example, disregarding edit counts for each article may provide a greater breadth of recommended work but at the tradeoff of decreased relevance for users.

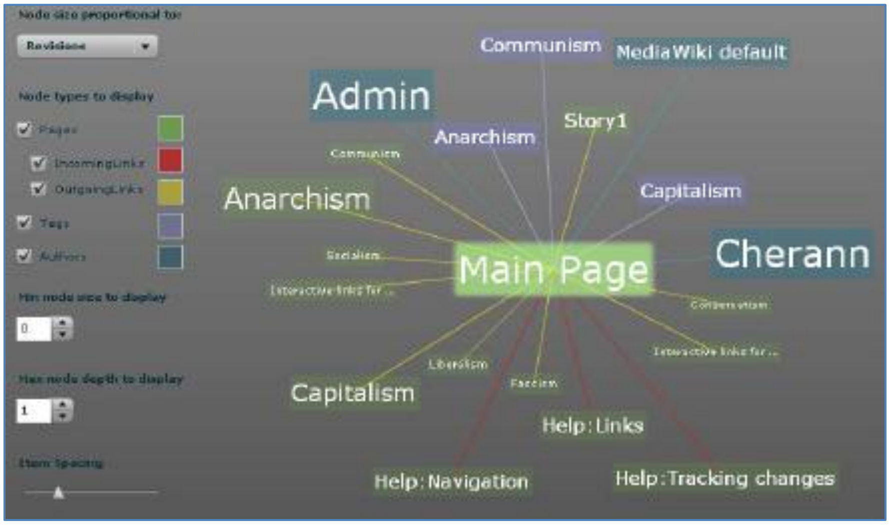
As previously mentioned, the results reported by Cosley et al. focus only on what we consider to be collaboration initiation rate, i.e. the likelihood of contributing to an article. Their work reports that 2.5 to 4.3 percent (roughly 30 to 40 out of 1150) of the recommendation pages for each algorithm were followed and edited within two weeks of being presented to the user – a modest success rate. The problem of collaboration quality, i.e. the quality of the edits made or of the collaboration between the authoring users, was not addressed by this solution.

2.1.2 Annoki

The Annoki platform by Tansey and Stroulia (2010) is a suite of MediaWiki extensions geared towards improving task-based collaboration (i.e. software engineering projects). The functionality that the platform provides over traditional wikis include: 1) namespace-based access control and an easy-to-use interface for managing permissions; 2) annotations (i.e. tags and aliases/nicknames) and a simplified template editor; 3) visualizations for overall wiki structure, page content, and user contributions; and 4) additional features such as calendar extensions and LaTeX export. In short, Annoki strives to improve task-based collaboration primarily by providing productivity-enhancing tools. A few features – tags, WikiMap, and Wiki Contribution Analysis – improve awareness of peer activity and facilitates information discovery. These are of particular interest to our work.

Tags in Annoki are largely implemented in a similar manner to tags in Web 2.0 applications. Users may annotate pages with tags to associate them with particular categories of pages. Each tag has its own wiki page which is automatically populated with links to all wiki pages marked with the same tag. This is akin to Wikipedia’s automatically-generated “category pages” for locating other items sharing the same category. Annoki also features a wiki-level tag cloud which displays all the tags used in the wiki in varying sizes based on frequency of use. While simple, this mechanism enables the discovery of potentially-related pages.

WikiMap (in Figure 2.2) is a tool that visualizes the structure of a wiki, displaying how the elements of the wiki (e.g. pages, users, and tags) are related to the particular centered element via connected nodes. The user may also click on any of the items to navigate to the corresponding page or re-center the map on a new element. The links connecting the center element are color-coded based upon element type, and the size of the node reflects the "importance" of the element. For a user node, this corresponds to the number of edits made; for a tag, this corresponds to the frequency of its use; and for pages, the user can choose between weighting based on the number of revisions, the number of contributing authors, the number of page views, and the number of links to/from the page.



Node size proportional to:
Revisions

Node types to display
☑ Pages 🟩
☑ IncomingLinks 🟥
☑ OutgoingLinks 🟨
☑ Tags 🟪
☑ Authors 🟦

Min node size to display
0

Max node depth to display
1

Item Spacing

Figure 2.2: An example of Annoki'sWikiMap, showing author, page, and category nodes, centered on the page "Main page" (Tansey and Stroulia, 2010)

The Wiki Contribution Analysis visualization tool displays the specific contributions of wiki users for particular articles as shown in Figure 2.3. In addition to displaying statistics on insertions, deletions, and internal and external links made, Annoki introduces the notion of *sentence ownership* to attribute sentences to specific authors, and the number of sentences *owned* in the article is also displayed. Sentence ownership is given to a revision author if: 1) the sentence written is not in a previous revision, or 2) the author changed more than 50% of the words in the sentence.

As of 2010, the Annoki platform was used as both an independent collaboration platform and as base for other systems that require domain-specific features. In the former use case, ten instances of the platform were installed for “various groups.” The heaviest use was seen by the Software Engineering Research Lab (SERL) at the University of Alberta over the course of two years. Table 2.2 lists some usage statistics from SERL and the other nine installations.

System	**Users**	**Pages** (non-redirect)	**Edits**	**Page Views**
SERL	197	2,365	19,828	209,798
Others	218	422	2,272	38,565

Table 2.2: Usage statistics for Annoki installations (Tansey and Stroulia, 2010)

While Annoki does not make use of user and data models for intelligent user interface content, it does have the potential to build such models and apply them in future extensions. User modeling includes wiki actions tracked (e.g. pages viewed and edited) as well as specifics of the edits (e.g. sentence ownership, links added/deleted). Such low-level tracking could be leveraged for user expertise modeling, described later in Chapter 2.2.2. Users’ interests can also be implicitly modeled through their “links” to other wiki content, as mapped by the WikiMap feature. Data models in Annoki include: links to and from other wiki pages and users, as shown in the WikiMap; keywords and associated topics through the tagging functionality; and a particular page “type” based on any templates or wiEGO graphical page structures used to create the page. Coupled with the user model, the page’s quality can be modeled as well.

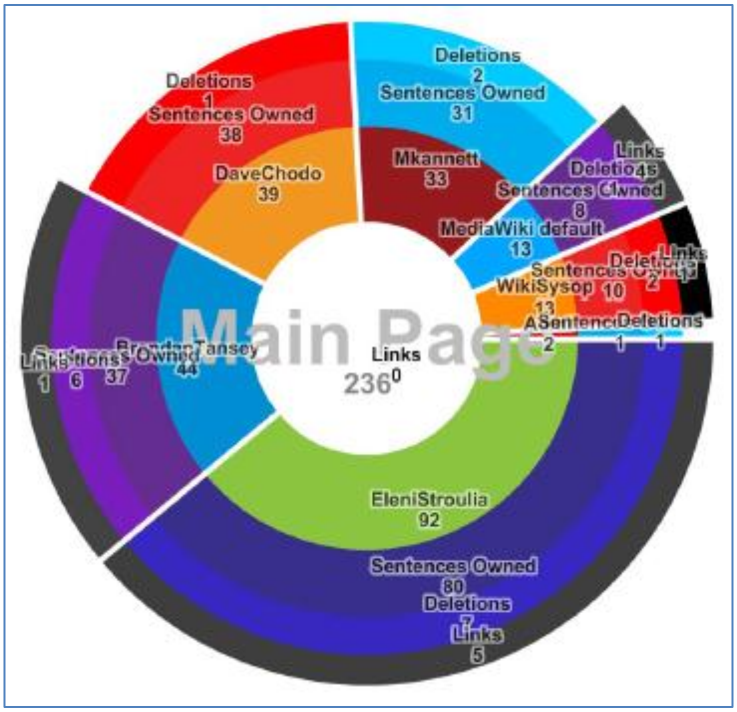


Figure 2.3: Graphical display of wiki page contributions in Annoki (Tansey and Stroulia, 2010)

Unfortunately, the platform currently lacks "intelligent" or automated features that could be leveraged to further improve ease of use and collaboration in the system. It is more a collection of tools that facilitate wiki management and content creation than it is a tool for directly promoting collaboration between its users, and the qualitative results reported reflect this. The usefulness of the namespace-based access control mechanism, the wiEGO visualization, and simplified template creation mechanism in particular were highlighted over the other features. It should be noted, however, that the template feature gave rise to a powerful collaboration tool. By creating and making use of a template for academic papers, SERL was able to circulate interesting or useful papers throughout the group, using the paper's corresponding wiki page to share thoughts and identify potential discussion partners.

Now, the three features specifically covered – the tag mechanism and corresponding tag cloud visualization, the WikiMap, and the Wiki Contribution Analysis – can all indirectly lead to improved collaboration, but require user initiative and motivation to do so. In the case of tags and the WikiMap, user exploration and navigation of similarly-tagged or connected elements, respectively, is facilitated with the visualizations. However, there is no distinction or prioritization for pages that may need work or pages seeking additional contributors. The Wiki Contribution Analysis component could be used to motivate users to periodically review the pages they've contributed towards or encourage others to increase participation. But again, this requires human motivation to make use of the information and contact other users.

The performance of the Annoki platform overall was not thoroughly reported aside from minor quantitative usage data and qualitative descriptions of which features were particularly helpful. Although it has been deployed and is currently in use for a few years, there are no quantitative results reported on its effectiveness in improving collaboration. We are thus unable to determine how much collaboration initiation rate and quality are improved with this platform.

2.1.3 Socs

The Socs prototype developed by Atzenbeck and Hicks (2008) is an application for Mac OS that "serves as a means to express, store, and communicate social information about people." The goal of the application is to improve collaboration through increased social and group awareness. Socs provides these in Wikipedia by 1) providing a *social space* visualization of the authors contributing to each wiki page and the social groups to which they belong, 2) retrieving information on authors' activity levels (i.e. how frequently s/he modified the page), 3) enabling the user to flag authors of interest, and 4) integrating with the Apple Address Book framework. The hope is that the integration of contributor and group awareness features, information visualization, and communication tools improves collaboration by encouraging communication with and among wiki page authors, which increases understanding of author intentions and provides an avenue for “implicit recommendations” of other works through communication with other authors.

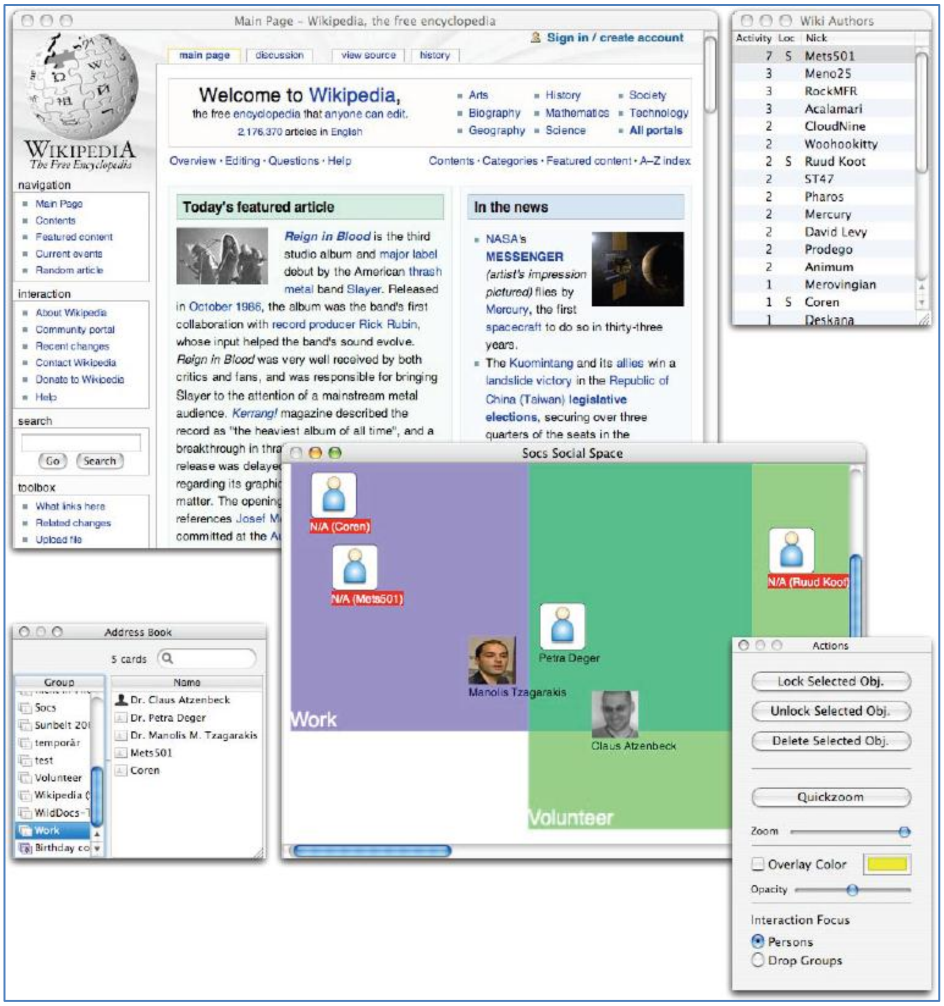
The cornerstone to the Socs prototype is the social space visualization, which presents the user’s people and groups of interest in a 2D area as seen in Figure 2.4. People are represented by markers, and groups are represented by colored rectangles. Membership to a group is represented by a marker’s presence within the corresponding rectangle, and presence in overlapping regions indicates membership in multiple groups. An algorithm is not used to programmatically discover group membership for each person on the social space – rather, groups and people on the space are limited to those already known by the user (i.e. in the user’s address book). Placement of the markers is then determined based on the groups that the user has manually *associated* them with. The space also utilizes other visual cues such as distance, alignment, color, and size to convey additional information to the user. The social space integrates with the Apple Address Book and Wiki (Page) Authors list via drag and drop functionality –people and groups from the address book and article authors from the authors list can be dropped into the user’s social space to visualize the relations between them and highlight their participation on wiki pages. Any changes to the social space (i.e. insertion and deletion of members and groups) are reflected in the system-wide address book, and contact with authors is facilitated by creating an e-mail to the associated person when a marker is clicked.

The second component to the Socs prototype is its awareness features. When the user navigates to a Wikipedia page (or another compatible website), Socs obtains its contributors and populates them in a list along with activity level (i.e. number of revisions made) for easy viewing. If an author is already in the Socs system, it is indicated in the “Loc” column of the window, and authors that are in the user’s current social space are also highlighted in the social space window. By highlighting authors in this manner, the user is: 1) made aware of acquaintances that took part in the article and the groups to which they belong and 2) provided with a simplified mechanism for contacting them if needed. The cost of communication is reduced since the article authors are already identified and tied to address book contacts.

Since Socs is primarily a social application, its primary modeling occurs on the user side, associating wiki contributors with various social groups that potentially overlap. Further, users are also linked to the pages that they’ve contributed towards, enabling easy access to other work performed by particular users. A visualization of these group associations can be seen in the “Social Space” window of Figure 2.4.

The proposed solution’s primary shortcoming, just like that described for the Annoki platform, is its lack of intelligent support. Since the application only highlights authors that manually marked by the user, the potential benefits of the tool is diminished since it does not identify, display, or recommend *new* social relations to groups or people that the user does not yet have on the social space. That is, there is no guided *discovery* of authors or social groups. While “strangers” may be manually added to the social space (and consequently, the user’s address book) from the Wiki Authors window, the user is not actively encouraged to communicate or collaborate with them. This is addressed in a component of our approach, which recommends new social relations via suggesting people to collaborate with.

There are currently no evaluation results reported on Socs’s performance, so we are unable to determine how well it improves collaboration initiation rate and quality.



Reign in Blood is the third studio album and major label debut by the American thrash metal band Slayer. Released in October 1986, the album was the band's first collaboration with record producer Rick Rubin, whose input helped the band's sound evolve. *Reign in Blood* was very well received by both critics and fans, and was responsible for bringing Slayer to the attention of a mainstream metal audience. *Kerrang!* magazine described the record as "the heaviest album of all time", and a breakthrough in thra… release was delaye… regarding its graphic… matter. The opening… references Josef M… committed at the Au…

- NASA's **MESSENGER** *(artist's impression pictured)* flies by Mercury, the first spacecraft to do so in thirty-three years.
- The Kuomintang and its allies win a landslide victory in the Republic of China (Taiwan) **legislative elections**, securing over three quarters of the seats in the

Activity	Loc	Nick
7	S	Mets501
3		Meno25
3		RockMFR
3		Acalamari
2		CloudNine
2		Woohookitty
2	S	Ruud Koot
2		ST47
2		Pharos
2		Mercury
2		David Levy
2		Prodego
2		Animum
1		Merovingian
1	S	Coren
1		Deskana

Figure 2.3: Screenshot of Socs social space, web browser, wiki authors list, address book (Atzenbeck and Hicks, 2008)

2.2 Other Relevant Wiki-Related Work

Although there are relatively few existing works with the express goal of improving wiki collaboration, other wiki-related research may be relevant to our goal of improving collaboration initiation rate and quality. In particular, research in determining article quality and user expertise is especially relevant to our interests, since their inclusion in our user and data models may improve the accuracy of our recommendations. We have included some of these measures and the ideas that they are based on in our own recommendation algorithm in Chapter 3. We specifically incorporate article quality based on Lih's (2004) "rigor" and the notion of page quality based on contributing users' expertise in Hu et al.'s models.

2.2.1 Article Quality

Article quality is relevant to data modeling and the recommendation-based approach of improving collaboration since it: 1) helps determine good quality articles to highlight (e.g. for recommending articles to view) and 2) helps determine which articles are of poorer quality and need work (e.g. for recommending articles to edit) (Huet al.2007). The approaches to calculating this can be broadly categorized based on the information used to make the calculation. Specifically, we examine Lih's (2004) metadata-based quality metrics and the article content-based quality models of Hu et al. (2007).

- Based on Metadata
    - Rigor (Lih 2004) – the number of edits made to an article.
    - Diversity (Lih 2004) – the number of unique editors for the page.
- Based on Article Content

- Basic (Hu et al. 2007) – article quality as a function of the expertise of contributing authors and the amount each author contributed to the article.
- PeerReview (Hu et al. 2007) – Basic model with text “review”; *all* unmodified text in a revision is considered “reviewed” by the author, boosting its quality.
- ProbReview (Hu et al. 2007) – PeerReview with a probabilistic model of text review; text that is *closer* to the revision author’s contribution is *more likely* to be reviewed.

Lih (2004) proposes two basic methods for benchmarking article quality based strictly upon metadata, i.e. without analyzing the content of the article: *rigor* and *diversity*. *Rigor* is the number of edits that the article has undergone, and its importance is based on the assumption that an article that has been edited more times undergoes a “deeper treatment of the subject or more scrutiny of the content.” *Diversity* is the number of unique authors contributing to the article, and greater diversity for an article is indicative of “more voices and different points of view” on its subject. Lih proposes finding benchmark values, i.e. high quality thresholds, for these measures by calculating the median rigor and diversity for a collection of benchmark Wikipedia articles.

Hu et al. (2007) developed three quality measurement models that calculate quality as a function of the expertise of its contributing authors: Basic, PeerReview, and ProbReview. The Basic model is based upon the assumption that higher expertise authors leads to a better quality article. An article’s quality is then the sum of the expertise of its contributing authors, with each author’s expertise weighted by the amount s/he has contributed to the page. However, an author’s expertise is also based on the quality of the pages that s/he contributed towards – thus, the two have a circular relation and reinforce one another. From this setup, values for quality and expertise are then calculated by first initializing them to a value and then iteratively computing them until they converge. PeerReview and ProbReview differ from Basic in that it introduces the notion of "reviewing" text in addition to authoring it. Text that is unchanged by an author in between revisions is considered to be reviewed and implicitly accepted, and the author's expertise is factored into the quality of the existing text. In the PeerReview model, it is assumed that all unmodified text is reviewed by the author. However, this assumption is not particularly accurate – users who contribute minor changes or contribute changes to only a specific area of the article. The ProbReview accounts for this by adding a probabilistic element to the "review" of unmodified text – it assumes that the unmodified text that is closer to the author's contributions are more likely to be reviewed than text further away from them.

These measures of article quality are related to our work since we have incorporated ideas suggested in both Lih's and Hu et al.'s works in our data models. Lih's rigor measurement (i.e. the number of edits) is used directly in our algorithm when calculating the recommendation score of the article due to its ease of implementation. We currently choose to exclude diversity from our data model since collaboration in our target domain is typically carried out in groups of a fixed size during its primary development. We also use Hu et al.'s idea of calculating article quality based on authors' expertise and the proportion of their contributions to the article. However, we use an alternative to convergence between article quality scores and author expertise, which may be computationally expensive when convergence is slow. Our alternative to this is further detailed in Chapter 2.2.2 and Chapter 3, and our recommendation algorithm is described in Chapter 3.

2.2.2 User Expertise Modeling

Closely related to the notion of article quality is the idea of determining a user's expertise, either through explicit feedback provided by other users or implicitly through the user's actions in the environment. For a wiki, this is often derived primarily from the user's contributions to wiki pages. In models that account for both page quality and user expertise, the two reinforce one another: the collective expertise of page authors contribute towards a page's quality, and the quality of each page in the user's editing history plays a role in determining his/her expertise.

As previously mentioned, Hu et al. (2007) make use of user expertise in their calculations of article quality. In the Basic model, user expertise is the sum of the qualities of the articles that the author has contributed towards, with each contributing term being discounted by the proportion of the text not authored by the user. That is, the quality of each article in the author's editing profile is multiplied by the percentage of the author's contribution. In the other two models, the expertise of users who have "reviewed" the author's text also contributes towards the author's expertise.

Similar to Hu et al.'s models for wiki article quality, the notion of author expertise is included in our user model and plays a role in our recommendation algorithm. As previously mentioned in 2.2.1, we calculate this in a manner different from what is proposed by Hu et al. since finding convergence may be computationally expensive. Instead, we calculate expertise based on contribution *longevity*. Its intuition is similar to that of PeerReview – text of high quality will be left unchanged (i.e. reviewed and accepted) in between revisions. The difference lies in how expertise is calculated. Hu et al. base this on the expertise of the authors who have "reviewed" the text, and this method requires convergence calculation. Our longevity approach is dependent only on time and does not require finding convergence. Contribution longevity and its use within our algorithm are further described in Chapter 3.

2.3 Recommendation Algorithms

Finally, research in existing recommender systems can be leveraged to improve wiki collaboration via recommendations for pages to view or edit. Recommendation algorithms are particularly relevant to our work since we wish to take a recommendations-based approach to improving collaboration between wiki users. Research in this area has largely been centered upon its use in e-commerce to suggest items for the user to purchase (e.g. on commercial websites) or on news and other special interest websites to suggest items to view. Examples of such algorithms include: 1) collaborative filtering, 2) content-based, and 3) hybrid recommendation algorithms.

Collaborative filtering leverages information from people similar to the target user in order to generate recommendations. However, it has a couple limitations, namely inaccurate recommendations for new users and new items, and inaccurate recommendations due to sparsity of ratings. Our approach contains collaborative filtering elements, but is not a pure collaborative filtering algorithm.

Content-based approaches utilize features to recommend items that are similar to items that the user has liked. While it generally lacks the same weaknesses as collaborative filtering, it has its own distinct limitations, including the need for a large feature set, indistinguishability of same-featured items, and a potential lack of diversity in recommendations. Our approach also contains content-based elements, but is not a pure content-based algorithm.

Hybrid recommendation algorithms combine multiple recommendation algorithms or elements from those algorithms to generate recommendations, with the idea that the varied strengths of the components compensate for their individual weaknesses. While recommendations from these algorithms have higher accuracy than their pure counterparts, they may be computationally more expensive to generate. Our approach (described further in Chapter 3) qualifies as a hybrid algorithm since it combines elements from Lih's rigor (Chapter 2.2.1) and Hu et al.'s page quality and user expertise (Chapters 2.2.1 and 2.2.2) in addition to elements from other algorithms mentioned in this subsection.

While existent, research in applying these algorithms to a wiki environment has not been as thoroughly explored as their use in the previously mentioned domains. Noteworthy examples of applications to wikis include: Cosley et al.'s SuggestBot (covered in Chapter 2.1.1) which utilizes a hybrid composite of multiple recommendation approaches, one of which is based on collaborative filtering and another is content based; and the works of Durao and Dolog (2009) and Sruirai et al. (2009)which utilizes a content-based approaches to suggest pages to view.

The contents of the user and data models leveraged by the recommenders are generally limited to only the requisite attributes needed to generate the recommendations, i.e. data used during the computation. Consequently, currently-popular algorithms may overlook wiki-specific factors (such as frequency and implicit quality of edits) and the synergy between factors in separate algorithms when applied to the wiki domain without modifications. While specific model contents are dependent upon the approach taken in hybrid recommenders, general statements can be made of user and data models for collaborative filtering and content-based recommendation. Descriptions of these can be found in their corresponding sub-chapters.

2.3.1 Collaborative Filtering

Adomavicius and Tuzhilin (2005) describe collaborative recommendation methods as predictions on the utility of an item for a particular user based on ratings that similar users have given it. That is, it recommends items that users with similar preferences – “neighbors” – have found favorable. Sarwar et al. (2000) generalize the collaborative filtering process into three parts: 1) the representation of input data, 2) neighborhood formation, and 3) recommendation generation. Input data are typically represented in most CF-based algorithms as an *M* by *N* customer-item matrix where *M* is the number of users and *N* is the number of items in the system. Each entry denotes a user’s affinity (e.g. through rating, number of views, etc.) for the item. The biggest differences between the various CF-based algorithms then lie in the techniques used in neighborhood formation and recommendation generation. For example, the popular user-based top-N variant of CF uses Pearson correlation to determine the *k* users most similar to the target user. Predicted ratings are then calculated by taking a weighted average of the ratings given by these *k* neighbors, and the top *N* items with the highest ratings are recommended to the target user. As another example, one of the recommenders used in Cosley et al.’s SuggestBot (described in Chapter 2.1.1) leverages collaborative filtering.

This recommendation approach is driven largely by its user modeling. The most common ratings-based CF approach uses a model where user interests are represented by the set of ratings provided throughout their entire history within the system. Recommendations are then generated by comparing user models and aggregating a "neighborhood score" for items not yet rated by the target user. The "rating" aspect of representing user interest can be swapped out or augmented with other indicators available in the application domain, such as item views, edits, etc. Since this approach is not driven directly by page content, it does not leverage a data model in its computation.

Limitations of general collaborative filtering include: 1) inaccurate recommendations for *new users* and *new items* (i.e. new users and items lack the history needed to generate accurate recommendations for them) and 2) inaccurate recommendations due to *sparsity* of ratings (i.e. there is a lack of jointly rated items due to a very large number of items in the system relative to the number of items rated by users) (Adomavicius and Tuzhilin 2005). The CF algorithms have a worst-case performance when operating on very large and very sparse matrices. Performance can be improved by a large factor with reduction techniques, but the accuracy of recommendations can suffer (Linden et al. 2003).This recommendation approach is relevant to our work since our algorithm contains collaborative filtering-like aspects in determining the recommendation score of an article. Chapter 3 describes our algorithm in greater detail.

2.3.2 Content-Based

Content-based recommendation methods leverage the features or characteristics of an item to predict whether the target user would like it, based on how favorably the user has received items with similar features. That is, in contrast to collaborative filtering which focuses on similarity between users, content-based algorithms focus on similarity between items. The intuition is that users are more likely to enjoy items that have similar qualities to items that the user already likes. For instance, a person who enjoys the *Harry Potter* series may be more likely to enjoy *The Lord of the Rings* than *Lawrence of Arabia* since the former arguably has more in common with it than the latter.

There are several different approaches for generating recommendations within this category of algorithm, including those based on clustering (e.g. Shepitsen et al. 2008) or on keyword term frequency-inverse document frequency (e.g. the fulltext search-based recommender in Cosley et al.'s SuggestBot). An algorithm described by Adomavicius and Tuzhilin (2005) aggregates the target user's tastes into a feature vector and finds its cosine similarity to the feature vector of candidate items.

This recommender type utilizes both a user and data model in its calculations. Here, user interests and data content are represented as a subset of some set of keywords or tags global to the entire system. Either can be built *explicitly* through manual listing of interests and related topics, or *implicitly* through text analysis of page content viewed.

The limitations of content-based recommendation include: 1) the reliance on large feature sets that must be known beforehand if automatic feature extraction is not possible (e.g. in multimedia domains), 2) indistinguishability between items with *identical features*, 3) *lack of diversity* in recommendations(i.e. recommendations are limited to items containing features the user favors with little chance for "serendipitous" recommendations outside of one's usual tastes), and 4) inaccurate recommendations for *new users* (Adomavicius and Tuzhilin 2005). The content-based recommendation approach is relevant to our work since our algorithm contains content-based aspects in determining the recommendation score of an article (i.e. a component based on keywords). Chapter 3 describes our algorithm in greater detail.

2.3.3 Hybrid Recommendation Algorithms

One solution to overcoming the shortcomings of collaborative filtering and content-based algorithms is to "hybridize" it by leveraging elements or results from other recommendation algorithms that lack the same weaknesses, e.g. basing the collaborative filtering partially on item traits as in content-based recommendation or vice versa (Adomavicius and Tuzhilin 2005). Adomavcius and Tuzhilin (2005) specify three different ways in which algorithms can be hybridized: 1) generating recommendations from multiple algorithms separately and combining their results, 2) adding elements from other algorithms to a single "main" algorithm, and 3) constructing a single unifying model that incorporates elements from multiple algorithms. Cosley et al.'s SuggestBot is one example, leveraging four different recommenders combined via the first approach (see Chapter 2.1.1 for more details). Experimental results comparing pure collaborative filtering and content-based recommendations against their hybridized counterparts have confirmed that the performance of the hybrid CF algorithms provides superior accuracy at the cost of additional computational complexity (Melville et al. 2002, and Han and Karypis 2005).

This is particularly relevant to our work since we leverage the third hybrid approach for our page recommendation algorithm, combining content-based elements such as the identification of similar pages via keywords, collaborative filtering elements such as the identification of peers with similar interests, and other elements such as author expertise, ratings, and other page metadata to suggest pages to view and edit. Rather than using only one of the “pure” algorithms previously described, we implement a hybrid one due to the large perceived cost of an incorrect collaboration recommendation. We appraise the cost of a false positive in this domain as greater than the cost of a false positive for recreational browsing due to the increased costs and potential losses for poor quality collaborations, as outlined in Chapter 1. This places increased importance on recommendation accuracy.

No single element or pure algorithm leverages all relevant information available in a wiki, and thus a single element on its own is not sufficient to provide accurate recommendations. Chapter 3 justifies our decision and describes our hybrid algorithm and associated user and data models in greater detail.

Chapter 3: Proposed Approach

As previously described in chapter 2, the existing works geared towards improving wiki collaboration either lack intelligent features that adapt to the users' profiles or fail to address both collaboration initiation rate and quality. We thus propose our own hybrid recommendation algorithm that leverages and unifies aspects of directly and tangentially related works to address these problems. The result is an algorithm that considers: 1) keywords- and/or tag-based similarity, 2) ratings-based collaborative filtering, 3) links between pages, 4) author reputation and expertise, and 5) the number of page views and edits to provide recommendations that are relevant to user (and thus encourages contribution) and of sufficient quality.

This chapter is organized in the following manner. We introduce our proposed algorithm in Chapter 3.1. In Chapter 3.2 we describe the Wikipedia Page Recommendation feature added to support the use of the Biofinity Intelligent Wiki in a classroom setting.

3.1 Page Recommendation Algorithm

Our hybrid page recommendation algorithm calculates a score for each page using a weighted mean of the individual component scores based on commonly used attributes of existing recommendation algorithms. To reiterate, these attributes are:

- Keywords-/tags-based similarity (e.g. Shepitsen et al. 2010, Cosley et al. 2010, Adomavicius and Tuzhilin 2005, etc.)
- Ratings-based collaborative filtering (e.g. Sarwar et al. 2000, etc.)
- User expertise (e.g. Hu et al. 2007, etc.)
- Links between pages (e.g. Cosley et al. 2010, Page et al. 1998, etc.)

- Number of page views and edits (e.g. Lih 2004, [older search engines], etc.)

These can be divided into two categories based on whether they contribute towards determining a page's *relevance* to the target user and its *quality*:

- Attributes determining relevance:
  - Keywords-/tags-based similarity
  - Links between pages
  - Ratings-based collaborative filtering
- Attributes determining quality:
  - Ratings-based collaborative filtering
  - User expertise
  - Number of page views and edits

Note that ratings-based collaborative filtering can be considered to fall into both categories – the ratings-based aspect determines page quality whereas the collaborative filtering with peers determines relevance to the target user.

In general terms, the page score of page $p_i = Score(p_i) = \boldsymbol{A}_{\boldsymbol{p}_i} \cdot \boldsymbol{W}$ where:

- Attribute scores: $\boldsymbol{A}_{\boldsymbol{p}_i} = \left[a_{key}, a_{rate}, a_{rep}, a_{links}, a_{views}, a_{edits}\right]$
- Attribute weights:$\boldsymbol{W} = [w_{key}, w_{rate}, w_{rep}, w_{links}, w_{views}, w_{edits}]$, where $w_{ttdk} + w_{rate} + w_{rep} + w_{links} + w_{views} + w_{edits} = 1$

The attribute weights will be initialized to the predetermined constants in Table 3.1. After calculating the page score, the top *n* highest-scoring pages will then be recommended to the user.

Attribute	**Weight**
Page Tags and Keywords w_{key}	0.25
Explicit Page Ratings w_{rate}	0.25
Author Reputation/Expertise w_{rep}	0.25
Links between Pages w_{links}	0.15
Number of Page Views w_{views}	0.05
Number of Page Edits w_{edits}	0.05

Table 3.1: Weights for each attribute in the weighted sum

The following subsections detail the values of these individual weights and the calculations made to obtain the individual attribute scores.

3.1.1 Page Topics, Tags, Disciplines, and Keywords

Page topics, tags, disciplines, and keywords are often used as primary attributes for generating recommendations in a wide variety of algorithms, e.g. Shepitsen et al's context-based hierarchical agglomerative clustering and many content-based recommendation algorithms (Shepitsen et al. 2008). These are often utilized in the following two ways: directly matching the terms to the target user's interests and indirectly matching *related* terms to the user's interests. Additionally, each topic may have varying levels of importance between different users.

We will thus leverage an existing algorithm developed by Shepitsen et al. that utilizes context-dependent hierarchical agglomerative clustering for personal recommendations (2008). It is selected since it is designed for the social tagging domain and makes use of the ideas mentioned in the previous paragraph. The algorithm is detailed as follows:

1. Calculate the cosine similarity between the user's interests and each resource:

   $S(u,r) = \cos(u,r) = \frac{\sum_{t\in T} tf(t,u) * tf(t,r)}{\sqrt{\sum_{t\in T} tf(t,u)^2} * \sqrt{\sum_{t\in T} tf(t,r)^2}}$, where

   - *T* is the set of all tags used in the system
   - *u* and *r* are vectors over the set of tags, with *u* representing the user's interests and *r* representing a wiki page
   - *tf(t,v)* is the tag frequency of tag *t* in vector *v* - for wiki pages, a tag frequency for a particular tag will only be 0 or 1

It should be noted that *T* can grow to be fairly large as the system grows, with the number of tags in the system being orders of magnitude larger than the number of pages and users. Further, not all tags will be relevant to all pages and users, resulting in relatively sparse vectors. Since tags that aren't relevant to the page or user do not figure into the calculation, we can limit the iterations to the union between the user's interests and the page's tags.

2. Calculate the relevance of the documents to the user:

   i. Calculate the target user u's interest in each cluster c:

   $uc_w(u,c) = \cos(u,c)$

   ii. Calculate each resource's closest clusters:

   $rc_w(r,c) = \cos(r,c)$

   iii. Calculate the user's modified interest in each resource:

   $$I(u,r) = \sum_{c\in C} uc_{w(u,c)} * rc_{w(r,c)}$$

Here, *Tags(i)* is defined to be the set of tags that an item *i* is associated with, where *i* is either a resource *r* or cluster *c*. Similarly, *Interests(u)* is the set of tags that a user *u* is observed to have interest in. We compute this by counting the tags associated with the pages that the user created, viewed, edited, positively rated, and discussed.

3. Calculate personalized rank scores

   $\alpha_{TTDK} = S'(u,r) = S(u,r) * I(u,r)$ where

   - *S'(u,r)* is the cluster-adjusted user-resource tag similarity
   - *S(u,r)* is the user-resource tag similarity computed in Step 1
   - *I(u,r)* is the target user's interest in resource *r* based on clustering

Details for how the tag clusters used in Step 2 of the procedure are found, as well as additional details on the algorithm, can be found in Shepitsen et al. (2008).

A key assumption that the algorithm had is that the recommendation is generated for single-tag queries – that is, the vector of user interests *u* only contains a single tag. It consequently lacks applicability to generating recommendations relevant to *all* user interests, and simply iterating this process over all user-interested tags may not scale up well. We thus adapted the algorithm to cover the entire spectrum of the user's interests.

The weight for this attribute w_{key} will be initially set to 0.25 due to the relative importance of topics, etc. in determining whether a page is suitable to the target user.

3.1.2 Explicit Page Ratings

It is found by Papagelis and Plexousakis (2005) that recommendations based on explicit ratings by users are generally more accurate than those determined through implicit measures. Thus, we can provide more-accurate recommendations by leveraging the explicit page ratings provided by other users. There is a possibility for frustration bias to factor into the ratings if the system actively and persistently queries the user to obtain these ratings. However, this bias can be ignored since the ratings are voluntarily provided.

Our system used a binary voting system of "Likes" and "Dislikes." The net page rating for this particular system is then a simple difference between the number of Likes n_{like,p_i} and the number of Dislikes $n_{dislike,p_i}$.

$$\text{Net Page Rating for page } p_i = r_{p_i} = n_{like,p_i} - n_{dislike,p_i}$$

Since a page's contents change over time as users make revisions, it is possible that older ratings are not indicative of contemporary opinion towards it. We will thus weight the raw net page rating according to when it was made relative to the date the recommendation calculations are performed. For simplicity, we will consider time as a collection of discrete time periods where all ratings in the same period receive the same weight.

We will use a weighted harmonic mean to calculate the time-adjusted page rating, and thus, the contributing value to the page score:

$$a_{rate} = WHM(r_{p_i})$$

The weight for this attribute w_{rate} will be initially set to 0.25 due to the importance of explicit page ratings relative to the other attributes.

3.1.3 Reputation and Expertise of Author(s)

In trust and reputation-based systems, reputation impacts the perceived credibility of a user. This is analogous to trusting and valuing the opinions of domain experts.

Similarly, the reputation and expertise of the contributing authors should be considered when determining whether a page should be recommended.

We can trace the page content to the users responsible for each contribution by successively “diff-ing” each revision to the page to determine the changes made with each one. The page reputation derived from author reputations is then calculated as:

$$\alpha_{rep} = \sum_{u \in U_{p_i}} Reputation(u) * c_u, \text{ where}$$

- U_{p_i} is the set of all contributing users to page p_i
- $Reputation(u)$ is the reputation score of user *u*, consisting of a linear combination of expertise and explicit user ratings:

  $$Reputation(u) = 0.5 * Expertise(u) + 0.5 * Rating(u)$$

  Expertise and *Rating* functions are defined in chapters 3.1.3.1 and 3.1.3.2.
- $c_u = [0,1]$ is the proportion of the content in the latest revision authored by user *u*

The weight for this attribute w_{rep} will be initially set to 0.25 due to the perceived importance of author reputation in making recommendations.

3.1.3.1 User Expertise

The expertise of a user can be a key factor in determining the user’s reputation. Depending on the target user, similar levels of expertise (i.e. a peer relationship) or disparate levels of expertise (i.e. a mentor-mentee relationship) may be sought. Within our system, we define *expertise* to be a quality inherent in the revision contributions that the users make, distinguishing it from *participation* which encompasses any sort of action the user takes within the wiki. The implicit indicators of expertise that we consider include: views per contribution authored, ratings towards contributions authored, and longevity of page contributions. We combine these measures via a weighted sum.

$$a_{exp} = Expertise(u) = \sum_{attr\in\{views,ratings\}} \beta_{attr} * Contribution_{attr}(u)$$

To determine the expertise of a particular user, we calculate the user’s impact on the views and ratings received by the page. This is essentially done by scaling the views and ratings that a particular revision has received by the proportion of the user’s contribution, in terms of word count relative to the entire page.

The contribution of each attribute by user u, $Contribution_{attr}(u)$, is then defined as:

$Contribution_{attr}(u) = \sum_{e\in E_u} \sum_{i=e}^{j} Value_{attr}(i, i-1) * \gamma * c_{u,i}$ where

- E_u is the set of all edits made by the user *u*
- *e* is a particular revision made by the user *u* from E_u
- *j* is the last revision of the page containing revision *e*
- $Value_{attr}(i, i-1)$ is the value of the attribute *attr* between revisions *i* and *i*-1
    - For views: $Value_{views} = WHM(v_i) - WHM(v_{i-1})$, i.e. the change in the weighted harmonic mean of the number of page views between revisions *i* and *i-1*
    - For ratings: $Value_{ratings} = WHM(r_i) - WHM(r_{i-1})$, i.e. the change in the weighted harmonic mean of the page rating between revisions *i* and *i-1*

- γ is the time weighting applied. Since we wish to reward contributions for surviving subsequent revisions, this weight increases with revisions further away from revision e. This factor specifies the importance of a contribution's longevity.
- $c_{u,i}$ is the proportion of target user u's contribution towards revision *i* (i.e. the percentage of the revision that is content added by *u*). It is found by "diff"-ing revision *i* with revision *i*+1 and is bounded by [0,1].

Table 3.2 below details the weights and value used for each attribute. The weights for each contribution are preliminarily set based upon their perceived importance in determining expertise.

attr	$\boldsymbol{\beta_{attr}}$	$\boldsymbol{Value_{attr}(i, i-1)}$
Views	.5	Number of views the page received between revisions *i* and *i-1*, normalized by total number of views page received
Ratings	.5	Average rating per user using ratings received between revisions *i* and *i-1*

Table 3.2: Weights and values for attributes used in expertise calculation

3.1.3.2 Explicit User Ratings

Like with explicit ratings for wiki pages, users may also explicitly rate other users on the binary scale of "Likes" and "Dislikes." The net user rating for this rating scheme is then a simple difference between the number of Likes n_{like,u_i} and the number of Dislikes $n_{dislike,u_i}$.

$$\text{Net Rating for user} u_i, r_{u_i} = n_{like,u_i} - n_{dislike,u_i}$$

Since opinions on users may change over time as they improve and make contributions, it is possible that older ratings are not indicative of their current performance. We will thus weight the raw net user rating according to when it was made, relative to the date the calculations are performed. Again, we will consider time as a collection of discrete time periods where all ratings in the same period receive the same weight.

We will use a weighted harmonic mean to calculate the time-adjusted user rating, and thus, the contributing value to the user score:

$$a_{rate} = WHM(r_{u_i})$$

Where r_{u_i} is the net rating for user u_i, and the WHM function is as defined in Appendix A.

3.1.4 Links Between Pages

The Google search engine, the most widely-used internet search engine, makes use of a modified version of the publically available PageRank algorithm (Page et al. 1998). This algorithm essentially calculates the relative importance of a page by calculating the likelihood that a user browsing at random will reach it. That is, a page's importance is proportional to its in-bound links and inversely proportional to its outbound links. Essentially, we consider a page important if many other pages link to it.

The overall PageRank is determined as:

$$\alpha_{links} = I(p_i) = \sum_{p_j \in B_i} \frac{I(p_j)}{l_j}, \text{ where:}$$

- B_i is the set of all pages linking to p_i
- l_j is the number of outbound links from page p_j

The PageRank algorithm works as follows:

- Create a hyperlink matrix ***H***where $H_{ij} = \left\{\frac{1}{l_j} if\ p_j \in B_i, else\ 0\right\}$

- Form a “stationary vector” I^0whose components are PageRanks such that I^0 is an eigenvector of matrix $\boldsymbol{H}$ with eigenvalue 1.
- Repeatedly calculate $I^{k+1} = \boldsymbol{H}I^k$ until *I* converges, and this convergence scales linearly in $\log(n)$ where *n* is the number of directed links between the pages evaluated (Page et al. 1998). The post-convergence values in *I* are the PageRanks of each p_i.

The weight for this attribute w_{links} will be initially set to 0.15 , which is lower than the previous weights assigned thus far. This is due to the fact that the PageRank algorithm, on its own, does not consider the relevance of the connection between pages. While this can allow for the serendipitous discovery of strongly-linked concepts that users are unaware of, a recommendation is more likely to be followed if its relation to established interests are more apparent.

3.1.5 Number of Page Views

A page that is viewed more often is considered to be more popular, which may to some degree be indicative of the page quality. While its accuracy and precision for information retrieval are questionable as exemplified by early search engines, it plays a role in identifying which pages may be considered essential reading by the user base.

Like with explicit page ratings, we consider recent page views to be of greater importance than older page views. We will discount the number of views with time using the same weighting as for explicit ratings. The weighted harmonic mean a_{views} is then calculated in the same manner:

$$a_{views} = WHM(v_{p_i})$$

Where v_{p_i} is the number of views for page p_i, and WHM is as defined in Appendix A.

The weight for this attribute w_{views} will be initially set to 0.05, which is considerably lower than the previous weights assigned thus far. This is due to the facts that: 1) the number of page views can be easily manipulated, i.e. artificially inflated, and 2) early search engines using this attribute to return search results were not particularly successful.

3.1.6 Number of Page Edits

As suggested by Lih (2004), a page that is subject to many edits is more likely to be of higher quality after being refined many times, and the content on a frequently edited page may arguably be considered "fresher" than those updated less frequently. However, a high or low edit count may hold negative implications. For instance, a high edit count may be indicative of less value contributed per edit. Similarly, a low edit count may be indicative of an abandoned page when instead the page may be relatively "complete."

We thus use the number of page edits as the recommendation factor. Like with the number of page views and the explicit page ratings, we consider the more-recent edit counts to be of greater importance than older edit counts, and we will thus weight the time-adjusted page edits in a similar manner to views and ratings. The weighted harmonic mean a_{edits} is then calculated as:

$$a_{edits} = WHM(e_{p_i})$$

Where $e_{p_i}$is the number of edits for page p_i, and the WHM function is as defined in Appendix A.

The weight for this attribute $\boldsymbol{w_{edits}}$ will be initially set to 0.05, which is considerably lower than the previous weights assigned thus far. This is due to the facts that: 1) the number of page edits can be easily manipulated, i.e. artificially inflated; 2) the quality/value added of the edit is not considered, i.e. the edits may primarily be aesthetic or minor edits; and 3) the attribute has the uncertain implications previously described.

3.2 Wikipedia Page Recommendation

To further assist users with understanding page content and contributing to the Biofinity wiki, we have also implemented recommendations to Wikipedia pages based upon keywords located on the page currently viewed or edited. The goal of this feature is to improve student collaboration via easy access to peripheral information that can: 1) improve student comprehension of the page and related topics, and 2) aid the student in contributing content during the early revisions of the page. Due to time constraints in deploying the feature prior to student use, we opted for a fast, basic approach on this initial implementation that leverages the text processing capabilities of LingPipe.

The Wikipedia page recommendation process consists of three basic steps: 1) offline generation of a keyword dictionary, and 2) counting the occurrences of keywords for each revision of each Biofinity wiki page, and 3) ordering and presenting the results to the user.

3.2.1 Dictionary Generation

The keyword/phrase dictionary used for this feature is generated via depth-limited, breadth-first traversal of Wikipedia pages, starting from arbitrary root pages. The titles of each page visited are added to the dictionary with common stop words filtered out. Duplicates (i.e. pages linked to by more than one page) are added only once, and redirects (i.e. aliases and plural versions of pages) are linked to their target pages before both the redirect and target pages are added. Separate dictionaries are created for the biology lab and artificial intelligence courses.

3.2.2 Keyword Count

Each time a Biofinity wiki revision is saved, the revision text is scanned for dictionary keywords/phrases. For each keyword/phrase matched, a counter for it is incremented. Occurrences of a keyword/phrase originating from a Wikipedia redirect page count towards the redirect target's keyword/phrase. A keyword/phrase's score for a revision is the value of its counter at the end of the scanning and matching process.

3.2.3 Sort and Presentation

The top 20 Wikipedia keywords presented to users viewing or editing a page are ordered by their counts (as detailed in 3.2.2) with tie breakers handled by the word length of the keyword/phrase.

Chapter 4: Implementation

To fulfill our goals of obtaining wiki usage data and creating a framework to generate and present recommendations to wiki users, we developed and implemented a proprietary, full-featured wiki that integrates with the Biofinity Project. The wiki, currently dubbed as the Biofinity Intelligent Wiki, supports the following basic features:

- **Viewing, creating, editing, and deleting pages**
- **Page search** – indexing pages based on content and retrieving them based upon user query
- **File upload and download** – adding and retrieving files such as images, documents, and videos from the wiki

The wiki also supports the following Web 2.0 and social features:

- **Page tagging** – associate "tags" or key words with a wiki page to denote topics, relevant categories,
- **Page ratings** –express opinions on overall quality of a page's contents via a 1-5 scale
- **Page sharing via Facebook, Twitter, and the intra-wiki framework** – share pages to other users within the wiki, or to other social media outlets such as Facebook and Twitter
- **Comment/discussion threads** – generate comments and carry out threaded discussions on the page content

Additionally, the wiki also contains the following intelligent features:

- **Page and user modeling** – create and maintain a model of each page and user in the system, to facilitate implementation of intelligent features
- **User tracking via an agent framework** – a framework of "personal agents" to monitor and track user activity, with potential to carry out further autonomous action in future work
- **Recommendation framework** – a framework to automatically generate recommendations for wiki users, with potential to add multiple recommendation algorithms/techniques

The latter two intelligent features– user tracking and recommendation framework – are the ones that directly enable the investigation of the thesis topic. In particular, the user tracking feature is the cornerstone that provides data for our analyses in the next chapter.

Since the wiki is integrated with the Biofinity Project and is to be running on the same server, we are constrained in the server and database software used. Specifically, the Biofinity core ran using Glassfish v3 and a MySQL database, and the wiki was designed to operate in the same environment. Additionally, it was required that the wiki be encapsulated as a separate project and be packaged into a separate WAR file for easy deployment to the Glassfish server. By separating the wiki in this manner, changes to the wiki would not require changes to the core Biofinity site and vice-versa.

While the core Biofinity Project was written in Scala and leveraged the Lift framework, we wanted to use a more general and common language for the wiki in order to ease its development and to ease the implementation of future work for it. Java, HTML, and Javascript are an attractive alternative, since they are among the most common and popular tools used for web development on both the server and client ends. The decision is further simplified with the Google Web Toolkit (GWT), which enables the development of web applications written completely in Java, compiling the source files into equivalent HTML and Javascript code. These factors, combined with our familiarity with the language, led us to write the wiki almost completely in Java.

To summarize, the core technologies used by the Biofinity Intelligent Wiki are:

- Glassfish v3
- MySQL
- Java EE 6
- Google Web Toolkit

The rest of the chapter is arranged in the following manner. First, we will briefly describe the overall architecture of the Biofinity Intelligent Wiki in Chapter 4.1, including the interactions between the wiki client, server, and main Biofinity site. We will then delve into the specific architectures of the client, server, and database sides of the wiki in Chapters 4.2 through 4.4. Please note that the wiki features and implementation details discussed in this chapter may have changed after the time it is written.

4.1 Overall Architecture

Figure 4.1 below summarizes the Biofinity and Intelligent Wiki components and the interactions between them.

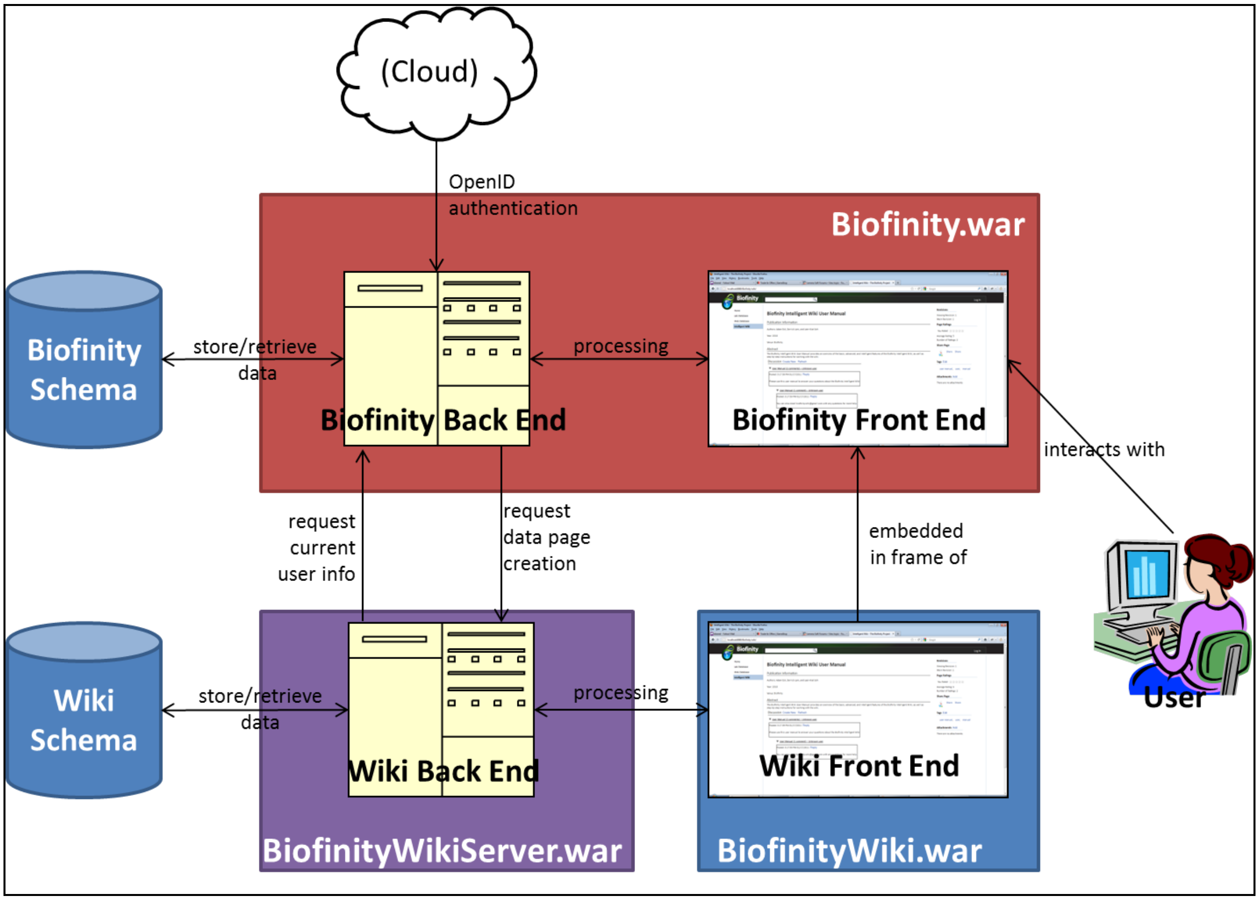


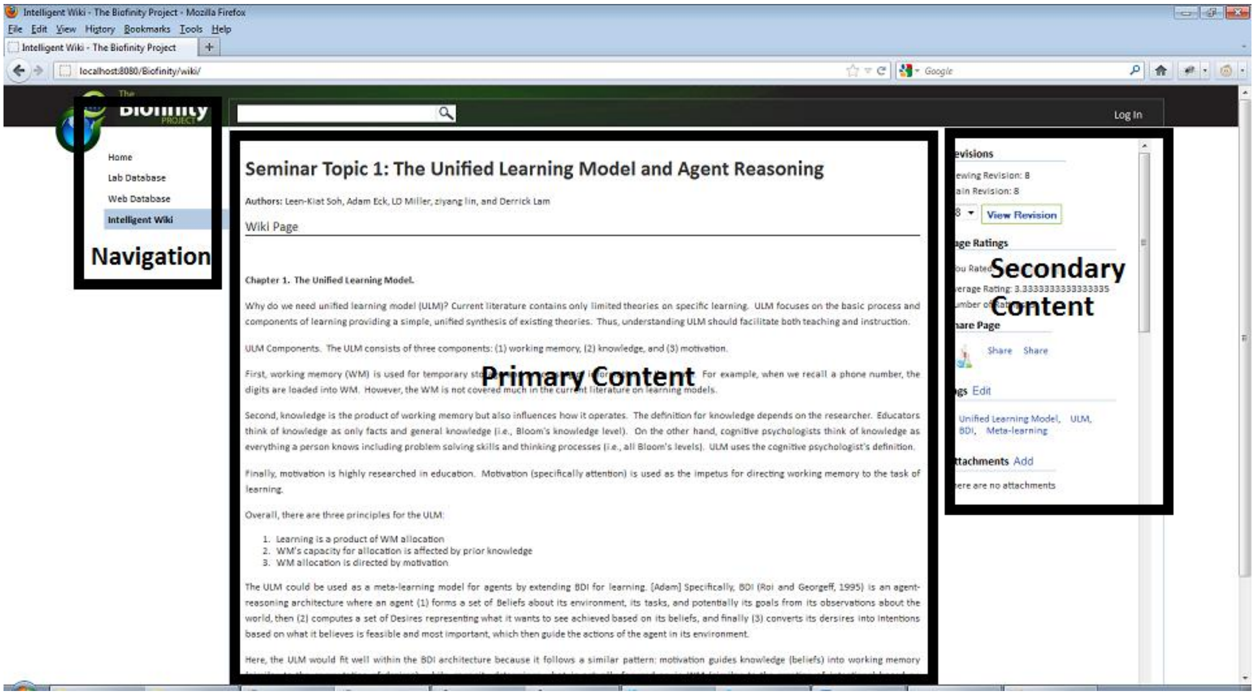
Figure 4.4: Overall Biofinity and wiki architecture

As previously mentioned, the wiki exists as a separate project from the Biofinity core, which is bundled in Biofinity.war in Figure 4.1. The Biofinity core and wiki store data in separate databases and rarely store or retrieve data from their counterpart, with few exceptions (e.g. querying user permissions). Further, the wiki front end and back end components are separated into their own packages, BiofinityWiki.war and BiofinityWikiServer.war, respectively. While the user interacts with the wiki front end through a frame in the Biofinity front end, the wiki server performs the bulk of the wiki-related processing.

The wiki's implementation has a few aspects and features tied to the Biofinity core. These include:

- Matching the Biofinity core's "look and feel"
- Integrating with Biofinity authentication and search
- Enabling the creation and linking of automatically-populated wiki pages from Biofinity data

First, the wiki's appearance must match the "look and feel" of the main site. Thus, the wiki front end is designed to make use of Biofinity's stylesheets and three-column page layout. Since one of the columns is reserved for navigation, the wiki page content – panels and text – is placed in the two remaining columns. Figure 4.2 illustrates this layout on a deployed wiki. Throughout the remainder of this chapter, we refer to the left sidebar as the *navigation sidebar*, the center area as the *primary content area*, and the left sidebar as the *secondary content area*.



Seminar Topic 1: The Unified Learning Model and Agent Reasoning

Authors: Leen-Kiat Soh, Adam Eck, LD Miller, ziyang lin, and Derrick Lam

Wiki Page

Chapter 1. The Unified Learning Model.

Why do we need unified learning model (ULM)? Current literature contains only limited theories on specific learning. ULM focuses on the basic process and components of learning providing a simple, unified synthesis of existing theories. Thus, understanding ULM should facilitate both teaching and instruction.

ULM Components. The ULM consists of three components: (1) working memory, (2) knowledge, and (3) motivation.

First, working memory (WM) is used for temporary st For example, when we recall a phone number, the digits are loaded into WM. However, the WM is not covered much in the current literature on learning models.

Second, knowledge is the product of working memory but also influences how it operates. The definition for knowledge depends on the researcher. Educators think of knowledge as only facts and general knowledge (i.e., Bloom's knowledge level). On the other hand, cognitive psychologists think of knowledge as everything a person knows including problem solving skills and thinking processes (i.e., all Bloom's levels). ULM uses the cognitive psychologist's definition.

Finally, motivation is highly researched in education. Motivation (specifically attention) is used as the impetus for directing working memory to the task of learning.

Overall, there are three principles for the ULM:

1. Learning is a product of WM allocation
2. WM's capacity for allocation is affected by prior knowledge
3. WM allocation is directed by motivation

The ULM could be used as a meta-learning model for agents by extending BDI for learning. [Adam] Specifically, BDI (Roi and Georgeff, 1995) is an agent-reasoning architecture where an agent (1) forms a set of Beliefs about its environment, its tasks, and potentially its goals from its observations about the world, then (2) computes a set of Desires representing what it wants to see achieved based on its beliefs, and finally (3) converts its dersires into Intentions based on what it believes is feasible and most important, which then guide the actions of the agent in its environment.

Here, the ULM would fit well within the BDI architecture because it follows a similar pattern: motivation guides knowledge (beliefs) into working memory

Figure 4.5: Biofinity Intelligent Wiki Layout

Second, the wiki must integrate with Biofinity's authentication system and search bar. Biofinity uses the Google OpenID authentication service[5], and users are redirected to a Google login page when attempting to log in to Biofinity. Since the Biofinity and wiki components have their own execution contexts and, consequently, their own session data instances for the same consumer, the session data pertaining to the current user needs to be synchronized across the two domains when the login occurs. That is, the Biofinity core notifies the Intelligent Wiki back end of the logged-in user via web service upon successful authentication. The Biofinity core also provides a web service for retrieving the current user and his/her current "lab", which the wiki leverages to renew its session data after it expires. The intelligent wiki also integrates with the search bar provided with the Biofinity core. While search bar events are handled by the core, wiki-related search queries are forwarded to the wiki server to process and return results. Details of the search indexing and results generation are detailed in Chapter 4.4.2.

The final integrated feature is the creation of *data pages* within the wiki. While viewing data in the Biofinity core, the user can request a wiki page to be created for it. The corresponding *data page* is a wiki page with a panel automatically populated with the corresponding information from the data in the Biofinity DB, and users can then expound upon the content with the wiki tools provided. Chapter 4.2.1.4 describes the format and features available for the data page in greater detail.

4.2 Wiki Front End Architecture

As previously mentioned, the intelligent wiki front end is primarily written with a subset of Java via the Google Web Toolkit, and although the GWT automatically

[5] http://code.google.com/apis/accounts/docs/OpenID.html

generates them from the Java source, we manually write some HTML and Javascript to supplement the generated code, e.g. for using the TinyMCE[6] rich text editor for edit mode. However, the amount of HTML and Javascript is relatively minimal – the front end is thus composed primarily of Java classes for each of the UI elements and their underlying representations. It interacts with the wiki back end by passing and receiving HTML messages and XML data through the back end's public-facing RESTful web services.

The front end is designed around the ideas of 1) providing distinct presentations or *pages* of the content to the user based upon the type of information involved, i.e. providing an *intelligent user interface*, and 2) reusing UI elements and features across these different presentations wherever possible as *panels* on the pages. In hierarchical terms, pages exist as top level items with subsets of panels as child elements.

Additionally, each page and panel has a corresponding data object that mirrors the representation used in the wiki back end. When the front end requests and receives XML data from the server, it immediately parses it into a data object whose values are used to populate the page or panel. Similarly, changes to the data object due to user interaction are sent to the server as XML data where it is parsed back into a data object.

To summarize, the Java classes in the wiki front end are categorized into three main groups, and we delve into their details in the upcoming sub-chapters:

- Pages (Ch. 4.2.1) – distinct presentations of information to the user

[6] http://www.tinymce.com/

  - Panels (Ch. 4.2.2) – re-usable features and UI elements, encapsulated in individual classes
  - Models – the underlying data representation of page and panel objects, as represented in the back end

As mentioned before, user tracking, i.e. logging user behavior and activities during sessions, is the primary source of data for our analyses in Chapter 5. Since every action and feature in the client requires requesting information from the server, we are easily able to determine what actions the user is performing, which wiki objects are involved, and when the action is performed. When applicable, the upcoming sub-chapters will also detail the tracking performed for each feature, as well as how it applies to our solution described in Chapter 3.

4.2.1 Pages

Information in the Biofinity Intelligent Wiki is primarily presented via *pages* that are equivalent to web pages on a website. While most of these pages exist to display particular information to the user, pages with user editable content each have two distinct layouts, one for viewing the information and one for editing the information. The content, panels, and functionality to be enabled in the *view mode* and the *edit mode* are determined by the type of the page requested. As of the writing of this thesis, the following distinct pages exist in the wiki:

- Editable Content Pages
  - Wiki Page (View/Edit)
  - Data Page (View/Edit)
  - Publication Page (View/Edit)

- Information Pages
  - User Page
  - Access Error Page
  - Consent Page
  - Main Page
  - Not Logged In Page
  - Results Page

While they may differ in terms of content and functionality available, they all have common events and members that every page should manage. These include keeping track of panels loaded for the page, requesting the root container to resize page contents to the Biofinity frame, and providing a means to determine whether a pop-up dialog box is currently open. We thus encapsulate these common elements into the *WebPage* interface and *AbstractPage* superclass from which all pages inherit from.

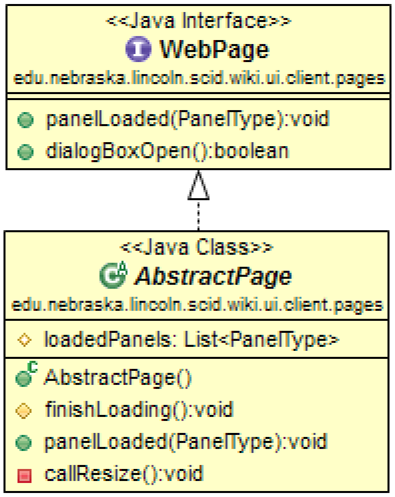

<<Java Interface>>
WebPage
edu.nebraska.lincoln.scid.wiki.ui.client.pages
panelLoaded(PanelType):void
dialogBoxOpen():boolean
<<Java Class>>
AbstractPage
edu.nebraska.lincoln.scid.wiki.ui.client.pages
loadedPanels: List<PanelType>
AbstractPage()
finishLoading():void
panelLoaded(PanelType):void
callResize():void

Figure 4.6: Class diagram for *WebPage* interface and *AbstractPage* superclass

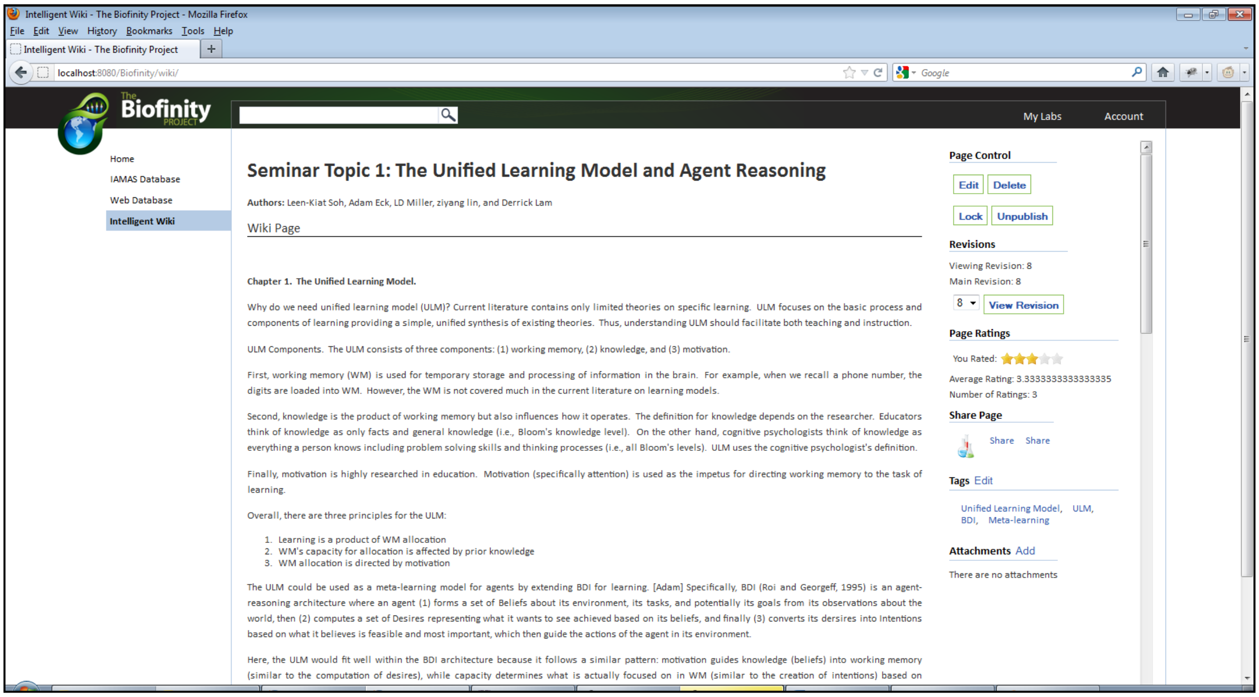
Generally, each of the pages follows a particular initialization and loading sequence when users request it. First, the user performs the page request, either by clicking a wiki hyperlink or manually entering a URL. These URLs often contain parameters specific to the page being loaded, including the page ID and particular revision ID to view. Starting with this information, the wiki client then queries the server for additional page-related information and processes the initial response with page-specific event handlers. Depending on the page, additional server requests/responses may be necessary before the initialization continues. Finally when sufficient information has been obtained, the page creates its UI elements, and instantiated panels may make their own server requests and perform their own processing before loading is finished. In a sense, this enables the asynchronous loading of the page's components, since the individual parts may finish loading before others, depending on which receive responses from the server first.

In terms of tracking, the system generally tracks views and edits made by wiki users to each of the pages. Broadly speaking, these impact the page's relevance to the user by determining topics and keywords of interest (Chapter 3.1.1) and the quality of a page through number of views/edits (Chapters 3.1.5 and 3.1.6) and through the user's expertise (3.1.3). The number of views and edits also impact the modeling and classification of a user, in terms of favoring active or passive activities and in terms of overachieving or minimalist performance (Chapter 5).

In the upcoming subsections (4.2.1.x) we describe each of the listed pages' functionality available and processes in greater detail.

4.2.1.1 Wiki Page

4.2.1.1.1 View Mode



Seminar Topic 1: The Unified Learning Model and Agent Reasoning

Authors: Leen-Kiat Soh, Adam Eck, LD Miller, ziyang lin, and Derrick Lam

Wiki Page

Chapter 1. The Unified Learning Model.

Why do we need unified learning model (ULM)? Current literature contains only limited theories on specific learning. ULM focuses on the basic process and components of learning providing a simple, unified synthesis of existing theories. Thus, understanding ULM should facilitate both teaching and instruction.

ULM Components. The ULM consists of three components: (1) working memory, (2) knowledge, and (3) motivation.

First, working memory (WM) is used for temporary storage and processing of information in the brain. For example, when we recall a phone number, the digits are loaded into WM. However, the WM is not covered much in the current literature on learning models.

Second, knowledge is the product of working memory but also influences how it operates. The definition for knowledge depends on the researcher. Educators think of knowledge as only facts and general knowledge (i.e., Bloom's knowledge level). On the other hand, cognitive psychologists think of knowledge as everything a person knows including problem solving skills and thinking processes (i.e., all Bloom's levels). ULM uses the cognitive psychologist's definition.

Finally, motivation is highly researched in education. Motivation (specifically attention) is used as the impetus for directing working memory to the task of learning.

Overall, there are three principles for the ULM:

1. Learning is a product of WM allocation
2. WM's capacity for allocation is affected by prior knowledge
3. WM allocation is directed by motivation

The ULM could be used as a meta-learning model for agents by extending BDI for learning. [Adam] Specifically, BDI (Roi and Georgeff, 1995) is an agent-reasoning architecture where an agent (1) forms a set of Beliefs about its environment, its tasks, and potentially its goals from its observations about the world, then (2) computes a set of Desires representing what it wants to see achieved based on its beliefs, and finally (3) converts its dersires into Intentions based on what it believes is feasible and most important, which then guide the actions of the agent in its environment.

Here, the ULM would fit well within the BDI architecture because it follows a similar pattern: motivation guides knowledge (beliefs) into working memory (similar to the computation of desires), while capacity determines what is actually focused on in WM (similar to the creation of intentions) based on

Viewing Revision: 8
Main Revision: 8

You Rated: ★★★☆☆
Average Rating: 3.3333333333333335
Number of Ratings: 3

Unified Learning Model, ULM, BDI, Meta-learning

There are no attachments

Figure 4.7: Wiki page (view mode) UI

Wiki pages, as seen in Figure 4.4 above, are the page type most similar to the typical article on Wikipedia and other wikis on the Internet, with the primary content area consisting of formatted text and images. The other intelligent wiki features available, such as basic page control, revision control, page rating, page sharing, tags, attachment upload/download, and a discussion area, are encapsulated in the panels attached to the page in both the primary and secondary content areas. That is, users can:

- make edits based upon the currently-viewed page revision
- delete the currently viewed revision
- lock the page from further edits
- publish the page for viewing by users outside the user’s lab group
- change the currently-viewed revision

- set the currently-viewed revision as the “main” one first seen on page load
- evaluate the page in the form of a 1-5 rating
- share the page with other wiki users through the intra-wiki recommendation system
- share the page through other social platforms such as Twitter and Facebook
- manually tag the page with related key words
- attach files to the page and download them
- participate in threaded discussions

Additionally, the Wikipedia Recommendation panel is also attached to wiki pages, and it displays suggested related Wikipedia articles based upon the keywords extracted from the page’s text. Further explanation for each of these functionality and their associated panels are located in Chapter 4.2.2.

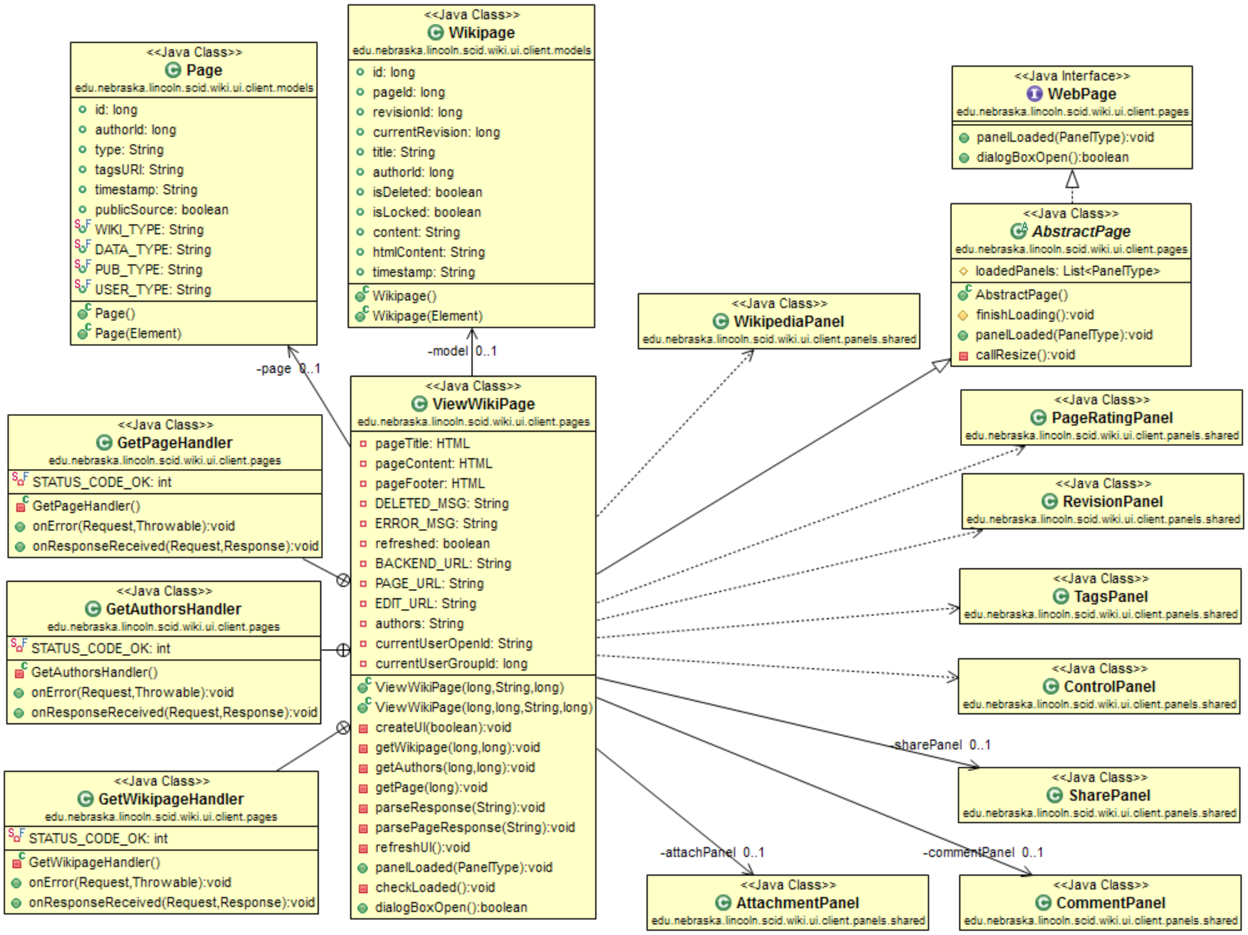
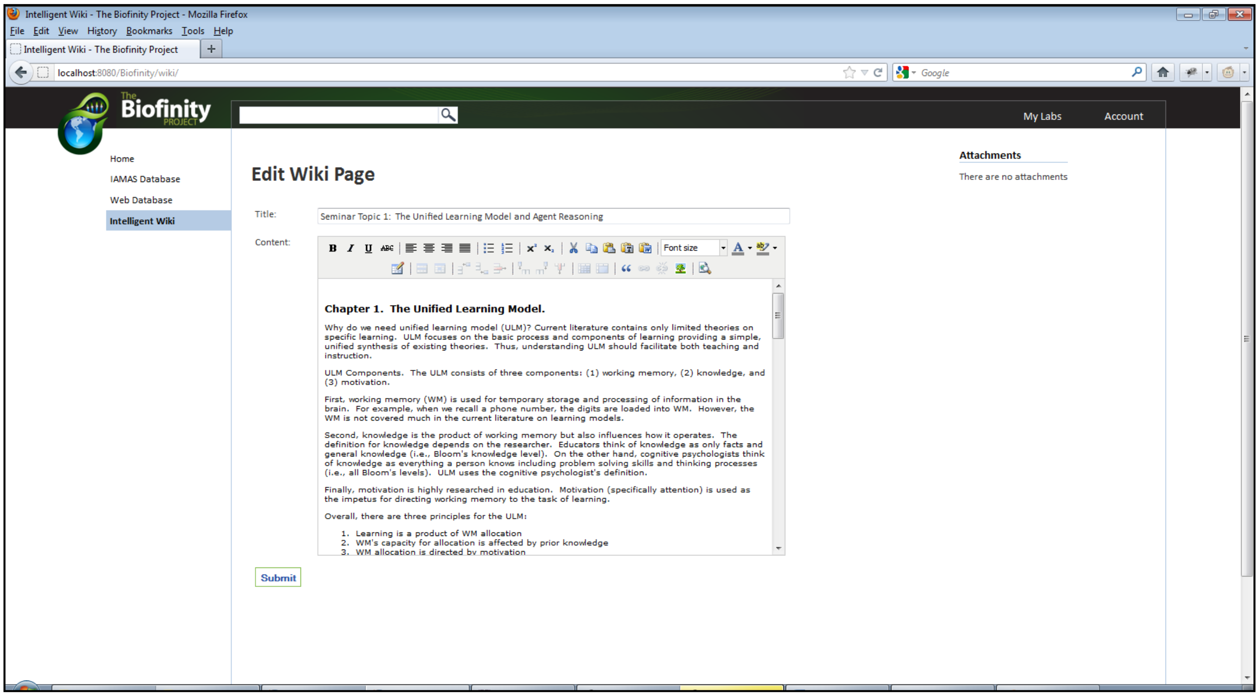


Figure 4.8: Wiki page (view mode) class diagram

Walking through the initialization process, the wiki page is first instantiated with a particular page ID, which is embedded as a URL parameter to the wiki client. It then queries the server for the basic page info associated with the ID, i.e. its deletion/lock status and its main revision (if one isn't specified in the URL). If the page is flagged as deleted or the user lacks permissions to view the page, then a corresponding message is displayed and processing stops. Otherwise, it requests the wiki page content for the particular page ID and revision. When that information is successfully returned, the wiki page instantiates its panels with the appropriate known page/revision information.

Viewing a wiki page creates a tracking entry for the user, which includes the page and revision viewed and the timestamp for the action. View actions contribute directly towards determining the page's quality (Chapter 3.1.5) and indirectly towards determining pages relevant to the user (Chapter 3.1.1).

4.2.1.1.2 Edit Mode


Intelligent Wiki - The Biofinity Project - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Intelligent Wiki - The Biofinity Project
localhost:8080/Biofinity/wiki/
Google
The Biofinity PROJECT
My Labs
Account
Home
IAMAS Database
Web Database
Intelligent Wiki
Font size
Submit

Edit Wiki Page

Title: Seminar Topic 1: The Unified Learning Model and Agent Reasoning

Content:

Chapter 1. The Unified Learning Model.

Why do we need unified learning model (ULM)? Current literature contains only limited theories on specific learning. ULM focuses on the basic process and components of learning providing a simple, unified synthesis of existing theories. Thus, understanding ULM should facilitate both teaching and instruction.

ULM Components. The ULM consists of three components: (1) working memory, (2) knowledge, and (3) motivation.

First, working memory (WM) is used for temporary storage and processing of information in the brain. For example, when we recall a phone number, the digits are loaded into WM. However, the WM is not covered much in the current literature on learning models.

Second, knowledge is the product of working memory but also influences how it operates. The definition for knowledge depends on the researcher. Educators think of knowledge as only facts and general knowledge (i.e., Bloom's knowledge level). On the other hand, cognitive psychologists think of knowledge as everything a person knows including problem solving skills and thinking processes (i.e., all Bloom's levels). ULM uses the cognitive psychologist's definition.

Finally, motivation is highly researched in education. Motivation (specifically attention) is used as the impetus for directing working memory to the task of learning.

Overall, there are three principles for the ULM:

1. Learning is a product of WM allocation
2. WM's capacity for allocation is affected by prior knowledge
3. WM allocation is directed by motivation

Attachments

There are no attachments

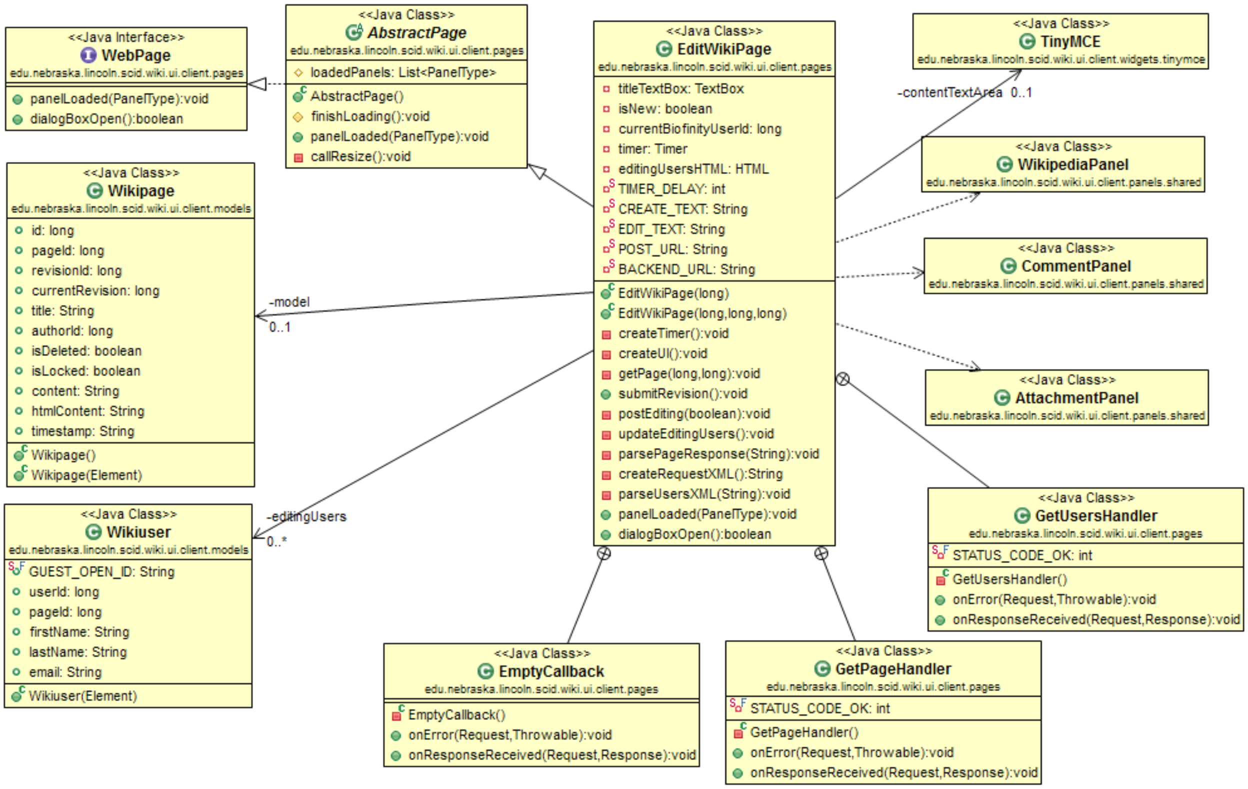
Figure 4.9: Wiki page (edit mode) UI

The edit mode of wiki pages, as seen in Figure 4.6, has considerably fewer elements than its counterpart, consisting of a two-field form, an attachments panel, a comments panel, and Wikipedia recommendations panel. While the title field is a regular text box, the content field is a TinyMCE WYSIWYG rich text editing text area, and its contents are transmitted to the server as plain text HTML upon revision submission. After a revision is successfully submitted, users are forwarded back to the view mode of the wiki page with the recently-added revision displayed.

The attachments panel is the same as the one used in the view mode, and users may upload additional files to the page. Any additions to the attachments list here are reflected upon the panel in the view mode. While it may seem unusual for the panel to be placed in the edit mode, this enables users to embed images within the content body via URL reference. Further explanation of how the attachments panel functions can be found in Chapter 4.2.2.2.

Similarly, the comments panel is the same as the one used in the view mode, and the comments and discussions carried out are also displayed in the comments panel. Additional comments can be made while in the edit mode, and changes made will be reflected in the view mode’s comments panel. Although the comments do not automatically refresh as new ones are made, his enables users to refer to the comments or carry out further discussion while editing the page by manually clicking the panel’s “Refresh” link. Further explanation of how the comments panel functions can be found in Chapter 4.2.2.3.

Finally, the Wikipedia recommendations panel is displayed as an aid to writing the revision by displaying hyperlinks to related Wikipedia articles. For convenience, clicking a link opens the corresponding article in a separate window, which provides a means for the user to quickly refer to the article to supplement their knowledge of related topics and prevents the user from losing his/her current work due to navigation. Further discussion on how the Wikipedia recommendations panel functions can be found in Chapter 4.2.2.



<<Java Interface>>
WebPage
edu.nebraska.lincoln.scid.wiki.ui.client.pages
panelLoaded(PanelType):void
dialogBoxOpen():boolean
<<Java Class>>
AbstractPage
edu.nebraska.lincoln.scid.wiki.ui.client.pages
loadedPanels: List<PanelType>
AbstractPage()
finishLoading():void
panelLoaded(PanelType):void
callResize():void
<<Java Class>>
EditWikiPage
edu.nebraska.lincoln.scid.wiki.ui.client.pages
titleTextBox: TextBox
isNew: boolean
currentBiofinityUserId: long
timer: Timer
editingUsersHTML: HTML
TIMER_DELAY: int
CREATE_TEXT: String
EDIT_TEXT: String
POST_URL: String
BACKEND_URL: String
EditWikiPage(long)
EditWikiPage(long,long,long)
createTimer():void
createUI():void
getPage(long,long):void
submitRevision():void
postEditing(boolean):void
updateEditingUsers():void
parsePageResponse(String):void
createRequestXML():String
parseUsersXML(String):void
panelLoaded(PanelType):void
dialogBoxOpen():boolean
<<Java Class>>
TinyMCE
edu.nebraska.lincoln.scid.wiki.ui.client.widgets.tinymce
-contentTextArea 0..1
<<Java Class>>
WikipediaPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.shared
<<Java Class>>
CommentPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.shared
<<Java Class>>
AttachmentPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.shared
<<Java Class>>
Wikipage
edu.nebraska.lincoln.scid.wiki.ui.client.models
id: long
pageId: long
revisionId: long
currentRevision: long
title: String
authorId: long
isDeleted: boolean
isLocked: boolean
content: String
htmlContent: String
timestamp: String
Wikipage()
Wikipage(Element)
-model
0..1
<<Java Class>>
Wikiuser
edu.nebraska.lincoln.scid.wiki.ui.client.models
GUEST_OPEN_ID: String
userId: long
pageId: long
firstName: String
lastName: String
email: String
Wikiuser(Element)
-editingUsers
0..*
<<Java Class>>
GetUsersHandler
edu.nebraska.lincoln.scid.wiki.ui.client.pages
STATUS_CODE_OK: int
GetUsersHandler()
onError(Request,Throwable):void
onResponseReceived(Request,Response):void
<<Java Class>>
EmptyCallback
edu.nebraska.lincoln.scid.wiki.ui.client.pages
EmptyCallback()
onError(Request,Throwable):void
onResponseReceived(Request,Response):void
<<Java Class>>
GetPageHandler
edu.nebraska.lincoln.scid.wiki.ui.client.pages
STATUS_CODE_OK: int
GetPageHandler()
onError(Request,Throwable):void
onResponseReceived(Request,Response):void

Figure 4.10: Wiki page (edit mode) class diagram

Editing a wiki page creates a tracking entry for the user, which includes the page edited, the particular revision's ID, and the timestamp for the action. Edit actions contribute directly towards determining the page's quality (Chapter 3.1.6) and the user's expertise (Chapter 3.1.3), and contribute indirectly towards determining pages relevant to the user (Chapter 3.1.1).

4.2.1.2 Data Page

4.2.1.2.1 View Mode

Data pages are pages that are generated per user request from occurrence, event, location, or classification data in the Biofinity database. In terms of page layout and available features, data pages are nearly identical to wiki pages – the only difference is the addition of a *data panel* that displays the corresponding Biofinity data for which the page was generated. Due to this similarity, data pages are represented in the underlying implementation as having both a WikiPage and a DataPage – the editable wiki component of the page is encapsulated in the former whereas the Biofinity data-specific information, such as the data type and data entity ID are contained in the latter. This representation is also used on the server side and in the database.

Since they contain different fields, each of the four page types has a different data panel, although the fields and values for all four are displayed in "<heading> <value>" format. For more information on the data panels, refer to Chapter 4.2.2.10. Note that the Biofinit data cannot be edited directly through the wiki interface – the user must navigate to and edit the data through their Biofinity lab instead.

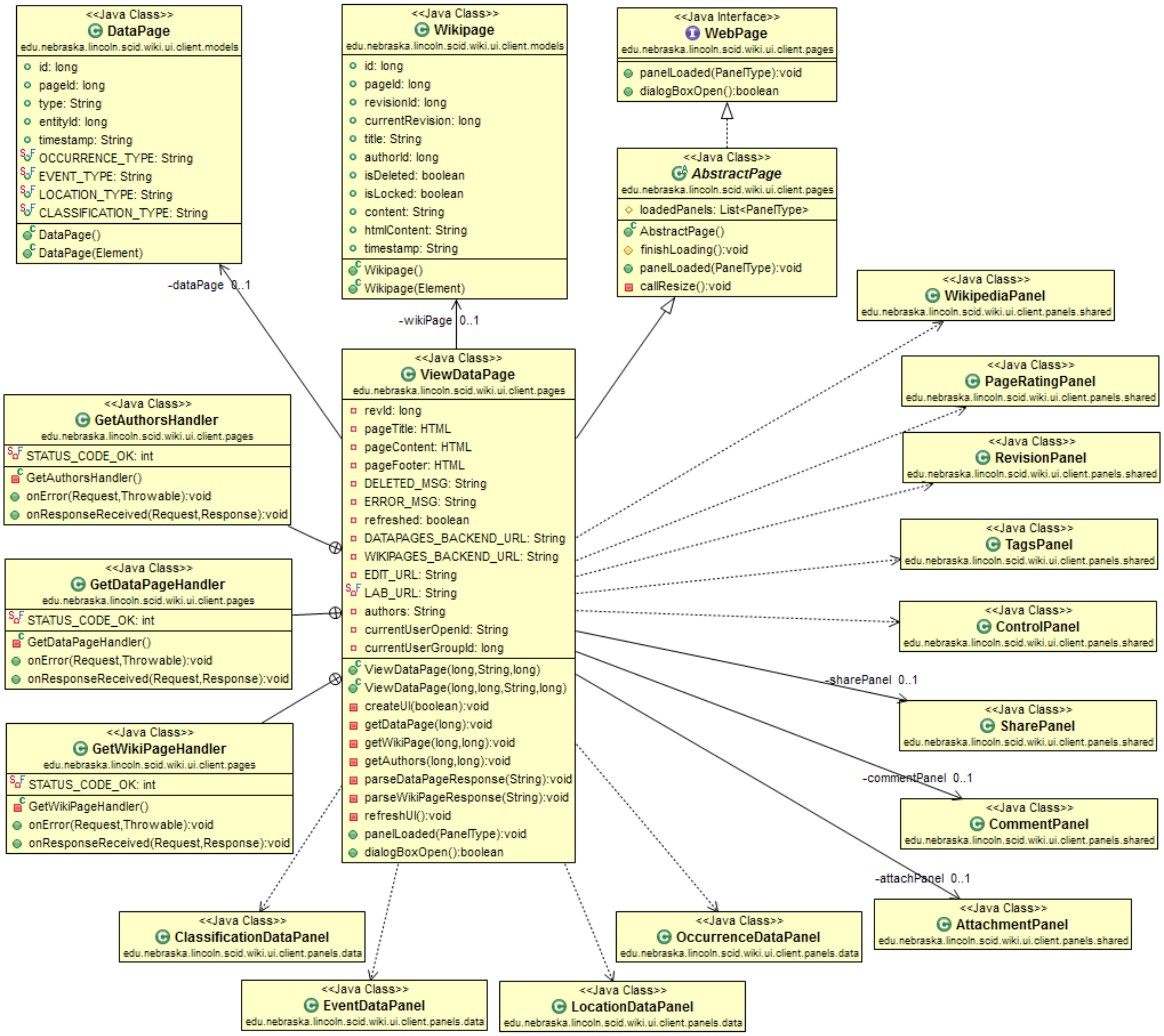

<<Java Class>>
DataPage
edu.nebraska.lincoln.scid.wiki.ui.client.models
id: long
pageId: long
type: String
entityId: long
timestamp: String
OCCURRENCE_TYPE: String
EVENT_TYPE: String
LOCATION_TYPE: String
CLASSIFICATION_TYPE: String
DataPage()
DataPage(Element)
<<Java Class>>
Wikipage
edu.nebraska.lincoln.scid.wiki.ui.client.models
id: long
pageId: long
revisionId: long
currentRevision: long
title: String
authorId: long
isDeleted: boolean
isLocked: boolean
content: String
htmlContent: String
timestamp: String
Wikipage()
Wikipage(Element)
<<Java Interface>>
WebPage
edu.nebraska.lincoln.scid.wiki.ui.client.pages
panelLoaded(PanelType):void
dialogBoxOpen():boolean
<<Java Class>>
AbstractPage
edu.nebraska.lincoln.scid.wiki.ui.client.pages
loadedPanels: List<PanelType>
AbstractPage()
finishLoading():void
panelLoaded(PanelType):void
callResize():void
<<Java Class>>
WikipediaPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.shared
-dataPage 0..1
-wikiPage 0..1
<<Java Class>>
ViewDataPage
edu.nebraska.lincoln.scid.wiki.ui.client.pages
revId: long
pageTitle: HTML
pageContent: HTML
pageFooter: HTML
DELETED_MSG: String
ERROR_MSG: String
refreshed: boolean
DATAPAGES_BACKEND_URL: String
WIKIPAGES_BACKEND_URL: String
EDIT_URL: String
LAB_URL: String
authors: String
currentUserOpenId: String
currentUserGroupId: long
ViewDataPage(long,String,long)
ViewDataPage(long,long,String,long)
createUI(boolean):void
getDataPage(long):void
getWikiPage(long,long):void
getAuthors(long,long):void
parseDataPageResponse(String):void
parseWikiPageResponse(String):void
refreshUI():void
panelLoaded(PanelType):void
dialogBoxOpen():boolean
<<Java Class>>
PageRatingPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.shared
<<Java Class>>
RevisionPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.shared
<<Java Class>>
TagsPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.shared
<<Java Class>>
ControlPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.shared
-sharePanel 0..1
<<Java Class>>
SharePanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.shared
-commentPanel 0..1
<<Java Class>>
CommentPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.shared
-attachPanel 0..1
<<Java Class>>
AttachmentPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.shared
<<Java Class>>
GetAuthorsHandler
edu.nebraska.lincoln.scid.wiki.ui.client.pages
STATUS_CODE_OK: int
GetAuthorsHandler()
onError(Request,Throwable):void
onResponseReceived(Request,Response):void
<<Java Class>>
GetDataPageHandler
edu.nebraska.lincoln.scid.wiki.ui.client.pages
STATUS_CODE_OK: int
GetDataPageHandler()
onError(Request,Throwable):void
onResponseReceived(Request,Response):void
<<Java Class>>
GetWikiPageHandler
edu.nebraska.lincoln.scid.wiki.ui.client.pages
STATUS_CODE_OK: int
GetWikiPageHandler()
onError(Request,Throwable):void
onResponseReceived(Request,Response):void
<<Java Class>>
ClassificationDataPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.data
<<Java Class>>
OccurrenceDataPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.data
<<Java Class>>
EventDataPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.data
<<Java Class>>
LocationDataPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.data

Figure 4.11: Data page (view mode) class diagram

Initialization of data pages also occurs in a manner similar to wiki pages. After the ViewDataPage class is instantiated with a page ID by the client, the client requests basic page information from the server and parses the server response for the main revision number and page type. The process differs slightly at this point – in addition to requesting the corresponding wiki page information, it also requests data page-specific information, such as the associated data entity ID and type. Based on these, one of the four data panel classes (Chapter 4.2.2.10) is instantiated to display the Biofinity data. The remainder of the process, i.e. creating the UI and initializing panels, is the same as for wiki pages.

Viewing a data page creates a tracking entry for the user, which includes the page and revision viewed and the timestamp for the action. View actions contribute directly towards determining the page's quality (Chapter 3.1.5) and indirectly towards determining pages relevant to the user (Chapter 3.1.1).

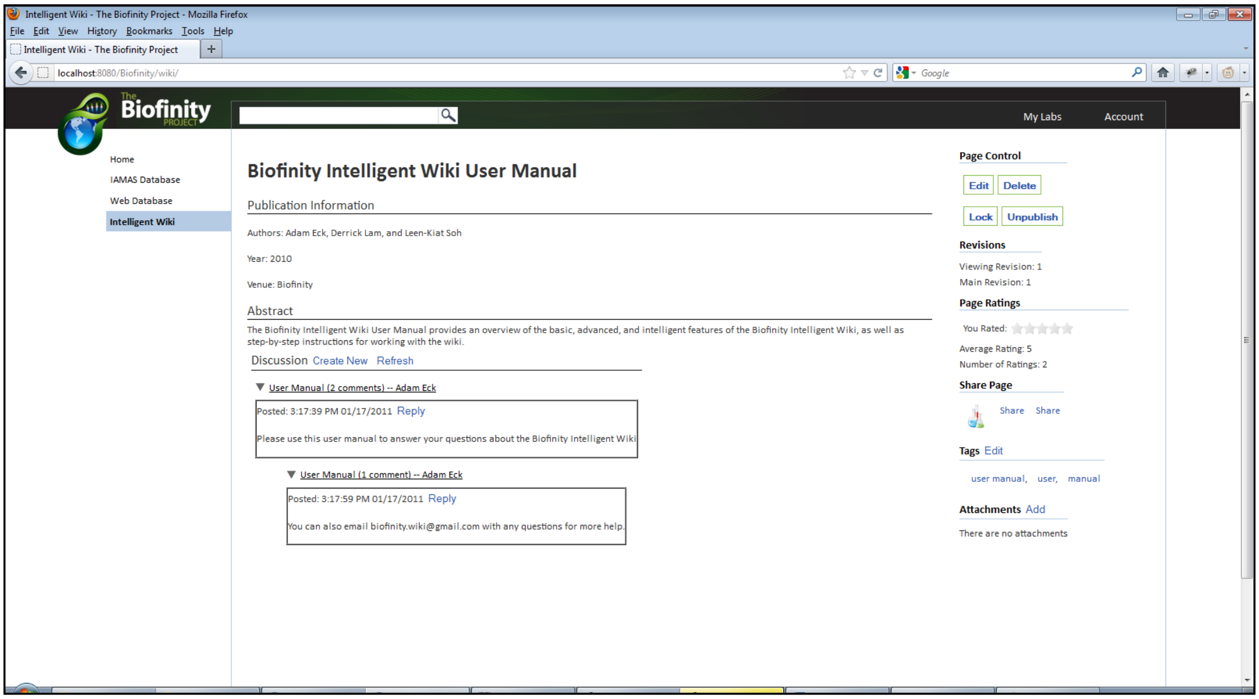
4.2.1.2.2 Edit Mode

The edit mode for data pages is the same as that of wiki pages, with the same wiki revision form, the same panels displayed, and the same event handling. That is, no unique edit functionality or information is introduced to the edit mode for data pages, and consequently, the edit mode for data pages reuses the EditWikiPage class in its entirety. For further information, please refer back to Chapter 4.2.1.1.

Editing a data page creates an entry similar to ones for wiki pages, which includes the page edited, the particular revision's ID, and the timestamp for the action. Edit actions contribute directly towards determining the page's quality (Chapter 3.1.6) and the user's expertise (Chapter 3.1.3), and contribute indirectly towards determining pages relevant to the user (Chapter 3.1.1).

4.2.1.3 Publication Page

4.2.1.3.1 View Mode


Intelligent Wiki - The Biofinity Project - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Intelligent Wiki - The Biofinity Project
localhost:8080/Biofinity/wiki/
Google
The Biofinity PROJECT
My Labs
Account
Home
IAMAS Database
Web Database
Intelligent Wiki
Page Control
Edit
Delete
Lock
Unpublish
Share
Share
Tags Edit
Attachments Add
Discussion Create New Refresh
Reply

Biofinity Intelligent Wiki User Manual

Publication Information

Authors: Adam Eck, Derrick Lam, and Leen-Kiat Soh

Year: 2010

Venue: Biofinity

Abstract

The Biofinity Intelligent Wiki User Manual provides an overview of the basic, advanced, and intelligent features of the Biofinity Intelligent Wiki, as well as step-by-step instructions for working with the wiki.

Discussion

▼ User Manual (2 comments) -- Adam Eck

Posted: 3:17:39 PM 01/17/2011 Reply

Please use this user manual to answer your questions about the Biofinity Intelligent Wiki

▼ User Manual (1 comment) -- Adam Eck

Posted: 3:17:59 PM 01/17/2011 Reply

You can also email biofinity.wiki@gmail.com with any questions for more help.

Revisions

Viewing Revision: 1

Main Revision: 1

Page Ratings

You Rated: ☆☆☆☆☆

Average Rating: 5

Number of Ratings: 2

Share Page

Tags

user manual, user, manual

Attachments

There are no attachments

Figure 4.12: Publication page (view mode) UI

Publication pages, an example of which is shown in Figure 4.9, are laid out in a manner similar to wiki and data pages, with their main body and comments panel in the primary content area and with panels for wiki features in the secondary content area. The main body consists of publication information, such as the publication authors, year of publication, and venue, as a well as an abstract for the publication. The publication itself can be uploaded to the page as an attachment or added as a URL in the abstract body.

As for panels attached to the page, publication pages generally have the same panels as its wiki page and data page counterparts: basic controls, revision control, page ratings, page sharing, tags, and attachments. Note that it does not have a Wikipedia recommendation panel.

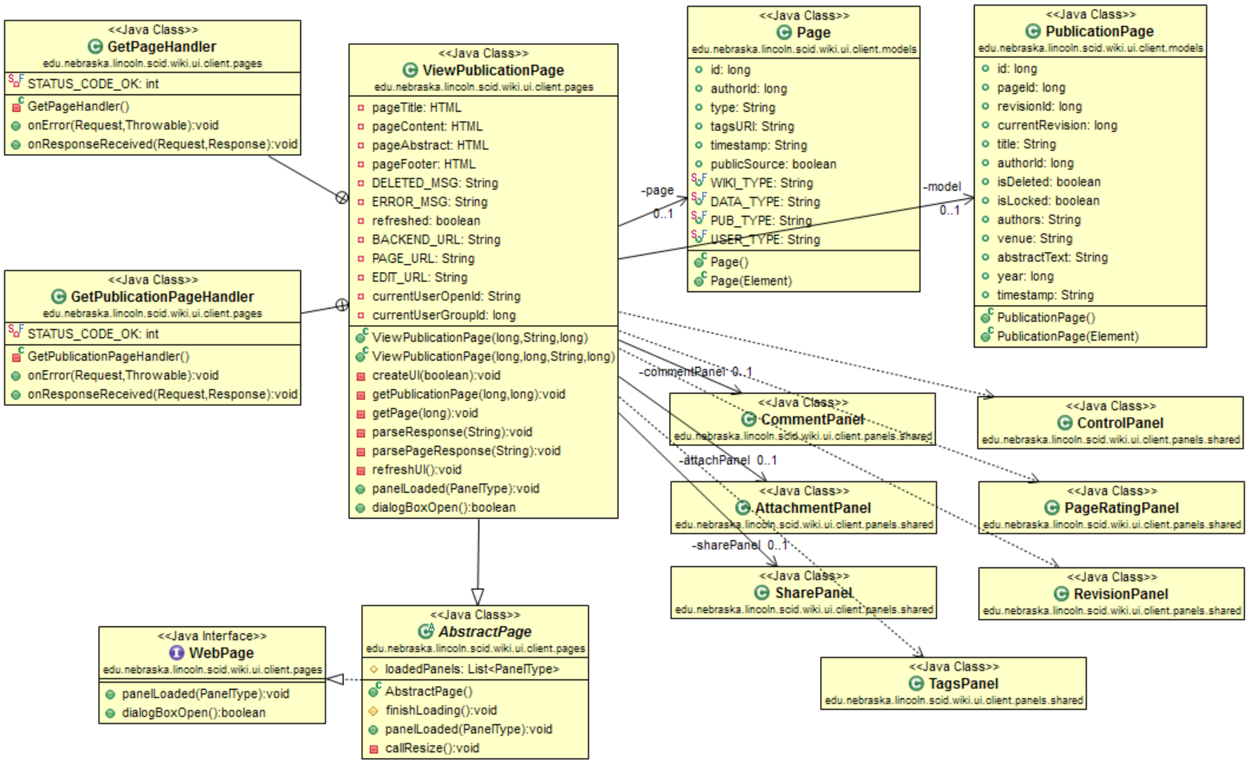
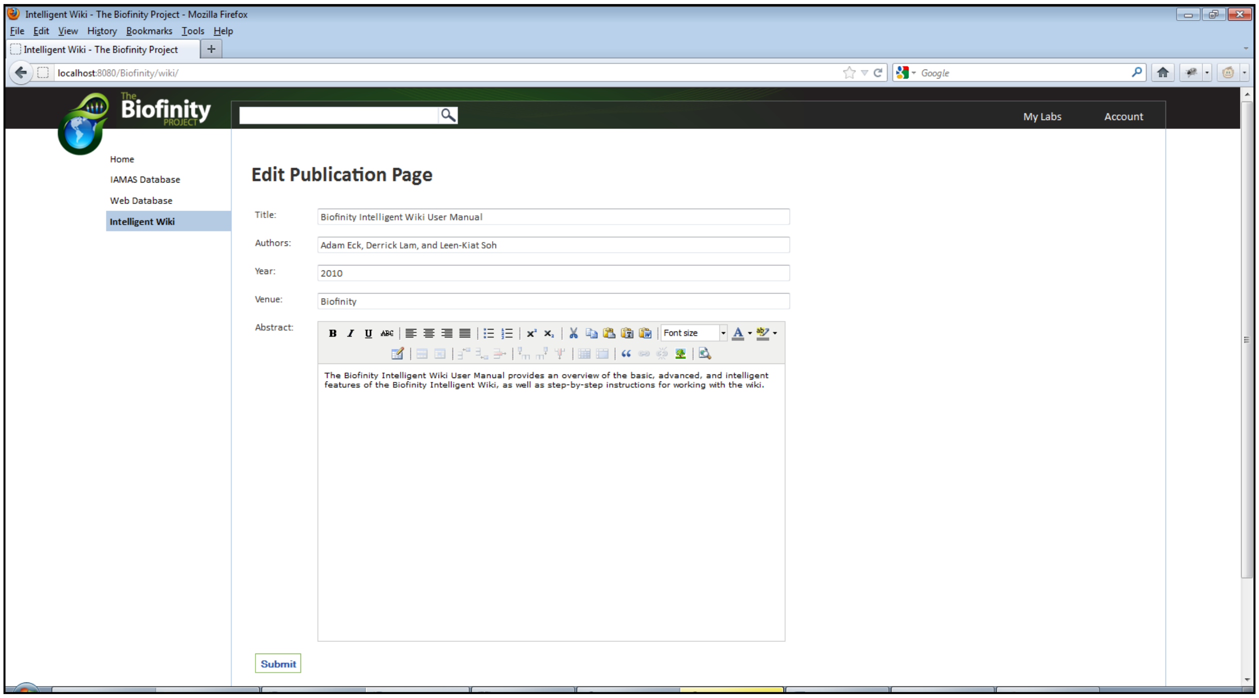


Figure 4.13: Publication page (view mode) class diagram

Tracking for viewing a publication page is similar to tracking for the previous types. An entry for it includes the page and revision viewed and the timestamp for the action. View actions contribute directly towards determining the page's quality (Chapter 3.1.5) and indirectly towards determining pages relevant to the user (Chapter 3.1.1).

4.2.1.3.2 Edit Mode


Intelligent Wiki - The Biofinity Project - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Intelligent Wiki - The Biofinity Project
localhost:8080/Biofinity/wiki/
Google
The Biofinity PROJECT
My Labs
Account
Home
IAMAS Database
Web Database
Intelligent Wiki
Font size
Submit

Edit Publication Page

Title: Biofinity Intelligent Wiki User Manual

Authors: Adam Eck, Derrick Lam, and Leen-Kiat Soh

Year: 2010

Venue: Biofinity

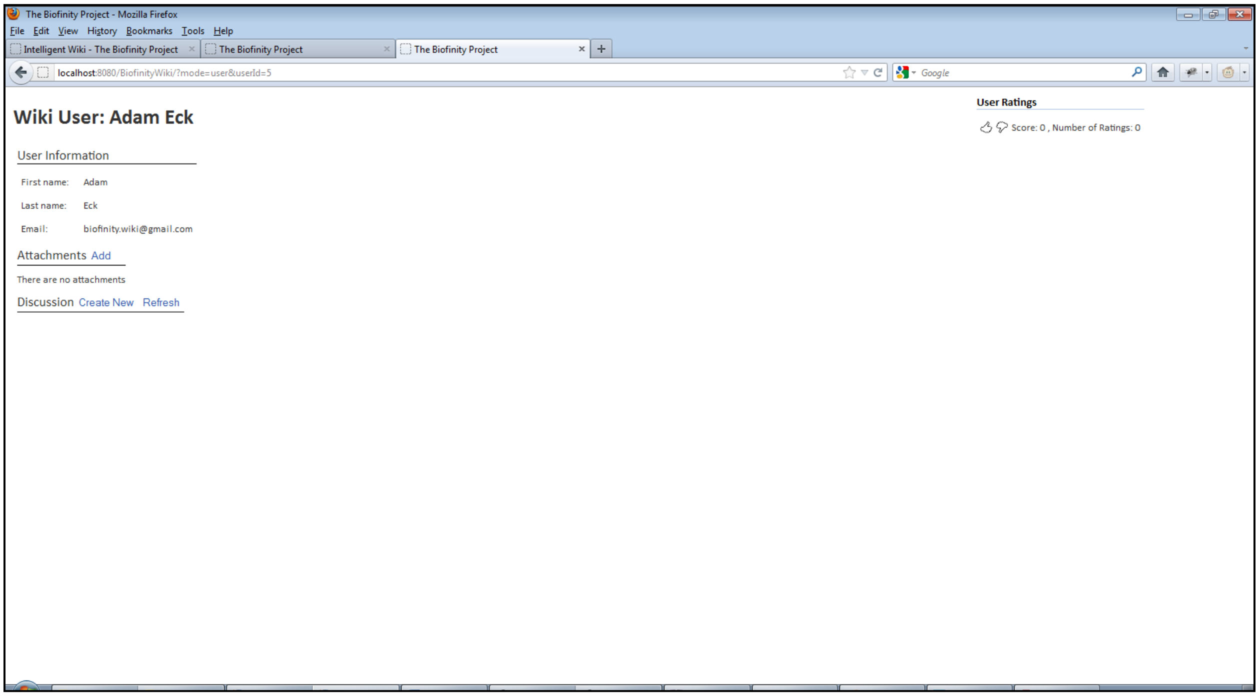
Abstract: The Biofinity Intelligent Wiki User Manual provides an overview of the basic, advanced, and intelligent features of the Biofinity Intelligent Wiki, as well as step-by-step instructions for working with the wiki.

Figure 4.14: Publication page (edit mode) UI

As can be seen in Figure 4.11, publication pages have their body contents split into multiple, specific fields rather than as one generic content field as with wiki and data pages. This is also reflected in the underlying model for the page in that each of these is its own attribute in the PublicationPage data class. Title, Authors, Year, and Venue all use a plain text box in the revision form. However, the Abstract field uses the TinyMCE rich text editor. Upon submission, the form's contents are sent to the server in XML format.

Editing a publication page creates a tracking entry for the user, which includes the page edited, the particular revision's ID, and the timestamp for the action. Edit actions contribute directly towards determining the page's quality (Chapter 3.1.6) and the user's expertise (Chapter 3.1.3), and contribute indirectly towards determining pages relevant to the user (Chapter 3.1.1).

4.2.1.4 User Page


The Biofinity Project - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Intelligent Wiki - The Biofinity Project
The Biofinity Project
The Biofinity Project
localhost:8080/BiofinityWiki/?mode=user&userId=5
Google

Wiki User: Adam Eck

User Information

First name: Adam

Last name: Eck

Email: biofinity.wiki@gmail.com

Attachments Add

There are no attachments

Discussion Create New Refresh

User Ratings

Score: 0 , Number of Ratings: 0

Figure 4.15: User page UI, as directly accessed outside of Biofinity UI

User pages, as pictured in Figure 4.12 when directly accessed outside of the Biofinity UI, are automatically generated whenever a new user is registered with the wiki, displaying the associated user's first name, last name, and e-mail address. This information is obtained from the Biofinity DB, which originally obtains the information from the associated user's Google OpenID profile. It should be noted that user pages have not been updated as frequently as the other page types and are disabled for the classroom deployment used to gather data.

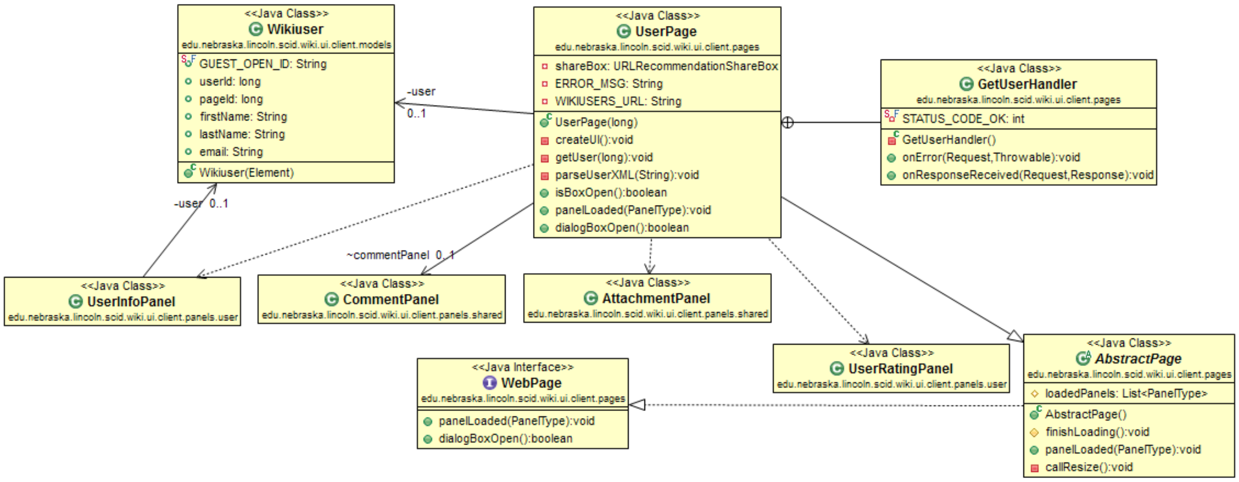

<<Java Class>>
Wikiuser
edu.nebraska.lincoln.scid.wiki.ui.client.models
GUEST_OPEN_ID: String
userId: long
pageId: long
firstName: String
lastName: String
email: String
Wikiuser(Element)
-user
0..1
<<Java Class>>
UserPage
edu.nebraska.lincoln.scid.wiki.ui.client.pages
shareBox: URLRecommendationShareBox
ERROR_MSG: String
WIKIUSERS_URL: String
UserPage(long)
createUI():void
getUser(long):void
parseUserXML(String):void
isBoxOpen():boolean
panelLoaded(PanelType):void
dialogBoxOpen():boolean
<<Java Class>>
GetUserHandler
edu.nebraska.lincoln.scid.wiki.ui.client.pages
STATUS_CODE_OK: int
GetUserHandler()
onError(Request,Throwable):void
onResponseReceived(Request,Response):void
-user 0..1
~commentPanel 0..1
<<Java Class>>
UserInfoPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.user
<<Java Class>>
CommentPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.shared
<<Java Class>>
AttachmentPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.shared
<<Java Class>>
UserRatingPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.user
<<Java Interface>>
WebPage
edu.nebraska.lincoln.scid.wiki.ui.client.pages
panelLoaded(PanelType):void
dialogBoxOpen():boolean
<<Java Class>>
AbstractPage
edu.nebraska.lincoln.scid.wiki.ui.client.pages
loadedPanels: List<PanelType>
AbstractPage()
finishLoading():void
panelLoaded(PanelType):void
callResize():void

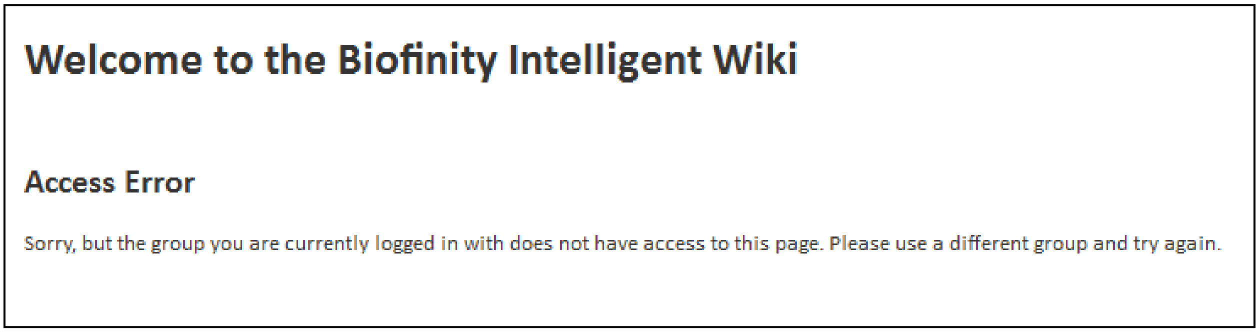
Figure 4.16: User page class diagram

When the user page is first loaded, it requests the user information from the server. After receiving a successful response, the GetUserHandler then triggers the process of parsing the wiki user information from XML to an instance of the Wikiuser data class. It should be noted that while user pages do have this underlying data class, their contents cannot be directly edited within the wiki. Rather, the user associated with the page will need to navigate to Account > Manage Account through the Biofinity interface to change it.

User pages currently have user information, attachments, comments, and user ratings panels associated with them. The user information panel is intuitive to include since it displays the main content of the page, as is the user ratings panel since it allows others to evaluate the user. As for the remaining two panels, the attachments panel is included for users to upload personal files (e.g. resume/CV), and the comments panel serves as a pseudo-messaging system.

The system does not currently track views to user pages.

4.2.1.5 Access Error Page



Welcome to the Biofinity Intelligent Wiki

Access Error

Sorry, but the group you are currently logged in with does not have access to this page. Please use a different group and try again.

Figure 4.17: Access error page UI

The Access Error Page is a fixed content page that is displayed when users attempt to access a private page that they do not have permissions for. This may also occur when the user is currently logged into the incorrect Biofinity lab. The page does not contain any panels and does not have any functionality associated with it. Consequently, views of this page are not tracked.

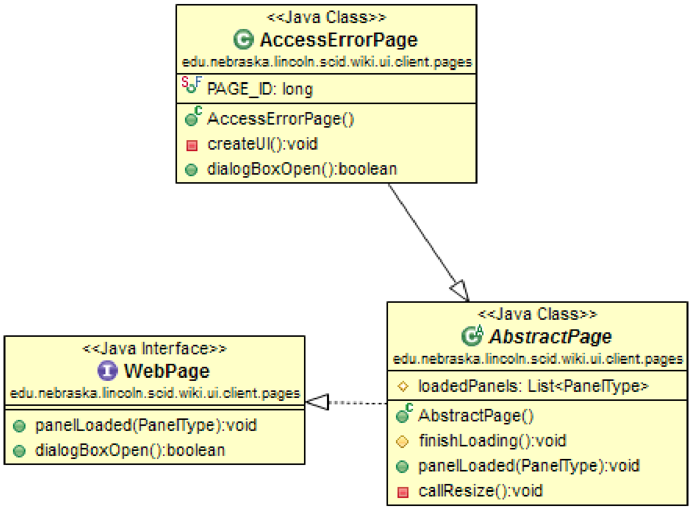
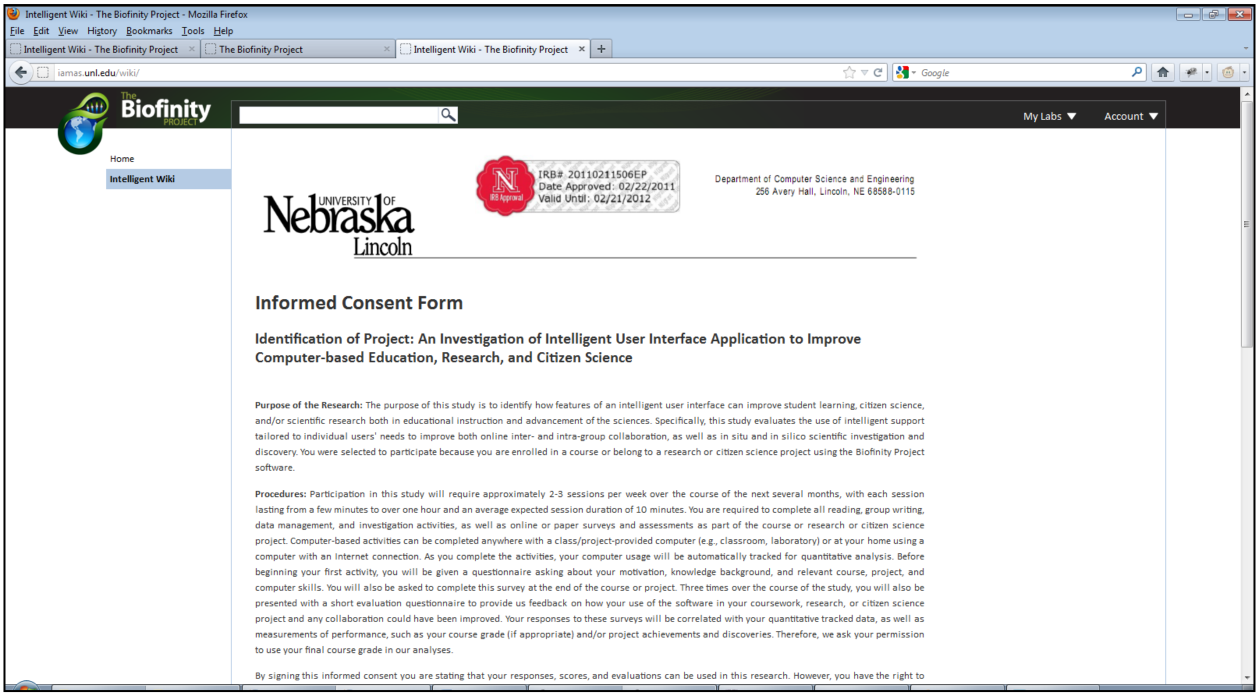


Figure 4.18: Access error page class diagram

4.2.1.6 Consent Page



Intelligent Wiki - The Biofinity Project - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Intelligent Wiki - The Biofinity Project
The Biofinity Project
Intelligent Wiki - The Biofinity Project
iamas.unl.edu/wiki/
Google
The Biofinity PROJECT
My Labs
Account
Home
Intelligent Wiki

IRB# 20110211506EP
Date Approved: 02/22/2011
Valid Until: 02/21/2012

Department of Computer Science and Engineering
256 Avery Hall, Lincoln, NE 68588-0115

UNIVERSITY OF Nebraska Lincoln

Informed Consent Form

Identification of Project: An Investigation of Intelligent User Interface Application to Improve Computer-based Education, Research, and Citizen Science

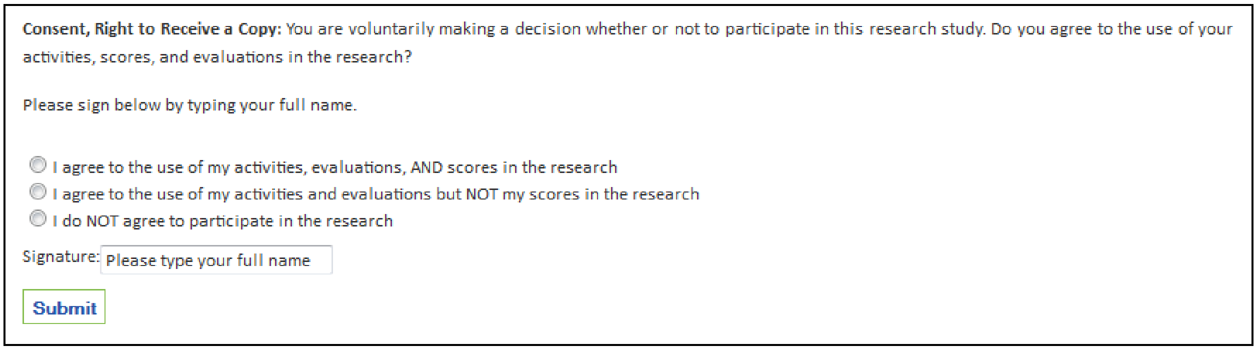
Purpose of the Research: The purpose of this study is to identify how features of an intelligent user interface can improve student learning, citizen science, and/or scientific research both in educational instruction and advancement of the sciences. Specifically, this study evaluates the use of intelligent support tailored to individual users' needs to improve both online inter- and intra-group collaboration, as well as in situ and in silico scientific investigation and discovery. You were selected to participate because you are enrolled in a course or belong to a research or citizen science project using the Biofinity Project software.

Procedures: Participation in this study will require approximately 2-3 sessions per week over the course of the next several months, with each session lasting from a few minutes to over one hour and an average expected session duration of 10 minutes. You are required to complete all reading, group writing, data management, and investigation activities, as well as online or paper surveys and assessments as part of the course or research or citizen science project. Computer-based activities can be completed anywhere with a class/project-provided computer (e.g., classroom, laboratory) or at your home using a computer with an Internet connection. As you complete the activities, your computer usage will be automatically tracked for quantitative analysis. Before beginning your first activity, you will be given a questionnaire asking about your motivation, knowledge background, and relevant course, project, and computer skills. You will also be asked to complete this survey at the end of the course or project. Three times over the course of the study, you will also be presented with a short evaluation questionnaire to provide us feedback on how your use of the software in your coursework, research, or citizen science project and any collaboration could have been improved. Your responses to these surveys will be correlated with your quantitative tracked data, as well as measurements of performance, such as your course grade (if appropriate) and/or project achievements and discoveries. Therefore, we ask your permission to use your final course grade in our analyses.

By signing this informed consent you are stating that your responses, scores, and evaluations can be used in this research. However, you have the right to

Figure 4.19: Consent page UI

The Consent page, as seen in Figure 4.16, displays the full text of the IRB Informed Consent Form, and it is used to obtain explicit permission from intelligent wiki users to include their tracking data, evaluations of the system, and/or their scores for the course (when applicable) for this thesis and other future studies. When it is first loaded, it checks to see whether the user has already filled out the consent form. If the user has not, then the form shown in Figure 4.17 is populated at the bottom of the page. Otherwise, the form is hidden.



Consent, Right to Receive a Copy: You are voluntarily making a decision whether or not to participate in this research study. Do you agree to the use of your activities, scores, and evaluations in the research?

Please sign below by typing your full name.

○ I agree to the use of my activities, evaluations, AND scores in the research
○ I agree to the use of my activities and evaluations but NOT my scores in the research
○ I do NOT agree to participate in the research

Signature: Please type your full name

Figure 4.20: Consent form UI

While the consent page does not contain any panels or an associated data class, its source file defines two event handlers: GetConsentHandler and PutConsentHandler. These handle the server response when querying for the user's consent status and submitting the consent form, respectively.

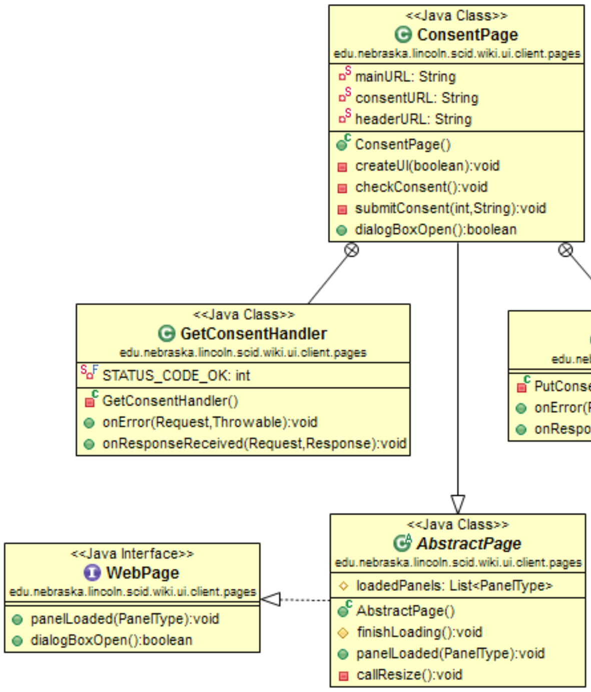
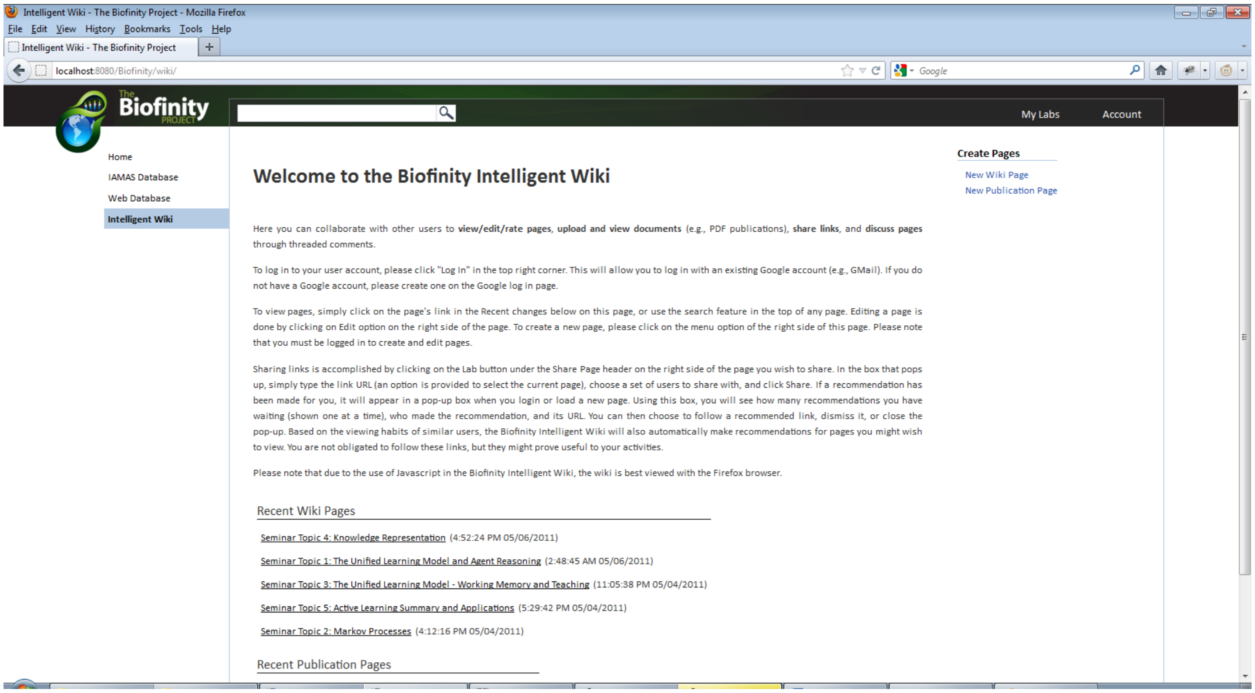

<<Java Class>>
ConsentPage
edu.nebraska.lincoln.scid.wiki.ui.client.pages
mainURL: String
consentURL: String
headerURL: String
ConsentPage()
createUI(boolean):void
checkConsent():void
submitConsent(int,String):void
dialogBoxOpen():boolean
<<Java Class>>
GetConsentHandler
edu.nebraska.lincoln.scid.wiki.ui.client.pages
STATUS_CODE_OK: int
GetConsentHandler()
onError(Request,Throwable):void
onResponseReceived(Request,Response):void
<<Java Class>>
PutConsentHandler
edu.nebraska.lincoln.scid.wiki.ui.client.pages
PutConsentHandler()
onError(Request,Throwable):void
onResponseReceived(Request,Response):void
<<Java Interface>>
WebPage
edu.nebraska.lincoln.scid.wiki.ui.client.pages
panelLoaded(PanelType):void
dialogBoxOpen():boolean
<<Java Class>>
AbstractPage
edu.nebraska.lincoln.scid.wiki.ui.client.pages
loadedPanels: List<PanelType>
AbstractPage()
finishLoading():void
panelLoaded(PanelType):void
callResize():void

Figure 4.21: Consent page class diagram

Since user consent does not (and should not) impact a user's experience with the intelligent wiki, additional tracking is not performed for this page.

4.2.1.7 Main Page


Intelligent Wiki - The Biofinity Project - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Intelligent Wiki - The Biofinity Project
localhost:8080/Biofinity/wiki/
Google
The Biofinity PROJECT
My Labs
Account
Home
IAMAS Database
Web Database
Intelligent Wiki

Welcome to the Biofinity Intelligent Wiki

Here you can collaborate with other users to **view/edit/rate pages**, **upload and view documents** (e.g., PDF publications), **share links**, and **discuss pages** through threaded comments.

To log in to your user account, please click "Log In" in the top right corner. This will allow you to log in with an existing Google account (e.g., GMail). If you do not have a Google account, please create one on the Google log in page.

To view pages, simply click on the page's link in the Recent changes below on this page, or use the search feature in the top of any page. Editing a page is done by clicking on Edit option on the right side of the page. To create a new page, please click on the menu option of the right side of this page. Please note that you must be logged in to create and edit pages.

Sharing links is accomplished by clicking on the Lab button under the Share Page header on the right side of the page you wish to share. In the box that pops up, simply type the link URL (an option is provided to select the current page), choose a set of users to share with, and click Share. If a recommendation has been made for you, it will appear in a pop-up box when you login or load a new page. Using this box, you will see how many recommendations you have waiting (shown one at a time), who made the recommendation, and its URL. You can then choose to follow a recommended link, dismiss it, or close the pop-up. Based on the viewing habits of similar users, the Biofinity Intelligent Wiki will also automatically make recommendations for pages you might wish to view. You are not obligated to follow these links, but they might prove useful to your activities.

Please note that due to the use of Javascript in the Biofinity Intelligent Wiki, the wiki is best viewed with the Firefox browser.

Recent Wiki Pages

Seminar Topic 4: Knowledge Representation (4:52:24 PM 05/06/2011)

Seminar Topic 1: The Unified Learning Model and Agent Reasoning (2:48:45 AM 05/06/2011)

Seminar Topic 3: The Unified Learning Model - Working Memory and Teaching (11:05:38 PM 05/04/2011)

Seminar Topic 5: Active Learning Summary and Applications (5:29:42 PM 05/04/2011)

Seminar Topic 2: Markov Processes (4:12:16 PM 05/04/2011)

Recent Publication Pages

Create Pages

New Wiki Page

New Publication Page

Figure 4.22: Main page UI

The main page (shown in Figure 4.19) is the first page seen by the user when the “Intelligent Wiki” link in the navigation sidebar is clicked, and its primary content consists of a hard-coded welcome message to the user along with brief instructions on how to use it. Since it consists of non-editable content, it does not have any underlying data classes and has very few features for the user to make use of. Consequently, few panels are attached to the page: only panels for creating pages, displaying pages recently edited by other group members, and obtaining user consent are attached to the main page. Details for each of the panels can be found in Chapter 4.2.2.

The main page itself does not have any tracking associated with it.

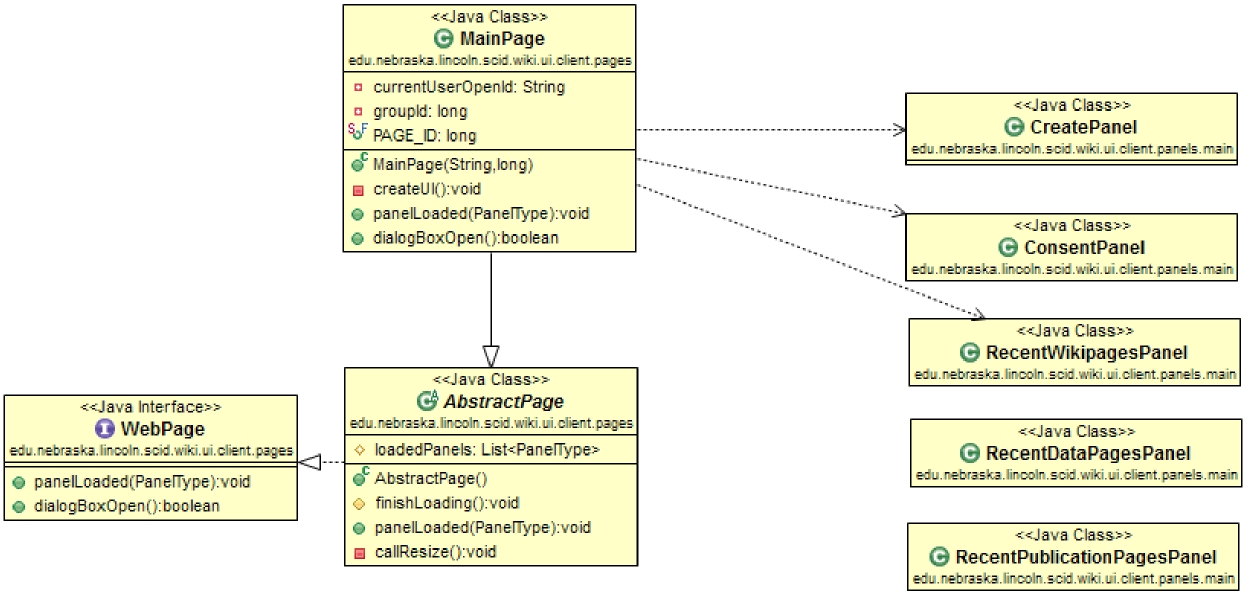
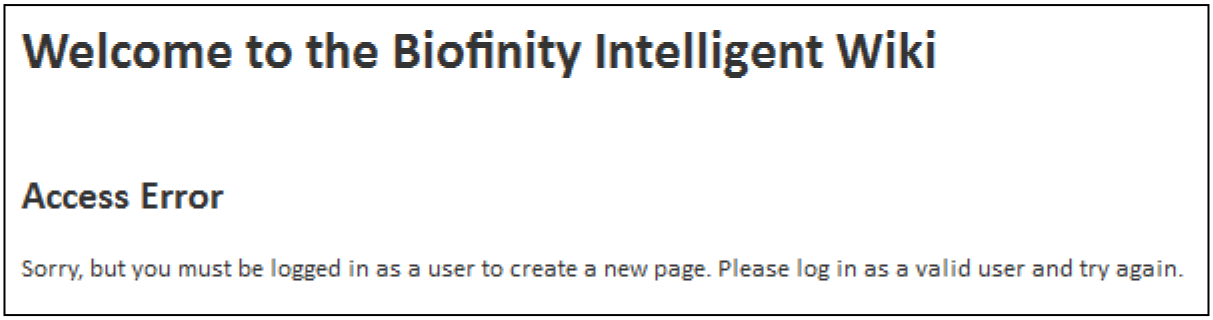

<<Java Class>>
MainPage
edu.nebraska.lincoln.scid.wiki.ui.client.pages
currentUserOpenId: String
groupId: long
PAGE_ID: long
MainPage(String,long)
createUI():void
panelLoaded(PanelType):void
dialogBoxOpen():boolean
<<Java Class>>
CreatePanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.main
<<Java Class>>
ConsentPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.main
<<Java Class>>
RecentWikipagesPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.main
<<Java Class>>
RecentDataPagesPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.main
<<Java Class>>
RecentPublicationPagesPanel
edu.nebraska.lincoln.scid.wiki.ui.client.panels.main
<<Java Interface>>
WebPage
edu.nebraska.lincoln.scid.wiki.ui.client.pages
panelLoaded(PanelType):void
dialogBoxOpen():boolean
<<Java Class>>
AbstractPage
edu.nebraska.lincoln.scid.wiki.ui.client.pages
loadedPanels: List<PanelType>
AbstractPage()
finishLoading():void
panelLoaded(PanelType):void
callResize():void

Figure 4.23: Main page class diagram

4.2.1.8 Not Logged In Page



Welcome to the Biofinity Intelligent Wiki

Access Error

Sorry, but you must be logged in as a user to create a new page. Please log in as a valid user and try again.

Figure 4.24: Not logged in page UI

Similar to the Access Error Page, the Not Logged In page simply displays an error and has no panels, data classes, or functionality associated with it. It occurs when a user attempts to create a new wiki or publication page when not logged into Biofinity, and no tracking entries are kept for viewing this page.

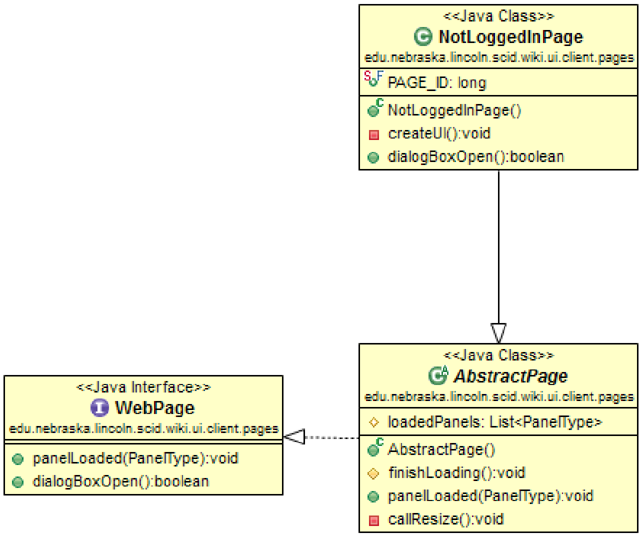
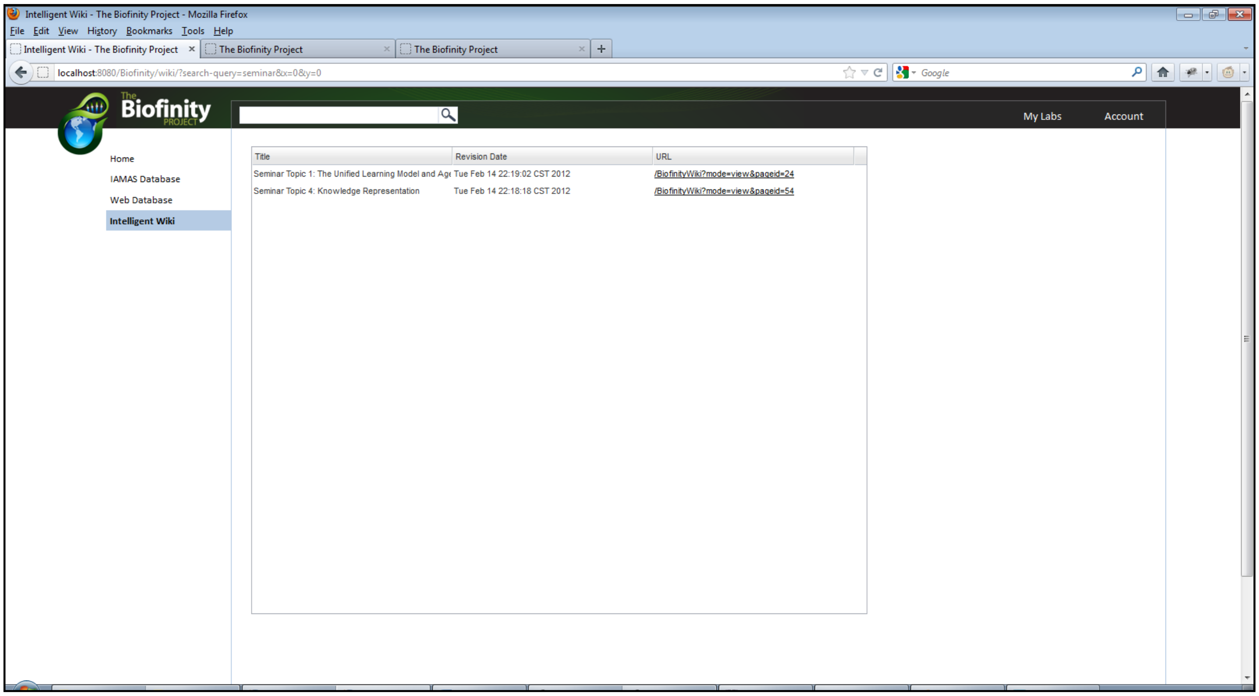


Figure 4.25: Not logged in page class diagram

4.2.1.9 Search Results Page


Intelligent Wiki - The Biofinity Project - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Intelligent Wiki - The Biofinity Project
The Biofinity Project
The Biofinity Project
localhost:8080/Biofinity/wiki/?search-query=seminar&x=0&y=0
Google
The Biofinity PROJECT
My Labs
Account
Home
IAMAS Database
Web Database
Intelligent Wiki

Title	Revision Date	URL
Seminar Topic 1: The Unified Learning Model and Age	Tue Feb 14 22:19:02 CST 2012	/BiofinityWiki?mode=view&pageid=24
Seminar Topic 4: Knowledge Representation	Tue Feb 14 22:18:18 CST 2012	/BiofinityWiki?mode=view&pageid=54

Figure 4.26: Search results page UI

The search results page (seen in Figure 4.23) is displayed after the user performs a keyword search via the search bar in the header while in the wiki section of the Biofinity system. The results are displayed in a table along with details such as the page's title, its date of last revision, and its URL. The results are originally ordered by their term frequency-inverse document frequency[7] (*tf-idf*) values as determined by the Lucene indexing and search engine, and the user can click the column headers to reorder the results. Since the SmartGWT ListGrid used to display the results can use XML data directly, the ResultsPage class does not parse the results into an equivalent Java object and is thus not associated with a data object. The results page also does not have any panels attached to it.

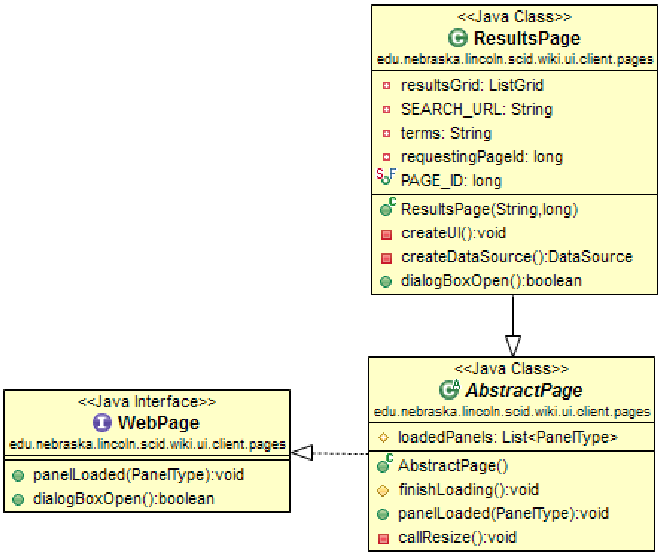


Figure 4.27: Results page class diagram

While a tracking entry isn't kept for viewing the search results page, the system does track the keywords used and the results returned for the search. These entries can contribute towards determining a page's topics, tags, discipline, and keywords, which then factors into determining a page's relevance to a user (Chapter 3.1.1).

4.2.2 Panels

As previously described, we encapsulate the various "recyclable" UI elements and interactive features and functionality into *panels*. The panels included on a page are

[7] Spärck Jones, Karen (1972). "A statistical interpretation of term specificity and its application in retrieval". *Journal of Documentation* **28** (1): 11–21.

initialized and added to the proper HTML content element during the execution of the page's "createUI" methods, i.e. after the client receives primary page-specific content from the server. Each panel then queries the wiki server separately and finishes loading once it receives any needed data, i.e. the panels can be loaded and displayed asynchronously.

Similar to how pages have an interface and abstract superclass, panels have their equivalents in the *Panel* interface and the *AbstractPanel* superclass implementing it.

4.2.2.1 Common Controls Panel

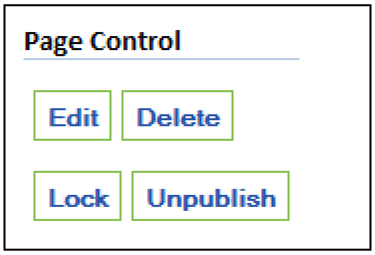


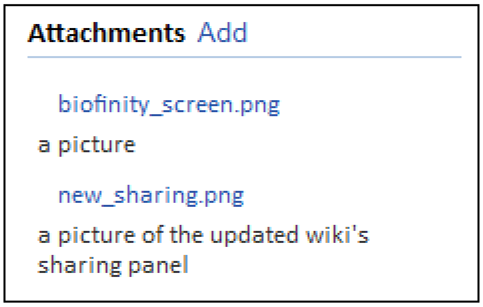
Figure 4.28: Common controls panel UI

The Common Controls panel in Figure 4.25 provides users with basic page management tools, including the options to edit the page, delete the currently viewed revision, lock the page from further editing, and publish the page to the public. Clicking the Edit button forwards the user to the page type's corresponding edit mode and pre-populates its fields with that of the currently revision. Clicking the Delete button marks the current revision as deleted and cannot be undone via the wiki interface. Additionally, the entire page is flagged as deleted if all of its revisions have been deleted. Unlike deletion, page locking and publishing can be undone, and the corresponding buttons will change to "Unlock" or "Unpublish" to revert the state.

Each of the features provided by this panel are tracked by the system, and the created tracking entry marks the action performed (one of edit, delete, lock, unlock, publish, or unpublish), the user performing the action, when it was performed, and the page and revision acted upon. The edit and delete actions in particular may directly impact the page's quality (Chapter 3.1.6) and the user's expertise (Chapter 3.1.3), and contribute indirectly towards determining pages relevant to the user (Chapter 3.1.1).

This panel only appears on the view modes for pages with editable contents, i.e. wiki pages, data pages, and publication pages.

4.2.2.2 Attachments Panel



Attachments Add

biofinity_screen.png

a picture

new_sharing.png

a picture of the updated wiki's sharing panel

Figure 4.29: Attachments panel UI

The attachments panel, seen in Figure 4.26, is used to upload and associate files to pages that the panel appears on, and to display download links to the files that have already been attached to the page. For image attachments in particular, hovering the mouse cursor over the hyperlink displays a thumbnail preview of the image, as seen in Figure 4.27 below.

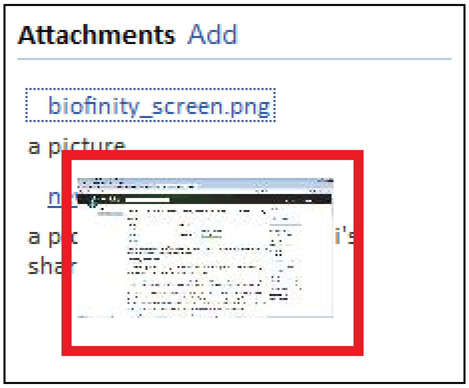


Figure 4.30: Attachments panel UI, with image preview highlighted

When the panel is instantiated, it requests the server for attachments currently attached to its parent page. The server then responds with the files' names, download URLs, and, when appropriate, thumbnail images. After receiving this response, the panel populates its attachments list. Note that the binary data for the (original) files are not transmitted during this initialization.

To upload a file, the user clicks the "Add" link in the panel header. The following pop-up appears for the user to select the target file and enter the caption text to display with it.

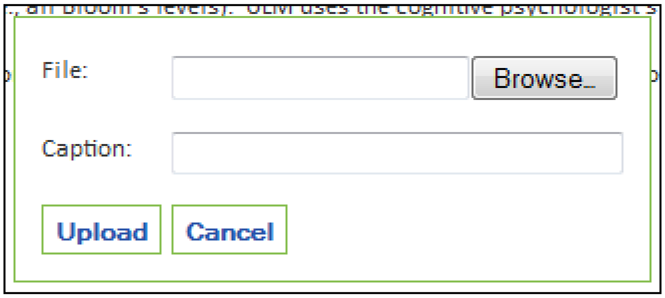


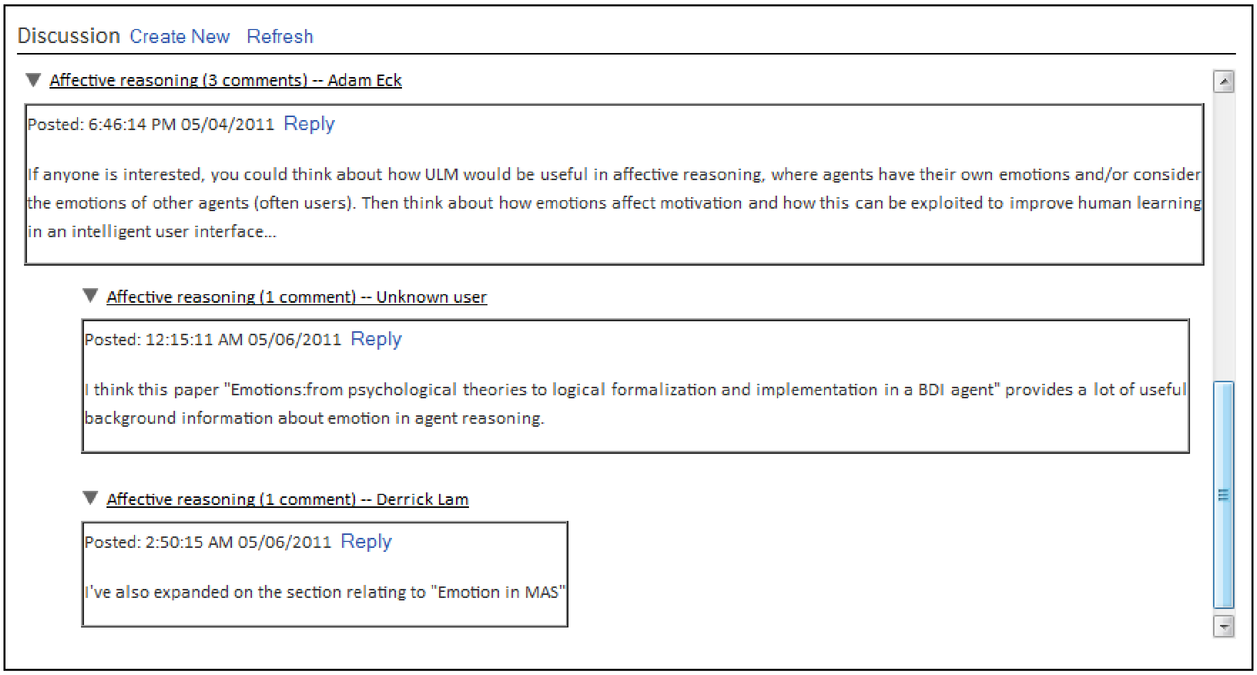
Figure 4.31 Attachment upload dialog box

To download a file, the user clicks the corresponding download hyperlink and specifies the save location via a browser-specific dialog box.

The system creates a tracking entry for uploads and downloads, which includes the ID of the attachment uploaded and the page that the attachment is associated with. Entries created from these actions are currently not used in recommendations.

This panel can be found on both the view and edit modes of pages with editable content, i.e. wiki, data, and publication pages, as well as on user pages.

4.2.2.3 Comments/Discussion Panel



▼ Affective reasoning (3 comments) -- Adam Eck

Posted: 6:46:14 PM 05/04/2011 Reply

If anyone is interested, you could think about how ULM would be useful in affective reasoning, where agents have their own emotions and/or consider the emotions of other agents (often users). Then think about how emotions affect motivation and how this can be exploited to improve human learning in an intelligent user interface...

▼ Affective reasoning (1 comment) -- Unknown user

Posted: 12:15:11 AM 05/06/2011 Reply

I think this paper "Emotions:from psychological theories to logical formalization and implementation in a BDI agent" provides a lot of useful background information about emotion in agent reasoning.

▼ Affective reasoning (1 comment) -- Derrick Lam

Posted: 2:50:15 AM 05/06/2011 Reply

I've also expanded on the section relating to "Emotion in MAS"

Figure 4.32: Comments panel UI, one root-level comment with two children comments

Unlike most of the other panels, the comments panel (as seen in Figure 4.29) is located in the primary content area due to its larger size, and it enables users to carry out threaded conversations. The comments are arranged in a tree-like manner, with comments at the root level being considered as the “beginning” of the threads. Direct replies to existing comments add “children” comments to them. Each comment may have any number of children comments, but only one parent and one root. This hierarchical nature of the comments is reflected via indentation – root-level comments are leftmost, and subordinate comments are indented at one level further than their immediate parent. Particular paths through the comments tree are thus particular "threads" of conversation.

The following features are provided for the comments panel: creating a new comment thread, replying to a particular comment, manual refreshing of comments, and collapsing of threads. New comment threads, i.e. new top-level comments, can be made by clicking the "Create New" link and filling in the dialog box (Figure 4.30) with the topic and comment body for the thread. Clicking the "Reply" link for a particular comment pops up a similar dialog box with the topic field pre-populated. Comment trees and sub-trees can be selectively collapsed and re-expanded by clicking the triangle icon next to the comment topic.

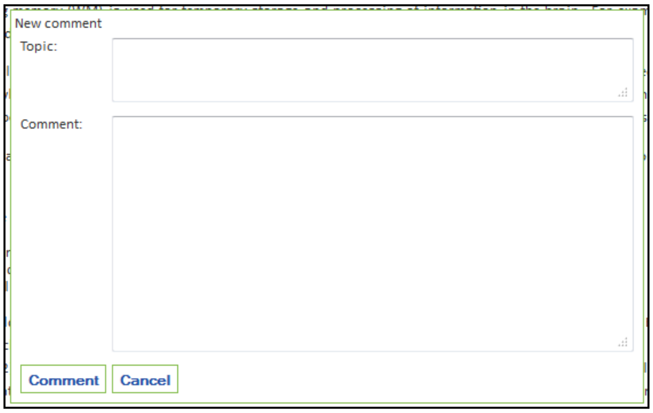


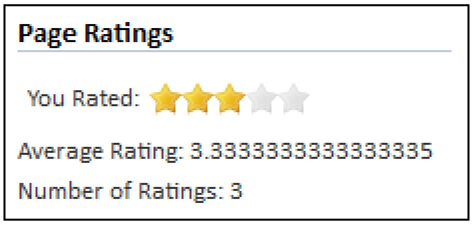
Figure 4.33: Create comment dialog box

Submission of a comment refreshes the contents of the panel without reloading the entire page. However, it should be noted that it will not refresh the panel for other users in real-time. Instead, users may click the "Refresh" link to manually update its contents. We opted for this approach since an automatic refresh requires periodic requests to the server or a persistent listener for comment-related server responses. Both require additional computational resources multiplied with more active wiki users, and the latter does not fit within the paradigm of RESTful services.

Tracking entries are made when a user comments upon the page or participates in existing discussions. While these currently do not impact our recommendations generated, these play a role in user and page modeling.

The comments panel can be found on both the view and edit modes of wiki, data, publication, and user pages.

4.2.2.4 Page Ratings Panel



Page Ratings

You Rated: ★★★☆☆

Average Rating: 3.3333333333333335

Number of Ratings: 3

Figure 4.34: Page ratings panel

The page ratings panel in Figure 4.31 appears on wiki, data, and publication pages, and it enables users to evaluate pages on a 1-5 star scale. The user can provide a rating by clicking the star rating to give it on the widget. This rating persists indefinitely and is loaded each time the user views the page. Should the user choose to re-evaluate the page, the user can give a different rating to overwrite the old one. That is, *a user can only contribute towards page's rating once per page*. By implementing the rating system in this manner, it prevents the practice of "spamming" ratings to guide the page's overall rating towards a particular score. While influencing page score is still possible through the use of additional accounts, the work involved in setting them up may discourage potential violators from doing so.

The panel also displays the page's average rating and the number of users contributing towards that score. By publicly displaying both, users can determine the relative relevance of a page's average score for themselves.

Tracking is performed when a user rates a page, and a user's rating can play a large impact on the recommendations made. First, they influence the page's relevance to the user since they indirectly indicate the user's inclination to the topics relevant to the page (Chapter 3.1.1). The ratings also impact the overall perceived quality of the page (Chapter 3.1.2).

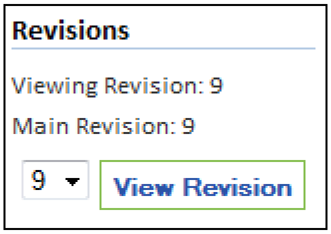
4.2.2.5 Page Stats Panel

The page stats panel was originally developed to display simple usage statistics for wiki, data, publication, and user pages. Upon initialization, the panel generates a request to the server, which then performs an online calculation of the requested information based on the tracking information collected. These include:

- Number of views
- Number of edits
- Number of files uploaded to it
- Number of comments
- Number of ratings
- Rating score
- Time elapsed since last view
- Time elapsed since last edit

It has since been removed after the wiki's prototyping stages and is disabled for classroom deployments. Thus, no tracking is performed for this panel since these stats cannot be viewed.

4.2.2.6 Revision Panel



Revisions

Viewing Revision: 9

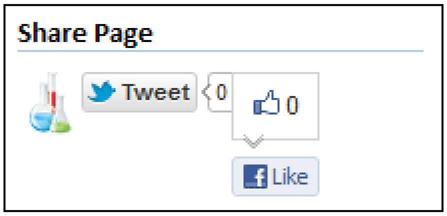
Main Revision: 9

Figure 4.35: Revision panel

The revisions panel (Figure 4.32) appears on the view mode of pages with editable content, such as wiki, data, and publication pages. Through this, users can view specific non-deleted revisions of a page as well as set another revision as the main one, i.e. the one first loaded when a specific revision for a page isn't specified. Further, this panel can be used to navigate to a specific revision for deletion. Selecting a revision in the dropdown box and clicking the "View Revision" button will reload the page with the contents of that particular revision.

The system creates tracking entries when the user sets the main revision for the page and when viewing a different revision. The revision changing aspect has no impact on modeling or recommendations, while the impact of the viewing aspect is as described previously for the appropriate page in Chapter 4.2.1.

4.2.2.7 Sharing Panel

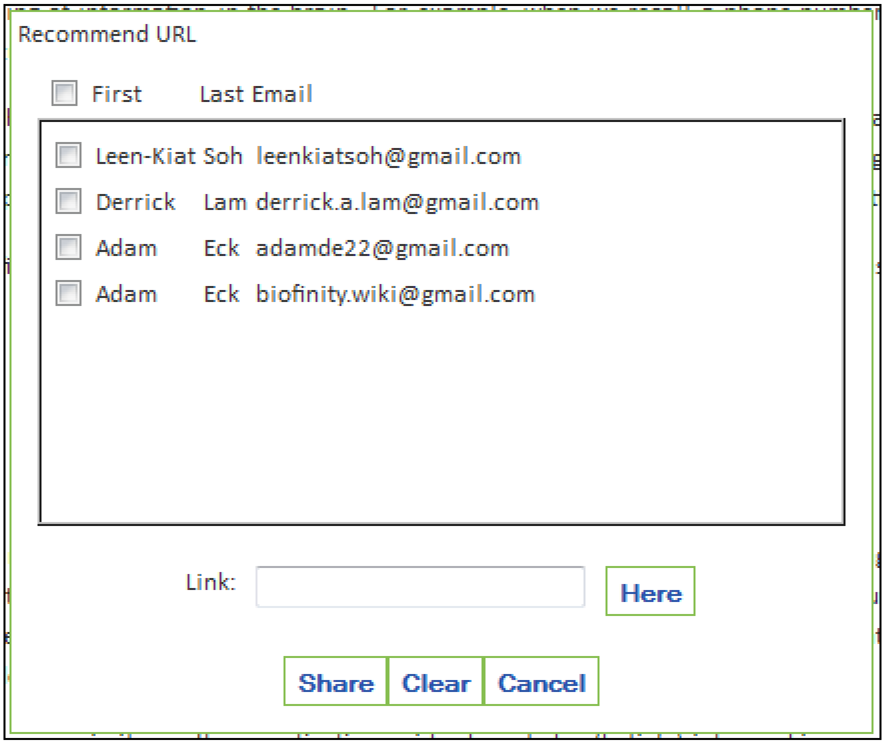


Share Page

Figure 4.36: Sharing panel

The Sharing Panel as seen in Figure 4.33 is used to share or send a hyperlink: 1) to another wiki user via the intra-wiki recommendation framework, 2) to the user's Twitter followers via a Twitter post to the user's account, or 3) to the user's Facebook friends via a post to his/her wall. Note that people following the shared link may not be able to access the recommended page if they do not have sufficient permission to do so.

To share a URL with another wiki user, the user first clicks on the MyLab icon (the flasks) and then enters the URL to share and the group peers to share them with in the subsequent pop-up dialog box (Figure 4.34, *URLRecommendationShareBox*). Peers are displayed with first name, last name, and e-mail address to help distinguish them from one another, i.e. when two users have the same first and last names. For convenience, a "Here" button has been included to enable the user to quickly obtain the URL of the currently-viewed page.



Recommend URL

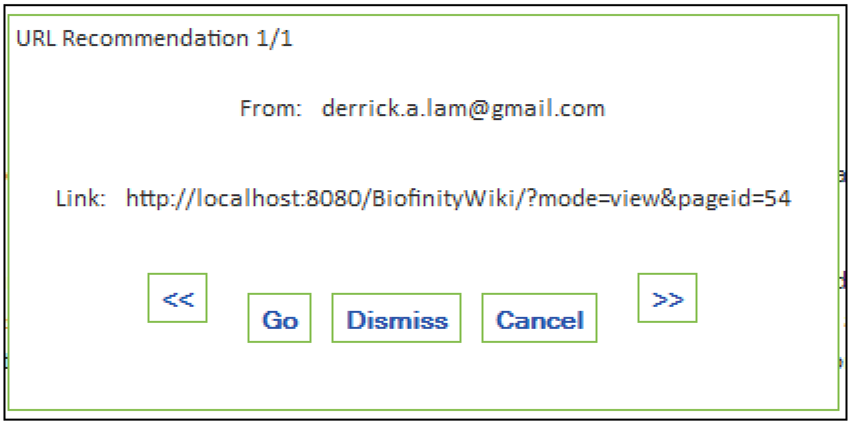
□	First	Last	Email
□	Leen-Kiat	Soh	leenkiatsoh@gmail.com
□	Derrick	Lam	derrick.a.lam@gmail.com
□	Adam	Eck	adamde22@gmail.com
□	Adam	Eck	biofinity.wiki@gmail.com

Figure 4.37: Intra-wiki URL recommendation dialog box

When a user receives a recommendation, a pop-up dialog (*URLRecommendationAlertBox*, Figure 4.35) is displayed to them in real time, provided that the receiving user is not currently "busy" with any work. We approximate this with the following set of rules to govern its display:

- If the alert box is currently showing, do not display it again
- If the user is currently editing a page (i.e. in edit mode), do not display the alert box
- If dialog boxes are currently open (e.g. when sharing pages with other users, when uploading an attachment, when making a comment), do not display the alert box
- If the alert box has been displayed within the last ten minutes and is currently closed, do not display the alert box

- If the above conditions are avoided and the user has recommendations that are not yet viewed or dismissed, display the alert box



URL Recommendation 1/1

From: derrick.a.lam@gmail.com

Link: http://localhost:8080/BiofinityWiki/?mode=view&pageid=54

Figure 4.38: Alert dialog box displaying recommendations received

Of particular note is that intra-wiki recommendations generated by the algorithms in Chapter 3.1 are displayed to users via this same alert interface. In those instances, the recommendations will be said to be from Biofinity.

Tracking is performed for the intra-wiki sharing and recommendation-following/dismissing aspects of this panel. While entries created from these actions currently have no impact in the recommendations made or the modeling performed, they can be leveraged in future work, e.g. minimizing/managing interruption of user activity and the ensuing frustration.

The sharing panel, and consequently the sharing dialog box, is only available on the view modes of pages with editable content, including wiki pages, data pages, and publication pages. However, the alert box can be displayed on any page if the display conditions are satisfied.

4.2.2.8 Tags Panel

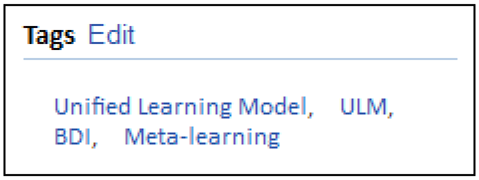
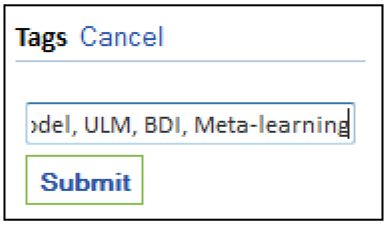


Figure 4.39: Tags panel

The tags panel (Figure 4.36) is used to associate a page with user-defined key words/phrases, and these may include words relevant to the page topics and areas of study. Each of these can be clicked to perform a search for other pages containing or tagged with these words, enabling users to find related content. To edit these, the user clicks the "Edit" link in the header. The list of tags then turns into a comma-separated list for the user to edit, as seen in Figure 4.37 below.



›del, ULM, BDI, Meta-learning

Figure 4.40: Editing tags

Upon submission, the list is sent to the server, which determines which tags have been added and removed by the edit.

This panel is available on the view modes of wiki, data, and publication pages.

4.2.2.9 Wikipedia Recommendation Panel

This panel currently appears only on wiki pages and is used to display hyperlinks to Wikipedia articles that may be related to the currently-viewed page. It provides users with easy access to additional related information, and clicking a link will open it in a new window. This feature provides two primary benefits to users: 1) improving comprehension of the page's contents by providing access to information on prerequisite and related topics, and 2) aid in the contribution of content to the Biofinity intelligent wiki.

As previously mentioned, the articles recommended are based upon keywords found in the page text after processing with LingPipe. For further information on how the page recommendations are generated, please refer back to Chapter 3.2.

Clicks to follow a Wikipedia recommendation are tracked by the system, and the tracking entry includes the keyword clicked and the page currently viewed before the click. While the use of this hyperlink to access Wikipedia is tracked, we cannot track further action taken by the user on Wikipedia. This action currently impacts neither modeling nor recommendation.

4.2.2.10 Data Panels

Data panels appear only on data pages and are populated with the Biofinity data that the pages are created from. Thus, they can only originate from the four supported Biofinity data types: classification, event, location, and occurrence data. The data panels for each of these have their own distinct fields to reflect the data type displayed. The fields for each of them are:

- Classification
    - Classification ID
    - Name
    - (repeated for each classification taxa)
        - Taxon Name

        - Taxon Rank
- Event
    - Event ID
    - Location ID
    - Event Date
    - Verbatim Event Date
    - Habitat
    - Sampling Effort
    - Sampling Protocol
- Location
    - Location ID
    - Longitude
    - Latitude
    - Verbatim Elevation
    - Continent
    - Country
    - State/Province
    - Locality Water Body
    - Island
    - Island Group
- Occurrence
    - Occurrence ID
    - Event ID

- Basis of Record
- Sex
- Life State
- Behavior
- Reproductive Condition
- Preparations

Upon initialization, they are provided with the Biofinity Entity ID of the data, which is used to request information to populate the fields via a Biofinity web service. Again, the information displayed in this panel cannot be edited directly through the wiki interface and instead needs to be modified through Biofinity.

The viewing of data panels is not tracked, although the viewing of the associated page is. Refer back to Chapter 4.2.1 for the impact of these tracking entries on our work.

4.2.2.11 Create Panel

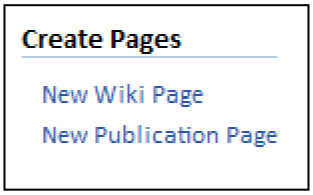
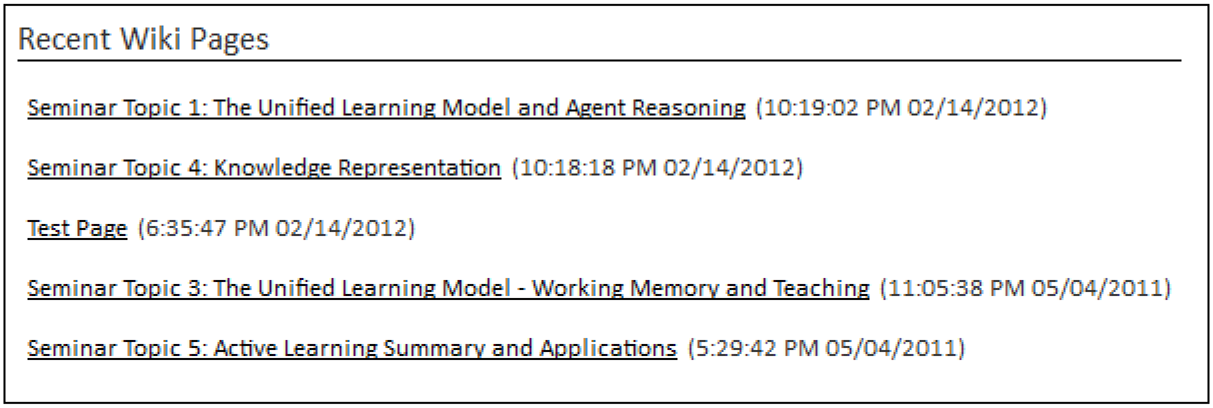


Figure 4.41: Create panel

The create panel (seen in Figure 4.38) is shown only on the main page and is used to display links for creating new wiki and publication pages. It is not shown when the user is not currently logged into the system. After clicking a link, the user is forwarded to the edit mode of the respective page type. Since there is no existing information for newly created pages, the fields in the revision forms are not pre-populated. Further, only the attachments panel is displayed for this “creation” mode.

Creating a page of either type causes a tracking entry to be created with the ID of the page and author involved.

4.2.2.12 Recent Pages Panels



Recent Wiki Pages

Seminar Topic 1: The Unified Learning Model and Agent Reasoning (10:19:02 PM 02/14/2012)

Seminar Topic 4: Knowledge Representation (10:18:18 PM 02/14/2012)

Test Page (6:35:47 PM 02/14/2012)

Seminar Topic 3: The Unified Learning Model - Working Memory and Teaching (11:05:38 PM 05/04/2011)

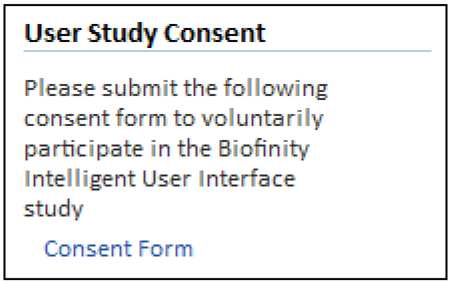
Seminar Topic 5: Active Learning Summary and Applications (5:29:42 PM 05/04/2011)

Figure 4.42: One of the three recent pages panels, RecentWikiPages

The recent pages panels display the five most recently edited wiki, publication, and data pages seen by peers in the user's current group. These panels only appear on the main page, beneath the welcome text in the primary content area. While the panel pictured is specifically for recent wiki pages, all other recent pages panels display their contents in a similar format.

The recent pages panels themselves do not require any tracking entries to be made for them. However, the viewing of the pages listed is tracked. Refer back to Chapter 4.2.1.x for additional information on the entries created and their impact.

4.2.2.13 Consent Panel



User Study Consent

Please submit the following consent form to voluntarily participate in the Biofinity Intelligent User Interface study

Consent Form

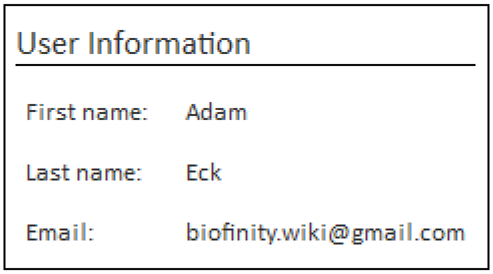
Figure 4.43: Consent panel

The consent panel (Figure 4.40) is only used on the Main Page and is placed in the sidebar area upon page load. When initialized, it obtains the current user's consent status from the wiki server. If the user has not yet filled one out, it displays the above message and provides a link to the Consent Form Page. Otherwise, it displays a thank-you message. Like with the consent page, no particular tracking is performed for the consent panel.

4.2.2.14 Users Panel

The users panel displays all wiki users belonging to the current group of the current user. After querying for and receiving this information from the server, each peer is displayed in "<first name> <last name>" format in a comma-separated list, and each name is a hyperlink to that user's corresponding user page. This panel was originally placed on the main page although it is currently unused. No particular tracking is performed for this panel.

4.2.2.15 User Info Panel



User Information

First name: Adam

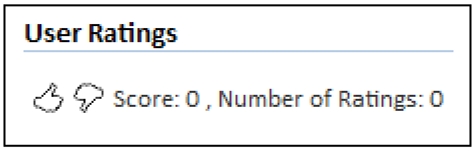
Last name: Eck

Email: biofinity.wiki@gmail.com

Figure 4.44:User info panel

The user info panel (Figure 4.41) is used in user pages to display the first name, last name, and e-mail address of a particular user. Since this panel only appears on user pages, the information displayed is that of the page's corresponding user. As previously mentioned, its contents cannot be directly modified through the wiki – instead, users must edit it through Biofinity via Accounts > Manage Accounts. There are no tracking entries related to this panel.

4.2.2.16 User Rating Panel



User Ratings

👍 👎 Score: 0 , Number of Ratings: 0

Figure 4.45: A user rating panel, as seen on a user page

The user rating panel (Figure 4.42) provides users with the opportunity to evaluate other users on a binary scale, i.e. "like" and "dislike." Like with page ratings, each user can only provide one rating at most for each other user, and this rating can be changed at any time. The overall "score" for the user is calculated by subtracting the number of "likes" from the number of "dislikes."

Tracking entries are made when a user provides a rating for another user, and this rating can be used in determining the ratee's expertise (Chapter 3.1.3).

4.2.2.17 User Stats Panels

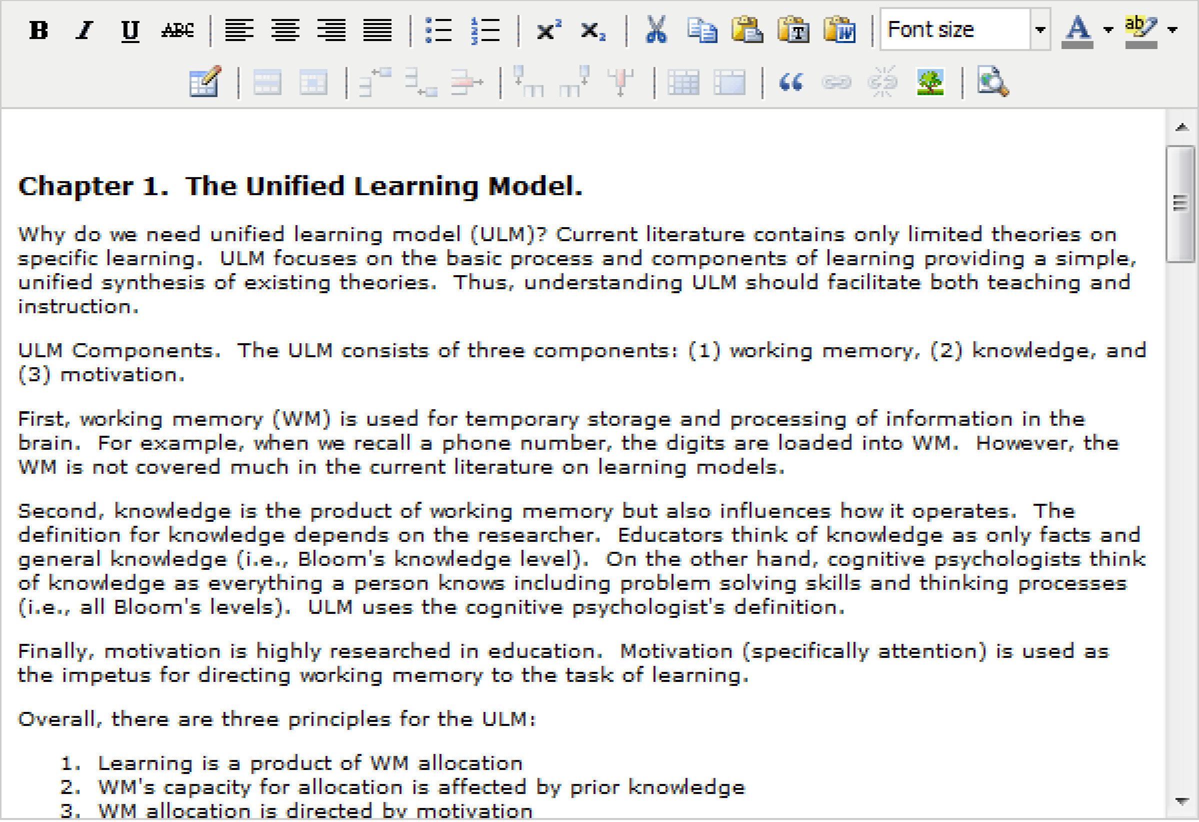
The user stats panels were developed during the wiki's early stages and displayed simple statistics on a particular user's behavior on the wiki. These were to be included on each user page, but were removed due to accuracy concerns and concerns that displaying them may influence user behavior.

These fell into three categories, including stats on the actions taken, on the user's recommendation activities, and on the user's session information. These are calculated online as a user page is loaded. In further detail, the statistics displayed include:

- Action Stats
    - Number of page views
    - Number of edits
    - Number of uploads/downloads
    - Number of comments
    - Number of page/user ratings made
    - Number of searches performed
    - Number of tags added/removed
    - Numerical rank for each of the above, relative to all other wiki users
- Recommendation Stats
    - Number of recommendations made
    - Number of recommendations followed
    - Numerical rank for each of the above, relative to all other wiki users
- Session Stats
    - Number of logins

- Total session duration
- Average session duration
- Numerical rank for each of the above, relative to all other wiki users

4.2.2.18 TinyMCE Editor



Chapter 1. The Unified Learning Model.

Why do we need unified learning model (ULM)? Current literature contains only limited theories on specific learning. ULM focuses on the basic process and components of learning providing a simple, unified synthesis of existing theories. Thus, understanding ULM should facilitate both teaching and instruction.

ULM Components. The ULM consists of three components: (1) working memory, (2) knowledge, and (3) motivation.

First, working memory (WM) is used for temporary storage and processing of information in the brain. For example, when we recall a phone number, the digits are loaded into WM. However, the WM is not covered much in the current literature on learning models.

Second, knowledge is the product of working memory but also influences how it operates. The definition for knowledge depends on the researcher. Educators think of knowledge as only facts and general knowledge (i.e., Bloom's knowledge level). On the other hand, cognitive psychologists think of knowledge as everything a person knows including problem solving skills and thinking processes (i.e., all Bloom's levels). ULM uses the cognitive psychologist's definition.

Finally, motivation is highly researched in education. Motivation (specifically attention) is used as the impetus for directing working memory to the task of learning.

Overall, there are three principles for the ULM:

1. Learning is a product of WM allocation
2. WM's capacity for allocation is affected by prior knowledge
3. WM allocation is directed by motivation

Figure 4.46: The TinyMCE WYSIWYG text editor, as seen in a wiki page’s edit mode

While the third-party developed TinyMCE rich text editor (Figure 4.43) is technically not a panel (i.e. does not extend AbstractPanel) and does not behave similarly to one (i.e. does not request from or post information to the wiki server), it is worth distinguishing as a UI component. As a full-featured WYSIWYG editor, it provides an editing interface similar to that of Microsoft Word, including features such as:

- Bold, italics, underline, and strikethrough text modifiers
- Left, center, right, and justify text alignment

- Bulleted and numbered lists
- Super-/sub-script
- Cut, copy, and paste functions
- Font size, text color, and background color
- Table creation and associated utility features
- Block quote formatting
- Hyperlinking
- Image insertion

A key feature of the editor is that it represents its contents as plain-text HTML. When revisions are submitted to the server, the revision contents are transmitted (with characters converted to hex equivalents when necessary) and stored in the database in this form.

While edits are made through this editor, tracking entries are not made for its use.

4.3 Wiki Database

The intelligent wiki database exists separately from the one used by the Biofinity core although both exist on the same server. The wiki back end is the only component with direct access to the wiki database, and connections to it are distributed from a connection pool managed by the Glassfish server.

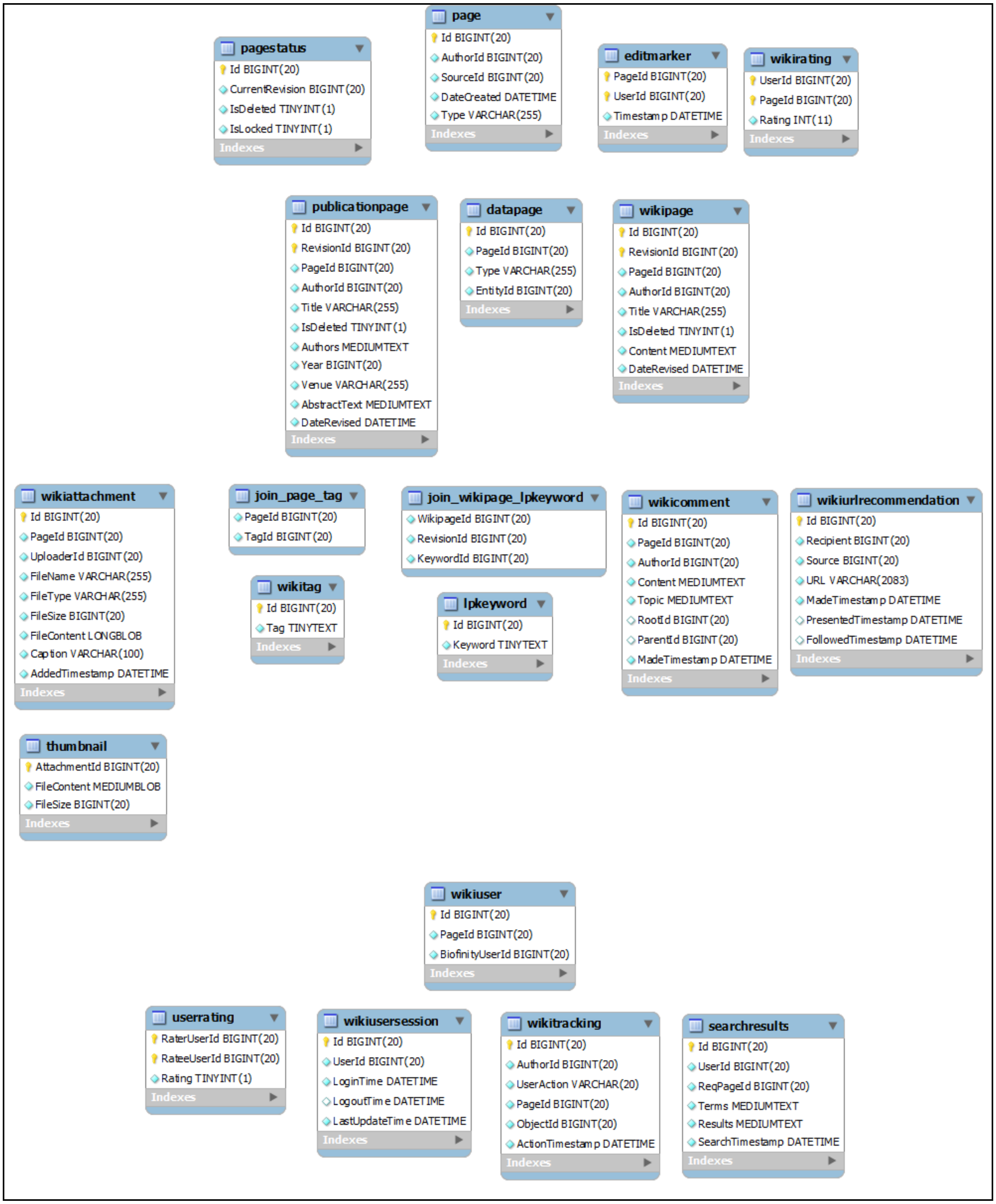


Figure 4.47: The Biofinity Intelligent Wiki's database schema

As seen in Figure4.44, there is a table for each of the objects and pages used in the wiki. While their purposes are self-explanatory from their names, their fields and relationships to the other tables may not be as straightforward. The subsections of this chapter (4.4.x) describe each of the tables in further detail.

4.3.1 Data Pages (datapage)

The “datapage” table stores information pertaining to data pages generated from data in Biofinity. Currently, the types of data generating a data page include event, occurrence, location, and classification data. While users can create and edit wiki content on a data page, this table only stores the link to Biofinity data. The wiki content is instead stored in the “wikipage” table. There is a 0..1-to-1 relationship between the entries in this table and the “page” table in that each data page has a corresponding entry in “page” but not vice-versa. Similarly, there is a 0..1-to-1 mapping between data page entries and wiki page entries. Table 4.1 details each of its fields.

Field	**Type**	**Description**
Id (PK)	BIGINT(20)	Identifier for the data page
PageId	BIGINT(20)	The corresponding page ID in table “page”
Type	VARCHAR(255)	The type of Biofinity data that the page is displaying. Currently can be one of *event*, *occurrence*, *location*, and *classification* data.
EntityId	BIGINT(20)	Entity ID of the corresponding data in the Biofinity DB

Table 4.1: Fields for Table "datapage"

4.3.2 Edit Markers (editmarker)

The “editmarker” table keeps track of the editing markers that warn users when others are editing a single page concurrently. That is, if an entry exists for the page that the user wishes to edit, then the user is warned of the concurrent editors before entering the page’s edit mode. There is a 0..1-to-1 relationship between its entries and the “page” table in that each entry is tied to a corresponding page ID, but not vice-versa. Its fields are detailed in Table 4.2.

Field	Type	Description
PageId (PK)	BIGINT(20)	The page to reserve a marker for
UserId	BIGINT(20)	The user possessing the marker
Timestamp	DATETIME	Indicates the edit marker issue date

Table 4.2: Fields for Table "editmarker"

4.3.3 Join Page to Tag (join_page_tag)

The "join_page_tag" table is a join table that links pages (table "page") to keyword tags (table "wikitag"), and there is a many-to-many relationship between them. Table 4.3 details the fields of the table.

Field	Type	Description
PageId	BIGINT(20)	The page to join the tag to
TagId	BIGINT(20)	The tag to join to the page

4.3.4 Join Wiki Page to LingPipe Keyword (join_wikipage_lpkeyword)

The "join_wikipage_lpkeyword" table is a join table that links particular wiki page revisions (table "wikipage") to keywords extracted by LingPipe (table "lpkeyword"). It joins wikipage-revision ID combinations with keyword IDs in a many-to-many relationship.

The table has since been revised to "Join Wiki Page to Automated Keyword" in later iterations of the wiki.

Field	Type	Description
WikipageId	BIGINT(20)	The wiki page to join the LingPipe keyword to
RevisionId	BIGINT(20)	The revision of the wiki page that the LingPipe keyword was generated for

KeywordId	BIGINT(20)	The LingPipe keyword to link to the page-revision combination

4.3.5 LingPipe Keywords (lpkeyword)

The “lpkeyword” table keeps track of all LingPipe keywords extracted from wiki page revisions, and the keywords are joined to specific wiki page revisions via the “join_wikipage_lpkeyword” table.

The table has since been revised to “Automated Keywords” in later iterations of the wiki.

Field	Type	Description
Id (PK)	BIGINT(20)	Identifier for the keyword
Keyword	TINYTEXT	The keyword text

4.3.6 Pages (page)

The “page” table contains the basic, immutable information about all pages in the wiki, and it provides a unique ID for the page to be referenced by regardless of type. All wiki objects in the wiki DB (except the join_wikipage_lpkeyword table) use this particular ID when referring to the page they are linked to.

Field	Type	Description
Id (PK)	BIGINT(20)	Identifier for the page
AuthorId	BIGINT(20)	The original/first page author
SourceId	BIGINT(20)	The Biofinity lab that the page is associated with
DateCreated	DATETIME	Created timestamp for page
Type	VARCHAR(255)	The page’s type, i.e. wiki, user, data, or publication

4.3.7 Page Status (pagestatus)

The “pagestatus” table contains information pertaining to the current status of the page, such as the default revision to display on page load and its locked/deleted status.

Note that the “IsLocked” field in this table is set when a user clicks the “Lock” button in the common page controls panel, and is *not* the lock induced by the “editmarker” table previously described. Pages that are locked cannot be edited until the lock is lifted, and deleted pages cannot be viewed or edited. The entries in this table have a 1-to-1 relationship with the entries in table “page.”

Field	**Type**	**Description**
Id	BIGINT(20)	The page that the status entry is for
CurrentRevision	BIGINT(20)	Default page revision to display when the page is first loaded
IsDeleted	TINYINT(1)	Flag indicating whether the page is deleted. 0 indicates that it is not deleted, and 1 indicates that it is.
IsLocked	TINYINT(1)	Flag indicating whether the page is locked from editing. 0 indicates that it is not locked, and 1 indicates that it is.

4.3.8 Publication Pages (publicationpage)

The “publicationpage” table stores content and revision information for special pages describing and organizing publications. They are different from the other page types in that its contents are split into distinct fields rather than being contained in a single generic field. As such, its structure is similar to that of the “wikipage” table in Chapter 4.2.14. There is a 0..1-to-1 relationship between the entries in this table and the “page” table.

Field	**Type**	**Description**
Id (PK)	BIGINT(20)	Identifier for the publication page
RevisionId (PK)	BIGINT(20)	Identifier for a particular revision of a publication page
PageId	BIGINT(20)	The publication page’s

		corresponding identifier in the “page” table
AuthorId	BIGINT(20)	The authoring user of the publication page revision
Title	VARCHAR(255)	The title of the publication page
IsDeleted	TINYINT(1)	Flag indicating whether the particular page and revision has been deleted
Authors	MEDIUMTEXT	The displayed list of authors for the publication
Year	BIGINT(20)	The year of publication
Venue	VARCHAR(255)	The publication venue
AbstractText	MEDIUMTEXT	An abstract for the publication
DateRevised	DATETIME	Timestamp for the publication page revision

4.3.9 Search Results (searchresults)

The “searchresults” table stores entries for tracking search behavior in the wiki, such as the terms used and the results returned. For simplicity, the search results are left in the XML form generated by the wiki server to be returned to the wiki client. Each entry in the table is associated with one user and one page whereas users and pages may have multiple search results associated with them.

Field	**Type**	**Description**
Id (PK)	BIGINT(20)	Identifier for the search result entry
UserId	BIGINT(20)	The user making the search
ReqPageId	BIGINT(20)	The current page when the search was performed
Terms	MEDIUMTEXT	The terms used for the search
Results	MEDIUMTEXT	The search results returned by the intelligent wiki back end
SearchTimestamp	DATETIME	The timestamp for the search

4.3.10 Thumbnails (thumbnail)

The “thumbnail” table stores thumbnail data generated for image wiki attachments. Its entries are automatically generated upon image upload and have a 1-to-1 relationship with image attachments in the “wikiattachment” table.

Field	Type	Description
AttachmentId (PK)	BIGINT(20)	The wikiattachment corresponding to the thumbnail
FileContent	MEDIUMBLOB	Binary data for the thumbnail
FileSize	BIGINT(20)	The file size of the thumbnail

4.3.11 User Ratings (userrating)

The “userrating” table stores the ratings that users have made towards other users. While the user rating feature was not enabled in the wiki deployment for gathering data, it exists to enable future work in recommending users to collaborate with. The entries in this table have a many-to-1 relationship with the users in the system – users may be associated with making or receiving multiple ratings of other users, but each rating entry is associated with only one rater/ratee.

Field	Type	Description
RaterUserId (PK)	BIGINT(20)	The user issuing the rating
RateeUserId	BIGINT(20)	The user being rated
Rating	TINYINT(1)	The rating given. 0 for “thumbs down”, and 1 for “thumbs up.”

4.3.12 Wiki Attachments (wikiattachment)

The “wikiattachment” table stores the files and associated metadata of items uploaded to the wiki. To preserve space, separate revisions of a same file are not currently supported, and uploads with the same file name on the same page will overwrite the existing one. The entries in this table have a many-to-1 relationship with pages and users.

Field	Type	Description
Id (PK)	BIGINT(20)	Identifier for the attachment
PageId	BIGINT(20)	The page that the attachment is added to
UploaderId	BIGINT(20)	The user that uploaded the attachment
FileName	VARCHAR(255)	The attachment file name
FileType	VARCHAR(255)	The attachment file type
FileSize	BIGINT(20)	The attachment file size
FileContent	LONGBLOB	Binary data for the attachment
Caption	VARCHAR(255)	Caption to display for the attachment
AddedTimestamp	DATETIME	Upload timestamp for the attachment

4.3.13 Wiki Comments (wikicomment)

The “wikicomment” table stores all information pertaining to comments/discussions occurring in the wiki. The inclusion of the “RootId” and “ParentId” fields enables comments to be made in a tree-like structure, which enables the representation of “nested conversation threads” in the table. There is a many-to-1 relationship between the entries in this table and pages/users. Pages and users can be associated with more than one comment, but each comment is only associated with one page and one user.

Field	Type	Description
Id (PK)	BIGINT(20)	Identifier for the comment
PageId	BIGINT(20)	The page that the comment was posted on
AuthorId	BIGINT(20)	The comment author
Content	MEDIUMTEXT	The body content of the comment
Topic	MEDIUMTEXT	The title of the comment/discussion thread

RootId	BIGINT(20)	The root comment of the thread tree this comment exists in
ParentId	BIGINT(20)	The direct parent of this comment in the thread tree
MadeTimestamp	DATETIME	Timestamp indicating when the comment is made

4.3.14 Wiki Pages (wikipage)

The “wikipage” table stores information pertaining to particular wiki pages and their revisions. In spite of its name, the table also stores the editable wiki information from data pages (i.e. each data page contains a wiki page). Its fields are largely similar to the fields of the “publicationpage” table, but it has one large generic content field instead of multiple smaller specialized fields. The entries of this table have a 0..1-to-1 relationship with the entries in the “page” table and a 1-to-0..1 mapping with the entries in “datapage.”

Field	Type	Description
Id (PK)	BIGINT(20)	Identifier for the wiki page
RevisionId (PK)	BIGINT(20)	Identifier for the particular revision of the wiki page
PageId	BIGINT(20)	The corresponding identifier for the wiki page in table “page”
AuthorId	BIGINT(20)	The authoring user of the revision
Title	VARCHAR(255)	The title of the wiki page
IsDeleted	TINYINT(1)	Flag indicating whether the particular page revision has been deleted (0 for not deleted, 1 for deleted)
Content	MEDIUMTEXT	The contents of the revision, as HTML
DateRevised	DATETIME	Timestamp for the revision

4.3.15 Wiki Ratings (wikirating)

The "wikirating" table stores ratings that users have given to particular pages, with ratings being on a scale from 1 to 5. It should be noted that using the IDs of both the rater and the page as the primary key, users cannot "spam" ratings to heavily influence the page's average, assuming that a reasonable number of other users have already voted. Instead, the user's newest rating will overwrite the old one given. There is a many-to-1 relationship between users/pages and entries in the "wikirating" table.

Field	Type	Description
UserId (PK)	BIGINT(20)	The user issuing the rating
PageId (PK)	BIGINT(20)	The page being rated
Rating	INT(11)	The numerical rating given (1-5)

4.3.16 Wiki Tags (wikitag)

The "wikitag" table keeps track of all tags used in the intelligent wiki, and the tags are joined to specific wiki page revisions via the "join_page_tag" table. Its entries have a many-to-many relationship with pages since each tag can be applied to multiple pages, and each page can be associated with multiple tags.

Field	Type	Description
Id (PK)	BIGINT(20)	Identifier of the tag
Tag	TINYTEXT	The text for the tag

4.3.17 Wiki Tracking (wikitracking)

The "wikitracking" table stores actions taken by every wiki user in the system, including the action performed, when it was performed, the page it occurred on, and the object acted upon. **In other words, this table stores tracking information of user behavior and contains the bulk of the data used in our analysis.** There is a many-to-1 relationship between the tracking entries and wiki users, the page involved, and the involved object. That is, particular users, pages, and objects may have multiple tracking entries associated with them, but a particular tracking entry will only be associated with one of each.

Field	**Type**	**Description**
Id (PK)	BIGINT(20)	Identifier for the tracking entry
AuthorId	BIGINT(20)	The user performing the action
UserAction	VARCHAR(255)	The action performed. Currently can be one of: *create, view, edit, delete, rate, rateUser, comment, upload, download, search, addTag, removeTag, publish, unpublish, lock, unlock, setCurrentRevision,* and *clickKeyword*.
PageId	BIGINT(20)	The page the action was performed on
ObjectId	BIGINT(20)	The object involved in the action. May be optional depending on UserAction.
ActionTimestamp	DATETIME	The timestamp of the action, i.e. when it was performed

4.3.18 Wiki URL Recommendations (wikiurlrecommendation)

The “wikiurlrecommendation” table stores all the recommendations made within the system, whether they be from user to user via the intra-wiki “Share” button or from the system to users. Since the recommendation entries are stored via URL as opposed to page IDs, recommendations to content outside the wiki can also be stored. There is a many-to-one relationship between users and URL recommendations – each user can make and receive multiple recommendations, but each wiki URL recommendation is only associated with two users.

Field	**Type**	**Description**

Id (PK)	BIGINT(20)	The unique identifier of the recommendation
Recipient	BIGINT(20)	The ID of the user to receive the recommendation
Source	BIGINT(20)	The ID of the user sending the recommendation
URL	VARCHAR(255)	The URL of the recommended item
MadeTimestamp	DATETIME	Timestamp when the recommendation was made, i.e. sent by the source
PresentedTimestamp	DATETIME	Timestamp when the recommendation was shown to the recipient
FollowedTimestamp	DATETIME	Timestamp when the recommendation was followed by the recipient

4.3.19 Wiki Users (wikiusers)

The “wikiusers” table assigns a unique ID to each user of the intelligent wiki and associates it with the user’s corresponding Biofinity user ID and an automatically-generated user page. References to users in the other tables (e.g. as AuthorId, UserId, RateeId, etc.) use this wiki-specific ID and *not* the user’s corresponding Biofinity ID.

Field	**Type**	**Description**
Id (PK)	BIGINT(20)	Identifier for a user in within the intelligent wiki
PageId	BIGINT(20)	The user’s corresponding user page
BiofinityUserId	BIGINT(20)	The user’s corresponding Biofinity user ID

4.3.20 Wiki User Sessions (wikiusersession)

The “wikiusersession” table stores log in/out times of each user’s sessions. While the system can consistently detect when the user logs in, the logout time may be less straightforward to determine if the user doesn’t log out manually. For example, when users forget to log out, the session entry could be left “open”, i.e. with no logout time, up through the user's next login. One approach taken to avoid this is to catch browser close events via Javascript and obtaining the timestamp when this occurs. In the event that the Javascript solution fails, the server sets the logout time to the last update time if 15 minutes have elapsed since then.

There is a many-to-one relationship between the entries in this table and wikiusers. Its fields are detailed in Table 4.20 below.

Field	**Type**	**Description**
Id (PK)	BIGINT(20)	Identifier for user session
UserId	BIGINT(20)	The user to whom the session belongs
LoginTime	DATETIME	The timestamp when the user logs in
LogoutTime	DATETIME	The timestamp when the user logs out
LastUpdateTime	DATETIME	The timestamp of the last time the user session was updated. Matches LogoutTime when the session is closed.

Chapter 5: Results

We have deployed the Biofinity Intelligent Wiki for use in the collaborative writing assignments of two classes at the University of Nebraska-Lincoln: Artificial Intelligence Applications (RAIK 390) during the Spring 2011 semester and Multiagent Systems (MAS 475/875) during the Fall 2011 semester. By gathering and analyzing the usage data of the students in these courses, we aim to: 1) gain a better understanding of the relative importance of each algorithm component and 2) identify trends in student behavior that can be leveraged in future instruction.

Section 5.1 first describes the logistics of the classes and their collaborative assignment within the wiki. From there, it summarizes the primary activities performed that we would like to track along with the rationale for choosing them. These attributes are the basis upon which we perform further analysis, including **active vs. passive** and **minimalist vs. overachiever** metrics for their activity profiles. Section 5.2 then delves into the results themselves, presenting the processed data within various contexts and highlighting noteworthy trends. We then summarize our findings in Section 5.7.

5.1 Logistics

This section describes the classes to which the Biofinity Intelligent Wiki has been deployed, as well as the collaborative writing assignments that make use of it. It also covers and justifies our specific points of observation and evaluation, including: the specific student activities to track and corresponding metrics of interest, our derived metrics of **active vs. passive** and **minimalist vs. overachiever** for their activity profiles, and additional potentially-valuable views on the data.

5.1.1 Class Descriptions

RAIK 390 is an exclusive honors class taken by highly motivated juniors and seniors majoring in business, computer science, and/or computer engineering. Its objective is to provide its students with the background knowledge necessary to recognize the need for and apply artificial intelligence to business applications. It has a class size of 16 students, and 15 students provided consent for their usage data to be used in this thesis.

MAS 475/875 is an upper-level computer science course that introduces the theories and applications of multi-agent systems. It primarily consists of junior and senior undergraduate students and graduate students, all of whom major in computer science or computer engineering. Although it is a relatively large class of 29 students, 17 students provided consent for their usage data to be used in the thesis.

5.1.2 Assignment Description

The collaborative writing assignments for both classes share similar guidelines in spite of the differing course topics. Initially, each student individually writes a summary on a topic covered in the course, and these write-ups consist of:

- An *overview* of the topic, including motivations and underlying principles, etc.
- A list of *praises*: descriptions of what the student believes are the important/useful aspects of the topic
- A list of *critiques*: descriptions of what the student believes are the weaknesses of topic
- Its *applications*: how the topic relates to real-world applications

- An initial set of *questions*: to encourage discussion in the next phase of the assignment (MAS only)

After this "individual contribution" phase is completed, the students then move on to a 4-/5-week long "collaboration phase" where they contribute towards, rate, and tag other students' summaries and participate in threaded discussions. However, they are not permitted to directly modify content submitted by other users.

Halfway through the collaboration phase, the instructor provides initial feedback on the students' collaborative efforts. Students' collaborative activities are graded in the following manner:

- 60% Wiki Editing (at least three other essays, and amount and quality)
- 30% Threaded Discussions
- 10% Rating (RAIK) or Rating, Tagging, and Viewing (MAS)

5.1.3 Points of Observation/Evaluation

When tracking and evaluating student activity in the wiki, we specifically focus on the actions that we believe to be observable indications of collaboration in a wiki:

- **Edits/revisions** – Edits and revisions to wiki pages are obvious indicators of collaboration in a wiki, and we consider them to be one of the most important forms of contribution we can observe. To gather of this data, we track all revisions made in the wiki and extract the *number of revisions* and *word length of each revision*, then manually read and evaluate its *quality*.
- **Comments/discussion** – Comments and discussions carried out on each via the Comments Panels previously described in the Implementation chapter. Alongside

edits/revisions, we consider these to also be one of the most important forms of wiki collaboration that we can observe, since the discussions can give rise to new ideas and additional wiki content. Specifically, we will track all comments made in the wiki and count the *number of comments made* and their *lengths* (in number of words). We then manually determine their *quality*.

- **Keyword tagging** – Marking wiki pages with words that can be used to mark or summarize their contents. Although we track the user performing the tagging, the keywords used, and the pages tagged with the keywords, we only consider the *number of occurrences of the tagging action by each user*.
- **Ratings/evaluation** – Numerical ratings given by a user to a particular wiki page, ideally after reading and evaluating its contents. Although we track the user making each rating, the page being rated, and the rating given, we only make use of the *number of occurrences of the rating action by each user* rather than the actual rating provided or its accuracy.

Based on the numbers of edits, comments, keywords tagged, and ratings provided as well as the average edit and comment lengths, we **categorize** the students in each class with this **"raw" attribute data**. Doing so may highlight the particular attributes that are most valuable in grouping them as well as provide some information on student behavior relative to the assignment's requirements.

Additionally, we propose the following metrics to categorize the students based on their activity profiles:

- **Active vs. Passive** – Categorizing students based on the number and type of collaborative actions performed.
    - *Passive* – Greater focus on tagging and rating without as many or comments or edits made. That is, these profiles consist primarily of actions that require low cognitive cost, relative to the cognitive cost of contributing to page content or discussion.
    - *Active* – Greater focus on many edits and comments, more pro-active, i.e. first to perform activities on non-primary pages.
- **Minimalist vs. Overachiever** – Examines the degree to which students participate, i.e. performs the aforementioned collaborative actions, relative to the minimum requirements for the assignment.
    - *Minimalist* – Performs minimum number of edits required for assignment, try to "game the system" (e.g. make non-valuable comments to increase comment count)
    - *Overachiever* – Makes valuable edits and comments and more than minimum required, performs more types of collaborative activities often.

Finally, we examine the usage data for **patterns in students' page sets for editing and commenting** and for **cliques among students**.

5.2 Categorization from Raw Attribute Data

After gathering the students' usage data, we wish to group the students based on their activities within the wiki. Since the data is largely numeric, we can leverage a clustering algorithm to classify them based on the quantitative aspects of their profiles. A few considerations must be made when selecting one appropriate for our data:

- The optimal number of clusters needed to appropriately categorize each class's students is not known
- Representative instances to use as cluster centers are not known
- There are no labels to assign to instances
- The number of instances for each data set is relatively few, e.g. less than 20

Based on these, we have determined the X-means clustering algorithm to be most appropriate for us due to its ability to: 1) determine the optimal number of clusters (from a range between a user-specified minimum and maximum) needed for the best clustering results and 2) its ability to operate on data lacking labels and known representative instances, i.e., unsupervised learning. We use the implementation of the X-means algorithm provided by the WEKA machine learning software suite to cluster the students in each class. The following settings were used:

Parameter	**Value**
binValue	1
cutOffFactor	0.5
debugLevel	0
debugVectorsFile	weka-3-6
distance	Euclidean distance
maxIterations	1
maxKMeans	1000
maxKMeansForChildren	1000
maxNumClusters	(# students for class) / 2
minNumClusters (*MinK*)	1

useKDTree	false

Table 5.2: Parameters used for X-means clustering in WEKA

Unfortunately, the results from the X-means clustering were unusual and unintuitive, either due to the characteristics and attributes of our data sets or due to idiosyncrasies in the WEKA implementation of the X-means algorithm used:

- The optimal solutions consistently grouped the instances into *MinK* (or fewer) clusters, where *MinK* is the X-means parameter specifying the minimum number of clusters expected
- Cluster assignments for each instance are not consistent across many (30) different seeds, i.e. instances do not have the same peers

5.2.1 Maximal Pairs Algorithm

We devised the following process to classify the students based on the results obtained from the X-means classifier. It can be summarized in the following steps:

5.2.1.1 Determine the optimal number of clusters k

Since the number of clusters in the X-means results is consistently dependent on the *MinK* value specified by us, it is difficult to determine whether the resulting number of clusters is truly optimal. Fortunately, the X-means results in the WEKA package do return two measures of the clustering effectiveness along with the clustering configurations themselves: ***Distortion*** and ***Bayesian Information Criterion (BIC***) values. A good clustering solution ideally minimizes Distortion while maximizing BIC. However, we've empirically determined that both Distortion and BIC values decrease for increasing values of *MinK* for both our data sets.

We then defined the following weighted sum to find the optimal number of clusters *k* that struck a balance between the two:

$$0.5 \times \left(1 - \frac{Distortion_k - Distortion_{\ min}}{Distortion_{\ max} - Distortion_{\ min}}\right) + 0.5 \times \left(\frac{BIC_k - BIC_{min}}{BIC_{\ max} - BIC_{min}}\right)$$

The value of *k* to be targeted (and the value of *MinK* to be used with X-means for results) is the value of *k* that maximizes the above formula.

5.2.1.2 Put the instances into k clusters, based on the co-occurrence data

Since the clusters formed by X-means are also dependent on the particular seed used and cluster membership is relatively inconsistent, we defined a simple clustering scheme that builds from the different clustering results across many seeds: iteratively build/merge clusters based on pairs of instances that co-occur most frequently, then second-most frequently, etc. This is based on the intuition that instances appearing together frequently in the X-means results, i.e. *co-occur* in the same cluster across many different seeds, are more likely to "truly belong" to the same cluster. Similarly, instances that rarely co-occur in the same clusters are less likely to "belong" together.

The pseudo code for this is:

```
•   Initialize each instance to its own singleton cluster.
•   While number of unique clusters remaining in $C$ is greater than $k$:
        o   For each cluster $A$ in $C$:
                ▪   For each cluster $B$ in $C$:
                        •   If $A$ != $B$ AND HasStrongMaximalPair( $A$, $B$ ) == TRUE
                                o   Merge( $A$, $B$ )
```

After *k* is found via the process in Section 5.2.1.1, the goal is to iteratively merge the instances into *k* clusters. Ideally, the algorithm ends when exactly *k* clusters remain, and this value is specifically targeted to correspond with the X-means results obtained for *MinK* = *k*.

As previously mentioned, the central concept upon which instances and clusters are joined is the idea of grouping frequently co-occurring instances together. We define a ***maximal pair*** for an instance to be the set of instances with which it co-occurs the most, excluding itself and instances within its current cluster. We also define an instance to be a ***strong maximal pair*** of another if they are mutually maximal pairs of one another. That is, *when instances are strong maximal pairs of one another, there are no other instances with which either co-occur more often.* And thus those instances shall be joined into the same cluster.

The Definitions subsection delves into the more-formal definitions and discussions of the maximal pairs and strong maximal pairs concepts.

Figure 5.1 summarizes the flow and transformation of information from the raw usage data to the final clusters.

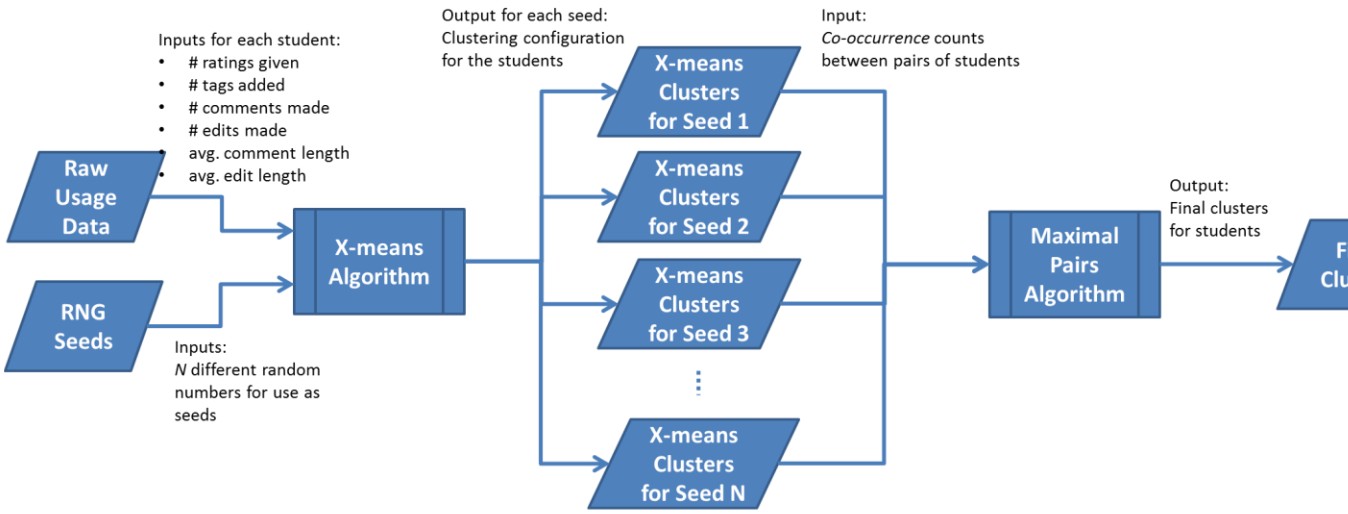

Inputs for each student:
# ratings given
# tags added
# comments made
# edits made
avg. comment length
avg. edit length
Raw Usage Data
RNG Seeds
Inputs:
N different random numbers for use as seeds
X-means Algorithm
Output for each seed:
Clustering configuration for the students
X-means Clusters for Seed 1
X-means Clusters for Seed 2
X-means Clusters for Seed 3
X-means Clusters for Seed N
Input:
Co-occurrence counts between pairs of students
Maximal Pairs Algorithm
Output:
Final clusters for students
F
Clu

Figure 5.48: Information transformation from raw usage data to final clusters

5.2.2 Definitions

This section more formally defines the concepts of ***maximal pairs*** and ***strong maximal pairs*** and the design decisions pertaining to their use in our maximum pairs-based clustering algorithm. Examples are also provided to aid in understanding of the concepts.

5.2.2.1 Preliminary Definitions

Before delving into the specifics of the two terms, we first define the various constants and functions used in describing them.

k – the *MinK* value used for X-means. This value is empirically determined by:

1) Running X-means with *MinK* ranging from 1 to (|*S*| / 2) and recording the average Distortion and BIC value across *N* seeds.

2) Determining the optimal value of *MinK* (and consequently, *k*) by maximizing:

$$0.5 \times \left(1 - \frac{Distortion_k - Distortion\ _{min}}{Distortion\ _{max} - Distortion\ _{min}}\right) + 0.5 \times \left(\frac{BIC_k - BIC_{min}}{BIC\ _{max} - BIC_{min}}\right)$$

N – the number of seeds used for the X-means portion of the process, i.e. the number of seeds used to generate co-occurrence data. While there are no specific guidelines for selecting a value for *N*, we generally want *N* >> *k*.

S – the set of all instances to be clustered.

C – the set of clusters currently unmerged during a particular iteration of the algorithm.

a.cluster – the ID of the cluster to which instance *a* belongs.

Co-occurrence(*a, b*) – the number of times instance *a* and instance *b* appear in the same cluster, when X-means is run with *MinK* = *k* across *N* seeds. This value has a range of 0 to *N*.

Merge(*A, B*) – the procedure to merge clusters *A* and *B*.

5.2.2.2 Maximal Pairs

A maximal pair for instance *a* is the instance *b* in *S* where the following condition holds:

- Co-occurrence (*a, b*) ≥ Max(Co-occurrence(*a, c*)) for all instances *c* where:
    - *c* ∈ *S*
    - *a.cluster != c.cluster*
    - *b != c*

The above definition suffices for finding the maximal pair of singleton clusters. Note that a given instance may have multiple maximal pairs. For example, two distinct instances *b* and *c* can both be a maximal pair of an instance *a* when Co-occurrence (*a*, *b*) = Co-occurrence(*a*, *c*).

We also define the maximal pairs for an entire cluster *A* as the most frequently co-occurring maximal pairs across all of the cluster's members. That is:

```
• Set of maximal pairs $M = \emptyset$, max_count = 0
• For each instance $a$ in $A$:
    o Instance $b$ = any instance in MaximalPairs( $a$ )
    o If Co-occurrence( $a, b$ ) > max_count
        ▪ max_count = Co-occurrence($a, b$ )
        ▪ $M = \{ \text{MaximalPairs}( a ) \}$
    o Else if Co-occurrence( $a, b$ ) == max_count
        ▪ $M = M \cup \{\text{MaximalPairs}( a ) \}$
• Return $M$
```

For example, consider the following co-occurrence table for cluster *A* = { *a1*, *a2*, *a3* } with outside-the-cluster instances *b*, *c*, *d*, *e*, and *f*:

	b	***c***	***d***	***e***	***f***
a1	5	10	2	7	8
a2	0	0	1	10	10
a3	8	8	2	3	1

Table 5.3: Co-occurrence table for cluster A = { a1, a2, a3 }

Each cell indicates the number of times (across the range of seeds used in running X-means) for which the instances in the corresponding row and column have been grouped in the same cluster by the X-means algorithm. As can be seen, instance *a1* co-occurs with instance *b* 5 times, with instance *c* 10 times, etc. The maximal pairs for *a1*, *a2*, and *a3* are then the instances with which they co-occurred the most, respectively:

- For *a1,* the maximal pair is *c* since it has the largest number of co-occurrences with a1 (10) compared to *b*, *d*, *e*, and *f*.
- For *a2,* the maximal pairs are both *e* and *f*, since they tie for the largest number of co-occurrences with *a2* (10)
- *a3* also has two maximal pairs since *b* and *c* tie for the largest number of co-occurrences with *a3* (8)

Then to determine the maximal pairs for the entire cluster, we iterate through the maximal pairs of its members:

- *a1*'s max pair of { *c* } is added to the set of max pairs for *A* since it is the first set of max pairs considered (whose co-occurrences exceed zero)
- *a2*'s max pairs of { *e*, *f* } is added to the set of max pairs for *A* since their co-occurrences tie that of *a1*'s max pair
- *a3*'s max pairs of { *b*, *c* } are *not* added to the max pairs for *A* since their co-occurrences with *a3* do not exceed the co-occurrences of the ones added thus far (8 < 10)

Thus, the maximal pairs of cluster *A* would be { *c*, *e*, *f* }.

This approach of using the *most frequently co-occurring* maximal pairs as the "representative" ones for the entire cluster was selected since it strikes a balance between the extremes of:

1. Including all maximal pairs for every cluster member as part of the cluster's maximal pairs set, and

2. Using only one member's maximal pairs when multiple candidates exist.

Option 1 is relatively lacking in restrictions compared to our chosen approach. While its effects are not noticeable for singleton clusters, it provides subtly different results in the returned maximal pairs. Returning back to the co-occurrences in Table 5.1, the maximal pair(s) for instances *a1, a2,* and *a3* are { c }, { e, f }, and { b, c }, respectively. Under this option, the cluster's maximal pairs would be the union of the 3 sets, { b, c, e, f }, meaning *b* is included regardless of its co-occurrence count with *a3*. This is problematic in a couple ways. First, since clusters are merged with maximal pairs as the basis, this enables the cluster to be merged with another on this "weaker" maximal pair. Second, it enables larger clusters to have a larger number of instances in their maximal pairs sets. This in turn increases their chances of merging with another cluster, leading to a "snowball effect" where bigger clusters continue to grow while smaller clusters are less likely to merge with one another.

Option 2 is restrictive towards the opposite extreme and also affects singleton clusters that have multiple maximal pairs. It can share the same weakness as (1) if the member chosen does not contain one of the cluster's "strongest" maximal pairs, although this can be remedied by only selecting among members that have a "strongest" pair. Secondly, this may introduce additional algorithm computational iterations without affecting its results. (See Extension (2) for further detail.)

We previously mentioned that the maximal pairs for a particular instance excluded itself and instances within its current cluster. It is apparent that the instance itself should not be a maximal pair candidate, since it would always co-occur with itself across all *N* seeds used for X-means. However, the reasoning for the cluster membership condition may not be as readily apparent:

- *a.cluster* != *b.cluster*

This additional condition serves two purposes. First, it removes unnecessary and/or redundant checks in the algorithm execution by skipping evaluation of instances that already belong to the same cluster. That is, since the maximal pairs concept is used to *merge* distinct clusters, it would be counterintuitive to examine instances within the same cluster as candidates. Second, it ensures that each cluster has an "outward-facing" maximal pair on every iteration of the algorithm. To better elaborate on this point, consider the following co-occurrence Table 5.3:

	a1	***a2***	***b1***	***b2***	***b3***	***c1***
a1	x	10	1	0	3	2
a2	10	x	5	6	1	3
b1	1	5	x	5	10	4
b2	0	6	5	x	10	3
b3	3	1	10	10	x	8
c1	2	3	4	3	8	x

Table 5.4: Co-occurrences between all instances in clusters *A* = {*a1, a2*}, *B* = {*b1, b2, b3*}, and *C* = {*c1*}

Using the definition of a cluster's maximal pair(s), we have the following maximal pairs for each cluster when *ignoring* the same-cluster condition:

Cluster	**Cluster Members**	**Cluster Maximal Pair(s)**
A	*a1, a2*	*a1,a2*
B	*b1, b2, b3*	*b1, b2, b3*
C	*c1*	*b3*

Table 5.5: Maximal pairs for clusters *A*, *B*, and *C* based on the co-occurrences in Table 5.2 when the same-cluster condition is ignored

Since cluster merges only occur when there are *mutually maximal pairs* between them, no merges can occur in this situation. When the same-cluster condition is followed, the maximal pairs for the clusters would then be:

Cluster	**Cluster Members**	**Cluster Maximal Pair(s)**
A	*a1, a2*	*b2*
B	*b1, b2, b3*	*c1*
C	*c1*	*b3*

Table 5.6: Maximal pairs for clusters *A*, *B*, and *C* based on the co-occurrences in Table 5.2 when the same-cluster condition is followed

Each of the clusters now have maximal pairs "outside" of themselves: cluster *A* has a max pair in cluster *B*, cluster *B* has a max pair in cluster *C*, and cluster *C* still has a max pair in cluster *B*. Since clusters *B* and *C* mutually have maximal pairs in one another, they can merge.

The following function definitions are used in the upcoming definition of strong maximal pairs:

MaximalPairs(*a*) – the procedure to obtain the set of maximal pairs for instance *a*. If there are multiple, all of them are included in the returned set.

MaximalPairs(*A*) – the procedure to obtain the set of maximal pairs for cluster *A*. Similar to the procedure for a single instance, this one also returns multiple instances when multiple maximal pairs are identified.

5.2.2.3 Strong Maximal Pairs

When instances *a* and *b* are a maximal pair of one another, i.e., mutually maximal pairs. That is, both:

- *a* ∈ MaximalPairs(*b*)
- *b* ∈ MaximalPairs(*a*)

Similarly, two clusters *A* and *B* can be considered to have a strong pair "joining" them, i.e. when both these conditions are simultaneously fulfilled:

- $\exists(a \in A)$ s.t. $a \in$ MaximalPairs(*B*)
- $\exists(b \in B)$ s.t. $b \in$ MaximalPairs(*A*)

As stated earlier, the algorithm uses this as the basis for merging clusters. When two clusters have a strong pair between them, they are merged. A triplet (or larger) can be merged when 2+ clusters have a mutual strong pair in a cluster whose maximal pairs span multiple clusters, as depicted in the following table:

Cluster	**Cluster Members**	**Cluster Maximal Pair(s)**
A	*a1, a2*	*c1*
B	*b1, b2, b3*	*c1*
C	*c1*	*a1, b3*

Table 5.7: Maximal pair situation where a cluster (*C*) can be merged with two other clusters (*A*, *B*)

Here, the maximal pair for clusters *A* and *B* is *c1*, and the maximal pairs for cluster *C* are both *a1* and *b3*. Thus, *A*-*C* and *B*-*C* are both strong cluster pairs.

The definition of the HasStrongMaximalPair function used previously in the introduction is then:

HasStrongMaximalPair(*A*, *B*) – the procedure to determine whether clusters *A* and *B* have a strong pair between them, according to our definition of strong maximal pairs.

5.2.2.4 Weak Maximal Pairs

We define a *weak maximal pair* to be a maximal pair that is *not* mutually maximal, i.e. not a strong maximal pair. It should be noted that a pair that is weak during one iteration of the algorithm may be strong in a later one, e.g., after merges occur between more-frequently occurring pairs.

5.2.3 Maximal Pairs Algorithm Results

This section details the cluster results of the maximal pairs algorithm. In addition to presenting the each class's cluster membership, we also describe our initial impressions of the results, examine each attribute's relative "strength" for the cluster assignments, justify any unusual traits, and discuss implications arising from them.

5.2.3.1 MAS Clusters

The categorization of the students in the MAS class is as follows:

Cluster ID	Members (Student ID)
Cluster 0	0, 2, 4, 6, 8, 14
Cluster 1	1, 5
Cluster 2	3, 7, 9, 10, 11, 13, 15, 16
Cluster 3	12

Table 5.8: Cluster assignments for the MAS class

Delving deeper, the attribute details for each of the clusters are:

MAS Cluster 0							
Student	**EditCnt**	**EditLength**	**NumCmts**	**CmtLength**	**NumTags**	**NumRates**	**Dist. From Centroid**
0	11	134.091	6	93.167	0	9	73.197
2	6	27.667	6	70.500	5	18	38.391
4	5	50.000	6	61.000	5	7	18.270
6	7	46.857	6	43.167	12	7	32.714
8	11	37.636	7	72.714	4	17	28.766
14	9	95.667	9	75.111	14	9	31.931
Avg μ_0	8.167	65.320	6.667	69.276	6.667	11.167	37.211
StDev σ_0	2.563	41.012	1.211	16.543	5.279	4.997	-

Table 5.9: Attribute details for MAS cluster 0

MAS Cluster 1							
Student	**EditCnt**	**EditLength**	**NumCmts**	**CmtLength**	**NumTags**	**NumRates**	**Dist. From Centroid**
1	7	47.43	5	93.6	27	9	32.117
5	6	89.17	4	140	12	7	32.117
Avg μ_1	6.500	68.298	4.500	116.800	19.500	8.000	32.117
StDev σ_1	0.707	29.513	0.707	32.810	10.607	1.414	-

Table 5.10: Attribute details for MAS cluster 1

MAS Cluster 2							
Student	EditCnt	EditLength	NumCmts	CmtLength	NumTags	NumRates	Dist. From Centroid
3	3	53.667	3	76.333	6	5	32.339
7	4	141.000	2	102.500	4	4	60.791
9	0	0.000	3	94.000	14	8	88.127
10	2	124.000	3	86.333	1	5	39.587
11	3	63.000	3	87.667	16	1	27.312
13	2	61.000	7	22.429	4	4	60.040
15	2	58.000	3	61.333	2	1	32.457
16	5	187.200	1	84.000	5	5	101.536
Avg μ_2	2.625	85.983	3.125	76.824	6.500	4.125	55.273
StDev σ_2	1.506	59.869	1.727	25.108	5.503	2.295	-

Table 5.11: Attribute details for MAS cluster 2

MAS Cluster 3							
Student	EditCnt	EditLength	NumCmts	CmtLength	NumTags	NumRates	Dist. From Centroid
12	0	0	0	0	0	0	0
Avg μ_3	-	-	-	-	-	-	-
StDev σ_3	-	-	-	-	-	-	-

Table 5.12: Attribute details for MAS cluster 3

5.2.3.1.1 Observations

The most notable impression of these cluster assignments is that Clusters 1 and 3 have fewer members than clusters 0 and 2. Looking into its attributes, the sole instance in Cluster 3 is a student that did not contribute to the wiki. It is good that the post-maximal pairs results singled out this extreme of lacking activity! On the other hand, Cluster 1 does not seem to represent a particular extreme of (quantitative) contributions to the wiki, and justification for its members is not as readily apparent.

To determine the relative importance of each attribute in the clustering, we examine each one individually.

- *Number of edits*

Clusters 0 and 1 seem to jointly comprise the “high” (5+) edit counts. However, the division between clusters 0 and 1 based on this attribute alone is not strong: $\mu_{0,editcnt} \pm \sigma_{0,editcnt}$ overlaps with $\mu_{1,editcnt} \pm \sigma_{1,editcnt}$ (within one stdev).

Clusters 2 and 3 seem to comprise the “low” (< 5) edit counts. Some separation between clusters 2 and 3 seems to exist: $\mu_{2,editcnt} \pm \sigma_{2,editcnt}$ does not overlap with $\mu_{3,editcnt} \pm \sigma_{3,editcnt}$ (i.e. the two clusters do not overlap within one stdev from their means). However, $\mu_{2,editcnt} \pm 2\sigma_{2,editcnt}$ overlaps with $\mu_{3,editcnt} \pm 2\sigma_{3,editcnt}$ (i.e., overlaps within two stdevs).

- *Edit length*

Excluding cluster 3, there appears to be no correlation between edit lengths and cluster assignments. The clusters have a mix of “relatively low” and “relatively high” edit lengths. That is, edit lengths for clusters 0, 1, and 2 all overlap one another within one stdev of their respective means. However, it may be possible for this attribute to be a distinguishing factor between clusters 2 and 3. This is currently uncertain since cluster 3 only has one member.

- *Number of comments*

Clusters 0 and 1 seem to comprise the “high” (4+) comment counts. Some separation between clusters 0 and 1 based on this attribute seems to exist: $\mu_0 \pm \sigma_0$ does not overlap with $\mu_1 \pm \sigma_1$ (i.e. doesn’t overlap within one stdev). However, $\mu_0 \pm 2\sigma_0$ overlaps with $\mu_1 \pm 2\sigma_1$.

Clusters 2 and 3 seem to comprise the "low" (< 4) comment counts, although student 13 appears to be an outlier for this attribute with a comment count of 7. Some separation between clusters 2 and 3 seems to exist: $\mu_2 \pm \sigma_2$ does not overlap with $\mu_3 \pm \sigma_3$. However, $\mu_2 \pm 2\sigma_2$ overlaps with $\mu_3 \pm 2\sigma_3$.

- *Comment length*

Excluding cluster 3, there appears to be no correlation between comment lengths and cluster assignments. The clusters have a mix of "relatively low" and "relatively high" comment lengths, and comment lengths for clusters 0, 1, and 2 all overlap one another within one stdev of their respective means.

Cluster 1 has a notably higher average comment length than the other clusters. Perhaps this can be a distinguishing factor between clusters 0 and 1? It may also be possible for this attribute to be a distinguishing factor between clusters 2 and 3. However, this is currently uncertain since cluster 3 only has one member.

- *Number of tags*

Excluding cluster 3, there seems to be no correlation between the cluster assignments and the number of tags provided. The students with high tag counts are distributed across clusters 0, 1, and 2, which suggests the attribute's lack of relevance in the cluster assignments. Tag counts for clusters 0, 1, and 2 all overlap one another within one stdev of their respective means.

- *Number of rates*

Clusters 0 and 1 comprise instances with a “high” (7+) number of ratings. However, there is no clear distinction in attribute values between clusters 0 and 1: $\mu_0 \pm \sigma_0$ overlaps with $\mu_1 \pm \sigma_1$ (i.e. overlap within one stdev from the means) for this attribute.

Clusters 2 and 3 seem to comprise instances with “low” numbers of ratings, although student 9 appears to be an outlier for this attribute (8 ratings). Some separation between clusters 2 and 3 seems to exist: $\mu_{2,editcnt} \pm \sigma_{2,editcnt}$ does not overlap with $\mu_{3,editcnt} \pm \sigma_{3,editcnt}$ (i.e. the two clusters do not overlap within one stdev from their means). However, $\mu_{2,editcnt} \pm 2\sigma_{2,editcnt}$ overlaps with $\mu_{3,editcnt} \pm 2\sigma_{3,editcnt}$ (i.e., overlaps within two stdevs from the means).

Interestingly, the action counts seem to play a larger role in determining cluster membership for this data set than edit/comment lengths.

To summarize, the following table summarizes the observed categorization of the students:

	Primary Categorization			**Secondary Categorization**
Clusters	*Edit Count*	*Comment Count*	*Ratings Count*	*Comment Length*
0	High	High	High	Low
1	High	High	High	High
2	Low	Low	Low	High
3	Low	Low	Low	Low

Table 5.13: Observed categorization of students for the MAS class

5.2.3.1.2 Justifications

- *Why are clusters 1 and 3 so small?*

For the singleton cluster 3, it was previously stated that the assignment appears to be suitable, since the student did not participate in the wiki collaborate phase. With values of 0 for every attribute, it is likely to be a relatively larger distance away from the other instances.

Also previously stated, cluster 1 did *not* appear to consist of students who appeared to be extreme outliers as with cluster 3. Rather, their values on the primary categorization attributes trend towards the lower end of the range for the corresponding attribute in cluster 0. And so rather than representing an exemplary peak of participation, the cluster may perhaps represent a niche "middle ground" between clusters 0 and 2.

- *Why do the clusters seem to be determined most from edit/comment/rating counts?*

One possible explanation for this emphasis on activity counts is that students may believe that the raw action counts factor into their grade. Thus, the number of edits may be artificially inflated via adding content in small increments and/or making multiple "minor" edits (e.g., fixing typos and text formatting) after the "meat" of the content is written. Comments may be similarly easy to make, particularly ones expressing agreement to an opinion or ones providing "obvious" remarks that require little insight. Since rates are also easy to perform, these may also be done in high quantity.

- *Why is the number of tags not a cluster indicator?*

Tags may be relatively tricky to contribute towards, as good ones are relevant to page content and/or related categories. It can be difficult to contribute additional tags when the few obvious tags are already entered, making it harder for students to "inflate" tag counts when they are not among the first collaborative contributors. As such, it would be difficult for this attribute to be correlated in the fashion of the other activity counts.

- *Why would edit and comment lengths have a lesser bearing on categorization? Is it due to the lack of correlation between edit/comment counts and lengths?*

Combined with the focus on activity counts, a general lack of correlation between edit and comment counts vs. lengths may contribute towards their apparent lack of importance in categorizing students in this data set. This may also be a result of student focus on activity count rather than length or quality of contributions.

5.2.3.1.3 Implications

What are the implications of the results for the MAS class? We’ve identified the following to be particularly prominent:

- *Need of a larger (consenting) sample size.*

As seen in the above results, half of the clusters are of considerably smaller size relative to the others, and a larger sample size will aid in confirming the validity of the observations and justifications drawn. In particular, having additional instances in clusters 1 and 3 will clarify the distinctions between clusters 0-1 and clusters 2-3, respectively. Additional instances for clusters 0 and 2 may also tighten the stdev/variances of the clusters, leading to more-specific categorization rules.

- *Potential usefulness of a custom (e.g., weighted) distance formula for X-means clustering.*

A potential avenue of investigation includes using a custom distance formula that emphasizes attributes that we value more, akin to the weighted sum used to emphasize edit and comment actions. While this introduces a bias to the clustering results, such clusters may be more valuable in the context of encouraging/emphasizing particular actions over others. Different activities might be weighted differently based on the instructor and assignment metrics, and the different weights serve to motivate students differently in their activities. So, a more prudent approach would be to incorporate these assignment scoring weights into the distance formula. For example, clusters based primarily on edit and comment count, length, and quality will be of greater interest for a collaborative writing course.

- *Leverage attributes of cluster membership to guide recommendations.*

The attribute range information for each of the clusters can be leveraged to guide recommendations for a user, relative to characteristics of its cluster peers. That is, if a student is identified to belong to a cluster that does not favor performing ratings or adding tags, then generating or presenting recommendations related to those actions could be a lower priority. It can also be used to guide “reminders” for particular actions when a student’s performance is lacking relative to its cluster peers. Caution should be taken to avoid “locking” students into a cluster – such recommendations may reinforce the student behaviors that place them into the cluster to begin with.

5.2.3.2 RAIK Clusters – Three Clusters

Performing the maximal pairs algorithm on the RAIK class data results in the following cluster assignments:

Cluster ID	Members (Student ID)
Cluster 0	0, 2, 3, 8, 11, 14
Cluster 1	1
Cluster 2	4, 5, 6, 7, 9, 10, 12, 13
Cluster 3	-

Table 5.14: Cluster membership for RAIK class

And the attribute details for each member, by cluster, are:

RAIK Cluster 0							
Student	RaikNumEdits	RaikEditLength	NumCmts	CmtLength	NumTags	NumRates	Dist. From Centroid
0	4	408.250	6	108.333	0	11	211.377
2	3	285.000	1	113.000	0	18	93.736
3	6	256.667	2	99.500	3	7	62.113
8	5	48.200	1	52.000	0	5	153.172
11	3	57.667	0	0.000	0	16	160.856
14	3	141.000	1	81.000	0	0	59.489
Avg μ_0	4.000	199.464	1.833	75.639	0.500	9.500	123.457
Std Dev σ_0	1.265	141.835	2.137	43.227	1.225	6.834	-

Table 5.15: Attribute details for RAIK cluster 0

RAIK Cluster 1							
Student	RaikNumEdits	RaikEditLength	NumCmts	CmtLength	NumTags	NumRates	Dist. From Centroid
1	5	106.400	4	33.000	34	16	0
Avg μ_1	-	-	-	-	-	-	-
Std Dev σ_1	-	-	-	-	-	-	-

Table 5.16: Attribute details for RAIK cluster 1

RAIK Cluster 2							
Student	RaikNumEdits	RaikEditLength	NumCmts	CmtLength	NumTags	NumRates	Dist. From Centroid
4	7	83.714	4	69.250	10	6	20.501

5	8	66.250	6	86.500	16	29	18.388
6	9	54.222	2	73.000	0	17	16.853
7	11	76.727	5	79.600	8	17	10.652
9	11	41.727	4	70.000	11	8	27.228
10	9	40.667	2	54.500	5	12	35.184
12	13	57.615	4	114.000	25	14	41.246
13	5	115.400	3	69.667	3	15	49.564
Avg μ_2	9.125	67.040	3.750	77.065	9.750	14.750	27.452
Std Dev σ_2	2.532	24.787	1.389	17.530	7.924	7.005	-

Table 5.17: Attribute details for RAIK cluster 2

5.2.3.2.1 Observations

In terms of quick, immediate impressions of the results, the presence of only three clusters in the results is particularly noteworthy. It seems unusual for our algorithm to end with three clusters when the target number of clusters for this data set (and consequently, the *MinK* used for X-means) is four.

As with the MAS class, the RAIK class also has a singleton cluster. However at a glance, it is difficult to tell whether this cluster assignment is appropriate for the instance, as its only noteworthy differences to the other clusters are:

- A relatively high number of tags.
- A relatively low average comment length.
- A relatively high average edit length.

Delving into the individual attributes, we have the following observations.

- *Number of edits*

Excluding the singleton cluster, there seems to be split based upon edit counts. Cluster 0 comprises the “low” edit counts and Cluster 2 comprises the “high” edit counts, and the separation is somewhat distinct: Clusters 0 and 2 do not overlap within one standard deviation, i.e. $\mu_0 \pm \sigma_0$and $\mu_2 \pm \sigma_2$ do not overlap. However, clusters 0 and 2 do overlap in two standard deviations, i.e. $\mu_0 \pm 2\sigma_0$ overlaps $\mu_2 \pm 2\sigma_2$.

Based on the above, Cluster 1 is aligned closest to cluster 0 for this activity.

- *Edit length*

Excluding the singleton cluster, there seems to be split along “high” and “low” edit lengths. The average edit length of cluster 0 is distinctly higher than that of cluster 2. However, the clusters overlap within one standard deviation, i.e., $\mu_0 \pm \sigma_0$ and $\mu_2 \pm \sigma_2$ overlap, due to cluster 0’s large standard deviation.

Note: outliers do exist in the cluster assignments, i.e. instances with “low” edit length in cluster 0 and “high” edit length in cluster 2. The following are the noted outliers of each cluster, and the subsequent cluster purity.

- Cluster 0: instances 8 and 11 (cluster purity = 4/6 = 0.667)
- Cluster 2: instance 13 (cluster purity = 7/8 = 0.875)

Cluster 1 is more closely aligned to cluster 0 than cluster 2 for this activity.

- *Number of comments*

There doesn’t seem to be a strong correlation between number of comments and cluster assignments. Clusters 0 and 2 overlap within one standard deviation. That is, the ranges for $\mu_0 \pm \sigma_0$ and $\mu_2 \pm \sigma_2$ overlap.

However, the following was observed:

- Majority of cluster 0 assignments (5/6) have 0-2 comments
- Majority of cluster 2 assignments (6/8) have 2+ comments

For this action, cluster 1 is more-closely aligned with cluster 2.

- *Comment length*

There doesn’t seem to be a strong correlation between comment length and cluster assignments. Clusters 0 and 2 overlap within one standard deviation, i.e., the intervals $\mu_0 \pm \sigma_0$ and $\mu_2 \pm \sigma_2$overlap.

A relatively low average comment length may be cluster 1’s distinguishing characteristic. However, this currently cannot be confirmed due to the lack of members in this cluster.

- *Number of tags*

Excluding the singleton cluster, there appears to be a separation based on the number of tags added.

Cluster 0 comprises the “low” tag counts, and cluster 2 comprises the relatively higher tag counts. The intervals for the two clusters do not overlap within one standard deviation. That is, $\mu_0 \pm \sigma_0$ and $\mu_2 \pm \sigma_2$do not overlap. However, the two clusters overlap within two standard deviations. That is, $\mu_0 \pm 2\sigma_0$ and $\mu_2 \pm 2\sigma_2$overlap. It should be noted that cluster 2 has a relatively large standard deviation on this action and contains instances whose values would be closer to those of the members of cluster 0.

For this action, cluster 1 is more-closely aligned with cluster 2. However, the number of tags its instance performed is the highest out of all instances, suggesting that this may be a defining characteristic of the cluster.

Overall, this attribute may be of secondary importance in determining cluster membership for this data set.

- *Number of rates*

There doesn't seem to be a strong correlation between number of ratings given and cluster assignments. The number of ratings for all three clusters overlaps within one standard deviation from their averages.

To summarize, the following table summarizes the observed categorization of the students:

	Primary Categorization		**Secondary Categorization**		
Clusters	*Edit Count*	*Edit Length*	*Comment Count*	*Comment Length*	*Tags Count*
0	Low	High	Low	High	Low
1	Low	High	Mid/High	Low	High
2	High	Low	High	Low/Mid	High

Table 5.18: Observed categorization of the RAIK class

5.2.3.2.2 Justification

- *Why 3 clusters?*

To re-iterate the pseudocode for our merging process:

```
•   While number of unique clusters remaining in C is greater than k:
        ○   For each cluster A in C:
                ▪   For each cluster B in C:
                        •   If A != B AND HasStrongMaximalPair( A, B ) == TRUE
                                ○   Merge( A, B )
```

Note that the check for the number of clusters remaining is on the outermost loop. That is, our algorithm currently stops after *all* clusters with strong max pairs are merged for a particular iteration of the while-loop. With this, it's possible for the number of clusters to go from above the target number to below during a single iteration. In this particular case, the final iteration of our algorithm started with five clusters and performed two merges, resulting in three clusters.

- *Why the emphasis on edit- and comment-related attributes?*

There seems to be particular emphasis on edit- and comment-related actions when categorizing the students in this class. This is justifiable since 90% of the collaborative contribution grade is based solely upon edits and comments, and this students aiming for a good score will prioritize these actions.

- *"Quantity vs. quality"*

A particular feature of this data set is that it seems to be split between two particular editing /commenting paradigms: 1) low count, high length and 2) high count, low length. The cluster results here demonstrate that there may be a tradeoff between number of edits/comments versus average edit/comment length – that is, due to limited time and/or cognitive resources, students will either make few large contributions or many small contributions to the wiki pages. This tradeoff is expounded upon further in Section 5.3.

- *Why the singleton cluster?*

Previously, our impression was that the purpose of the singleton cluster wasn't readily apparent. After examining the cluster instance's attributes relative to the others, it appears that *the cluster may represent the middle region between the relative extremes of low counts with high lengths and high counts with low lengths*. In particular, it shares the low count and high length paradigm for edits, but high count and low length for comments.

5.2.3.2.3 Implications

The following implications were derived from the maximal pairs algorithm results.

- *Need of a larger (consenting) sample size.*

As with the MAS results, our procedure resulted in another set of clusters where one is a singleton cluster. A larger sample size will also aid in: 1) confirming the validity of the observations and justifications drawn, and 2) further distinguishing cluster 1's attributes for this data set. See the corresponding implication for the MAS class for further detail.

- *Use of a custom (e.g., weighted) distance formula for X-means clustering.*

As with the MAS class, tailoring the distance formula used for generating the clusters may increase their value to users leveraging the cluster assignments for decision making. It may be particularly relevant in this case since the student collaborative contributions are graded in a slightly different manner between the MAS and RAIK class, i.e. the number of tags is not part of the grade for the RAIK class. Refer back to the corresponding section in the MAS results for additional discussion.

- *Leverage attributes of cluster membership to guide recommendations.*

As with the MAS class, recommendations can be guided based on student membership to these clusters. See the corresponding discussion in the MAS class results for further detail.

- *Possible need for a different stopping condition for our algorithm?*

This data set highlights the issue that it is possible for our procedure to end with fewer clusters than the target number desired. With this, the possibility exists for it to end with much fewer clusters than the target number, e.g., ending with one or two clusters when four or five are desired, and can arise when multiple multi-cluster maximal pairs are available. While the probability of our algorithm resulting in much fewer clusters than the target number is rare, we wish to re-evaluate the stopping condition and determine whether results significantly differ after changing it.

Due to this inconsistency between the target number of clusters and the number of clusters in our algorithm results, we subsequently evaluated the results arising from stopping our algorithm at the point when exactly four clusters remain.

The four-cluster results are largely similar to these three-cluster results, with two of the clusters formed by "reverting" one of the clusters from 5.2.3.2 to a pre-merge state. While this "split" identifies an additional level of granularity when categorizing students, it does not exhibit any strong deviations from the observations and conclusions drawn from the three-cluster data. For further detail, please refer to Appendix B.

5.2.4 Summary

In this section, we introduced and applied our Maximal Pairs Algorithm to the results obtained from the X-Means clustering algorithm, using the raw tracked attribute data collected from student activity as the attributes upon which the clusters are formed. To summarize the highlights of the categorization:

- For the MAS class, categorization appears to be primarily based upon the number of edits, comments, and ratings performed. Secondary categorization appears to be based on comment length.
- For the RAIK class, categorization appears to be primarily based on the number and average lengths of edits. Secondary categorization appears to be based on comment count, comment lengths, and number of tags contributed.

As we can see, attributes pertaining to edits and comments are a common factor in clustering the students. This is consistent with our expectations that student behavior is centered on these attributes, since those two collaborative elements comprise 90% of the student’s grade for the wiki assignment.

This leads to the question: why are there differences in categorization factors despite the collaborative wiki assignment being similarly structured and graded between the two classes? A possible answer is that the general approach or strategy taken by the students for each class is different.

With the RAIK students, emphasis is placed on edits which covers the number of edits, the average edit lengths, and the tradeoff between the two. A similar but lesser emphasis is placed on comments as a secondary categorization, with a similar tradeoff between number and length observed in the clusters formed. This is in line with the expectation for edits to be prioritized due to their 60% weight in the grade and for comments to be prioritized to a lesser degree due to their 30% weight.

The MAS class may be focused on the counts of the different actions instead, with students possibly believing that performing the actions more times results in a better grade. Particularly, the primary categorization for this class is based on number of edits, number of comments, and number of ratings, all three of which correspond to the three action types contributing towards the collaboration grade for the assignment. With this in mind, these actions are still in line with expectations for the assignment, even if the clusters are formed on a different basis than the RAIK class.

One interesting thing to note is that while the average number of tags created is relatively similar between the two classes, the average number of ratings performed for the MAS class is significantly lower (by approximately 50%). Recall in chapter 5.1.2 that the grading for the final 10% of the assignment differs slightly between the two classes. Specifically, the final 10% for the RAIK class is based solely on ratings provided whereas the final 10% for the MAS class consists of ratings, tags, and views. With this difference in criteria, *the RAIK class places greater emphasis on the number of ratings given, since the 10% is based solely on number of ratings.* On the other hand, this 10% is "spread" between ratings, tags, and views for the MAS class, and the average number of ratings is considerably lower as a result.

Overall, what does all of this suggest? On one hand, factoring in the idea of different strategies towards the assignment improves user modeling for recommendations, since students would be more receptive towards recommendations that are in line with their strategy. Also, this confirms to a certain extent that the instructor's expectations, e.g. grading requirements, for the wiki assignment will correlate to student behavior on the wiki.

Student behaviors appear to be motivated by factors such as the evaluation criteria of the wiki assignment, "quantity vs. quality", and the relative "ease" of making a particular contribution vs. others.

Student resources for working on the assignment, e.g., time and effort, are generally limited, and the total amount of such resources vary from student to student based on their individual schedules. Thus the factors listed may intuitively guide how they allocate these resources among the different wiki activities.

The discussion for the first item, evaluation criteria of the wiki assignment, has been previously covered in the categorization discussion in this sub-chapter (5.2.4). Please refer back to the previous paragraphs for details.

Working on the assumption that student resources for working on the wiki assignment are limited, "quantity vs. quality" refers to the tradeoff between the quality (e.g. thoroughness, meaningfulness, and level of insight displayed) of the contribution and the number of contributions made. Students with "more" total resources can appear to have more contributions of better quality than those with "less" total resources. The degree to which students focus on quality or quantity may be affected by the "ease of contribution", discussed next.

"Ease" of contribution pertains to the how "easy" it is to make a particular contribution towards a page, and this may vary depending on the action taken and its timing. For instance, it may be "easier" to make a relatively high-quality edit to a page that is initially poorly written, and "harder" to make an impactful edit to a well-written page. Additionally, it is "easier" to rate or write tags for a page than it is to make an edit or a comment. Timing plays a role in that students who act earlier have access to more "easy" contributions than those who contribute later.

While the instructor can influence student behavior via assignment requirements and evaluation criteria, student behavior can also be influenced by the behavior of their peers.

Clusters found via this approach can be leveraged to guide recommendations and profiling of students.

As previously alluded to in the discussions of the individual classes, the clusters found via this approach can be used to guide student profiling and recommendations. Since the clustering is based on behavior, the clusters can be used as "behavioral archetypes" for tailoring assistance towards *groups* of users. Recommendations can thus be generated according to the goals of the instructor, either by generating recommendations that the user would be likely to take (reinforcing the archetype) or helping the user develop new habits outside of her current behavior (breaking the archetype).

5.3 Active vs. Passive Activity Profiles

Relating the pre-analysis active/passive and minimalist/overachiever categories to the actions tracked within the wiki, we consider the categorization of active/passive actions to be based upon the *count* and *type* of actions performed. That is, we will count the number of active actions (i.e. comments and edits) and the number of passive actions (i.e. rates and add tags), and based on how the two compare for a given student, a categorization is made. However because the passive actions can be easily made on a larger order of magnitude than the active ones, we weight the total active actions prior to the comparison. We propose the following weighted sum to calculate the *weighted total collaborative actions*:

$$\begin{aligned}&\text{Weighted Total Collab. Actions}\\&= (3 * \#Edits) + (2 * \#\,Comments) + (\#\,Add\ Tag) + (\#Rate)\end{aligned}$$

Note that this categorization based on *action* ***types*** is different from categorizing based on overall activity level, i.e. number of actions performed. For instance, a user with only 1 edit as her sole collaborative activity has a type distribution of 100% active, whereas a user with 10 edits and 30 rates would have a type distribution of 50% active. In terms of activity level, the latter user has a greater activity count, but the former has a larger proportion of activities of the “active” type. However, the person with 10 edits and 30 rates is the “better” collaborator.

Tables 5.18 and 5.19 show the derived counts for the two action types in the RAIK and MAS classes, respectively.

5.3.1 RAIK Class

Student ID	Rate	Comment	Add Tag	Edit	Total Active Actions (Edits + Comments)	Weighted Total Active (3 * Edits + 2 * Comments)	Total Passive Actions (Rate + Add Tag)	% Active Actions
1	11	6	0	4	10	24	11	69%
2	16	4	34	5	9	23	50	32%
3	18	1	0	3	4	11	18	38%
4	7	2	3	6	8	22	10	69%
5	6	4	10	7	11	29	16	64%
6	29	6	16	8	14	36	45	44%
7	17	2	0	9	11	31	17	65%
8	17	5	8	11	16	43	25	63%
9	5	1	0	5	6	17	5	77%
10	8	4	11	11	15	41	19	68%
11	12	2	5	9	11	31	17	65%
12	16	0	0	3	3	9	16	36%
13	14	4	25	13	17	47	39	55%
14	15	3	3	5	8	21	18	54%
15	0	1	0	3	4	11	0	100%

Table 5.19: RAIK active/passive actions during collaboration phase

The clustering results from our Maximal Pairs Algorithm is as follows:

RAIK Cluster 0

Student ID	Rate	Comment	Add Tag	Edit	Total Active Actions (Edits + Comments)	Weighted Total Active (3 * Edits + 2 * Comments)	Total Passive Actions (Rate + Add Tag)	% Active Actions
1	11	6	0	4	10	24	11	69%
4	7	2	3	6	8	22	10	69%

5	6	4	10	7	11	29	16	64%
7	17	2	0	9	11	31	17	65%
8	17	5	8	11	16	43	25	63%
9	5	1	0	5	6	17	5	77%
10	8	4	11	11	15	41	19	68%
11	12	2	5	9	11	31	17	65%
15	0	1	0	3	4	11	0	100%

Table 5.20: RAIK active/passive action categorization – Cluster 0

RAIK Cluster 1								
Student ID	Rate	Comment	Add Tag	Edit	Total Active Actions (Edits + Comments)	Weighted Total Active (3 * Edits + 2 * Comments)	Total Passive Actions (Rate + Add Tag)	% Active Actions
6	29	6	16	8	14	36	45	44%
13	14	4	25	13	17	47	39	55%
14	15	3	3	5	8	21	18	54%

Table 5.21: RAIK active/passive action categorization – Cluster 1

RAIK Cluster 2								
Student ID	Rate	Comment	Add Tag	Edit	Total Active Actions (Edits + Comments)	Weighted Total Active (3 * Edits + 2 * Comments)	Total Passive Actions (Rate + Add Tag)	% Active Actions
2	16	4	34	5	9	23	50	32%
3	18	1	0	3	4	11	18	38%
12	16	0	0	3	3	9	16	36%

Table 5.22: RAIK active/passive action categorization – Cluster 2

5.3.1.1 Observations

The following were observed from the RAIK data.

- *Clusters*:
  - Cluster 0 – 9 students. Students whose weighted activity profiles consist primarily of "active" actions, i.e. edits and comments. Appears to contain % Active Actions ≥ 63%.

  - Cluster 1 – 3 students. Students whose weighted activity profiles are relatively "balanced" between those of Clusters 0 and 2.
  - Cluster 2 – 3 students. Students whose weighted activity profiles consist primarily of "passive" actions, i.e. tags and ratings. Appears to contain students with % Active Actions ≤ 38%.

- *Correlations*

In addition to the categorizations observed from performing our Maximal Pairs algorithm on the % active actions value, two key correlations were observed. First, there is a correlation of 0.531 between total active actions and total passive actions. Second, there is a negative correlation (-0.591) between the total number of unweighted collaborative actions and % active actions. These are relatively strong and worth investigating for underlying implications.

5.3.1.2 Justification

Biases towards particular actions?

It is likely that a bias towards the active actions is introduced in student behavior since the majority of the assignment grade is based on edits and comments. Students that have a high action count for both passive and active types may either be overachieving (see minimalist/overachiever section ahead) or may be compensating for a lack of quality in each of their actions performed (see next justification point).

Why the positive correlation between total active actions and total passive actions?

The correlation between total active actions and total passive actions (and consequently a possible cause for the high number of students with “balanced” action profiles) may be due to the following:

- Since grades are also based on overall contribution quality, students who are unable to make a sizeable contribution in “one shot” (i.e. late contributors) may attempt to compensate with smaller contributions in greater numbers.
- Lacks in edit and comment quality may also be (somewhat) compensated for with the minor activities (i.e. rating and tagging) since they also contribute a small portion to the collaboration grade (10%).
- There may be students who believe that action counts factor into their grade, so the number of edits may be artificially inflated via adding content in small increments. Since tags and rates are also easy to perform, these may also be done in high quantity.

Why the negative correlation between the total number of unweighted collaborative actions and % active actions?

The negative correlation between the total number of unweighted collaborative actions and % active actions may arise from the following:

- Students who are able to make sizeable, significant, and/or high quality contributions in a minimal number of edits/comments do not need to make additional actions for a “good grade.” As such, they may not be motivated to perform many passive actions (i.e. rates and tags). With the low number of edits and comments made dominating their low number of total actions, their profile will thus have a high percentage of active actions.

- As previously mentioned, a student that makes edits and comments lacking in quality may be motivated to perform as many actions as possible to compensate for them. Since ratings and tags are easier to provide, they can be carried out in higher quantity, thus increasing the number of total actions and decreasing the percentage of active actions in the users’ profiles.

5.3.1.3 Implications

Tradeoff between percentage active actions and collaboration initiation rate.

Students with a higher active action percentage have a lower collaboration initiation rate (i.e. fewer total collaborative actions, as suggested in the third justification point). Conversely, students with a lower active action percentage have a higher collaboration initiation rate (i.e. greater total collaborative actions, as suggested in the second justification point).

Cross comparison with the MAS class will be necessary to determine if this implication is true only for the RAIK class or if it may be applicable in general.

Usefulness and adequacy of the active action percentage metric?

The percent active metric (and subsequent categorization based upon it) is not adequate on its own to profile a student, since it does not account for the absolute quantity of the active/passive actions performed. However, this metric may be useful to instructors by providing the instructor with a quick overview of how the student is performing on the collaborative assignment. For instance in assignments emphasizing editing wiki pages and participating in discussions, a low % active actions may serve as a “red flag” identifying students who may be struggling with making such contributions.

The instructor may also tweak the weights used in our metric according to the assignment criteria and desired student behavior.

For future work, the raw total passive actions and (weighted) active actions can be leveraged for a metric addressing this deficiency.

5.3.2 MAS Class

Student ID	Rate	Comment	Add Tag	Edit	Total Active Actions (Edits + Comments)	Weighted Total Active (3 * Edits + 2 * Comments)	Total Passive Actions (Rate + Add Tag)	% Active Actions
1	9	6	0	5	11	28	9	76%
2	9	5	27	4	9	23	36	39%
3	18	6	5	3	9	24	23	51%
4	5	3	6	3	6	15	11	58%
5	7	6	5	4	10	26	12	68%
6	7	4	12	5	9	22	19	54%
7	7	6	12	4	10	26	19	58%
8	4	2	4	4	6	14	8	64%
9	17	7	4	5	12	31	21	60%
10	8	3	14	0	3	9	22	29%
11	5	3	1	2	5	13	6	68%
12	1	3	16	3	6	15	17	47%
13	0	0	0	0	0	0	0	0%
14	4	6	4	2	8	22	8	73%
15	9	8	14	8	16	40	23	63%
16	1	3	2	2	5	13	3	81%
17	5	1	5	4	5	11	10	52%

Table 5.23: MAS active/passive actions during collaboration phase

The results from clustering the students based on the % Active Actions value via the Maximal Pairs Algorithm is as follows:

MAS Cluster 0								
Student ID	Rate	Comment	Add Tag	Edit	Total Active Actions (Edits + Comments)	Weighted Total Active (3 * Edits + 2 * Comments)	Total Passive Actions (Rate + Add Tag)	% Active Actions
2	9	5	27	4	9	23	36	39%
10	8	3	14	0	3	9	22	29%
13	0	0	0	0	0	0	0	0%

Table 5.24: MAS active/passive action categorization – Cluster 0

MAS Cluster 1								
Student ID	Rate	Comment	Add Tag	Edit	Total Active Actions (Edits + Comments)	Weighted Total Active (3 * Edits + 2 * Comments)	Total Passive Actions (Rate + Add Tag)	% Active Actions
1	9	6	0	5	11	28	9	76%
3	18	6	5	3	9	24	23	51%
4	5	3	6	3	6	15	11	58%
5	7	6	5	4	10	26	12	68%
6	7	4	12	5	9	22	19	54%
7	7	6	12	4	10	26	19	58%
8	4	2	4	4	6	14	8	64%
9	17	7	4	5	12	31	21	60%
11	5	3	1	2	5	13	6	68%
12	1	3	16	3	6	15	17	47%
14	4	6	4	2	8	22	8	73%
15	9	8	14	8	16	40	23	63%
16	1	3	2	2	5	13	3	81%
17	5	1	5	4	5	11	10	52%

Table 5.25: RAIK active/passive action categorization – Cluster 1

5.3.2.1 Observations

The following was observed in the MAS data.

- *Clusters*:
  - The target *MinK* value for this class was determined to be two instead of the expected three.
  - Cluster 0 – 3 students. Consists of the students with the most “passive” activity profiles. All students in this category have a score $\leq$ 39%.
  - Cluster 1 – 14 students. Consists of students not in Cluster 0, i.e. % Active Action $>$ 39%.

- *Correlations*

In addition to identifying categories based on % active actions, we also check whether the correlations discovered with the RAIK data also apply to the MAS class. As with the other data set, there is a relatively strong correlation (0.519) between total active actions and total passive actions in the MAS class. However, unlike the RAIK class there is little/no correlation (0.083) between % active actions and unweighted total actions.

5.3.2.2 Justification

Biases towards particular actions?

As with the RAIK class, it is likely that a bias towards the active actions is introduced in student behavior since the majority of the assignment grade is based on edits and comments (60% edits, 30% comments, 10% passive actions). Students that have a high action count for both passive and active types may either be overachieving or may be compensating for a lack of quality in each of their (active) actions performed (see next justification point).

Why the positive correlation between total active actions and total passive actions?

The correlation between total active actions and total passive actions is also observed in the MAS class. Please refer back to 5.3.1.2 for additional details.

5.3.2.3 Implications

Usefulness and adequacy of the active action percentage metric?

As previously stated in the RAIK implications, the percent active actions metric (and subsequent categorization based upon it) is not adequate on its own to profile a student, since it does not account for the absolute quantity of the active/passive actions performed. See sub-chapter 5.3.1.3 for additional discussion.

5.3.3 Summary

In this sub-chapter, we introduced the active vs. passive approach for categorizing student activity profiles, and it is based on the concept of "active" (e.g., edits and comments) and "passive" (e.g., rates and tag adding) action *types*. This metric computes the *percentage* of actions in student's activity profiles that are of the "active" type. **Note that this is different from computing the frequency at which students perform actions.** There are three discussion topics introduced in this sub-chapter:

A positive correlation exists between the number of total active actions and total passive actions.

Students generally have limited resources (time, attention, etc.) when doing assignments for the class, so it is assumed that there will be a tradeoff between the number of active and passive actions due to the larger resource cost to perform the active actions. However, this is shown to be incorrect in both RAIK and MAS classes – contrary to expectations, there is a positive correlation between the two values. In order to reconcile this, **the concept of "making an edit/comment" needs to be separated from the concept of "making a *high quality* edit/comment."**

What is the motivation for a student to behave in this manner? Reiterating the justification from section 5.3.1.2:

- Since grades are also based on overall contribution quality, students who are unable to make a sizeable contribution in "one shot" (i.e., late contributors) may attempt to compensate with smaller contributions in greater numbers.
- Lacks in edit and comment quality may also be (somewhat) compensated for with the minor activities (i.e., rating and tagging) since they also contribute a small portion to the collaboration grade (10%).
- There may be students who believe that action counts factor into their grade, so the number of edits may be artificially inflated via adding content in small increments. Since tags and rates are also easy to perform, these may also be done in high quantity.

Tradeoff (or lack thereof) between the total number of unweighted collaborative actions and % active actions.

In the results for the RAIK class, a tradeoff was found between the % active actions metric and the total number of unweighted collaborative actions. At a glance, this result seems contradictory to the previous finding that total active actions and total passive actions are positively correlated – if the two are correlated and there are a large number of both active and passive actions in the user's activity profile, then wouldn't the active actions made increase the active %, resulting in a positive correlation between the two? To reconcile this, we need to consider that **although total active actions and total passive actions are positively correlated, active actions still cannot be performed as quickly as passive actions.**

Reiterating the justifications from sub-chapter 5.3.1.2:

- Students who are able to make sizeable, significant, and/or high quality contributions in a minimal number of edits/comments do not need to make additional actions for a "good grade." As such, they may not be motivated to perform many passive actions (i.e. rates and tags). With the low number of edits and comments made dominating their low number of total actions, their profile will thus have a high percentage of active actions.
- As previously mentioned, a student that makes edits and comments lacking in quality may be motivated to perform as many actions as possible to compensate for them. Since ratings and tags are easier to provide, they can be carried out in higher quantity, thus increasing the number of total actions and decreasing the percentage of active actions in the users' profiles.

Interestingly, the MAS class did not exhibit this negative correlation between % active action and unweighted total actions. Rather, the two appear to lack any correlation in this data set. Why might this be the case? The key may be in the slightly different grading criteria in the final 10% of the collaboration grade. Recall that for the RAIK class, the final 10% is comprised solely of the ratings action. In the MAS class, rates, tags, and views all comprise the 10%. **The addition of tags to the MAS assignment grading criteria motivates grade-conscientious students to also add tags. In the RAIK class, this action had no bearing on the grade and was thus ignored by students able to make high quality contributions in a minimal number of actions.**

5.4 Minimalist vs. Overachiever Activity Profiles

In this section, we re-introduce and expand upon our metric of determining student "effort" on the collaborative writing assignment. Determining the "minimalists" and the "overachievers" in the class has benefits that could aid in recommendation and teaching, such as identifying the work of overachieving students as "recommended reading" within the wiki, and alerting the instructor when students performing minimal work are identified.

As previously mentioned in Section 5.1.3, we defined the minimalist-overachiever scale to be based on the user's actions relative to the minimum requirements of the collaborative writing assignment. For both classes, this minimum requirement is to perform collaborative actions (editing, participating in threaded discussions, rating, and tagging) on three different wiki pages. The assignment also specifies that edits must be made to at least three other pages. Based on these two criteria, a student may choose to make three edits total, each on a different page, as the minimum effort.

There are a couple difficulties in basing this metric on the minimum requirements of the assignment. First, the baseline for minimalism in this particular assignment is relatively low. Second, the relative scale for the number of collaborative actions performed is also low (on the scale of tens), due to the limited time span of the assignment and limited student resources. With these two traits, it is difficult to determine whether a student is truly "overachieving" relative to the instructor's expectations. We thus decide to determine minimalists and overachievers by comparing the students' collaborative activities against that of their peers.

The process for determining minimalists and overachievers among the students in a class is as follows:

1. Cluster the students in the class based on each of the tracked attributes. This clustering identifies the different activity levels (e.g. high, medium, and low) among the students for each attribute, highlighting areas where a student is performing much or relatively little.
2. Map student placement within each attribute to a score between -1 (low) and +1 (high), based on the number of clusters for the attribute.

Number of Clusters	Categorizations
1	Don't Care (0)
2	Low (-1), High (+1)
3	Low (-1), Mid (0), High (+1)
4	Low (-1), Low-Mid (-0.33), High-Mid (+0.33), High (+1)
5	Low (-1), Low-Mid (-0.5), Mid (0), High-Mid (+0.5), High (+1)

Table 5.26: Possible activity level categorizations, based on the optimal number of clusters for an attribute.

3. Calculate the net minimalist/overachiever score for each student by summing their scores on the individual attributes.
4. Cluster the students based on the sums calculated in Step 3.

The principle behind this particular scoring scheme is for "High" placements and "Low" placements to counteract one another. Since underperformance on one attribute can be compensated for by overachieving in another, a positive net score is indicative of an overachiever, and a more positive score indicates stronger overachieving tendencies. A negative one indicates minimalist behavior, with a more negative score indicating stronger minimalistic tendencies. A net score near zero is indicative of a balance between the two.

In addition to evaluating the “quantitative” aspects of the students’ activities via the scoring discussed above, we also wish to consider the “quality” of the user’s contributions. We thus manually evaluate the content of each user’s prose (i.e., edits and comments) and include it as two additional categories contributing towards the net score: edit quality and comment quality.

In the following sub-sections, we apply this metric to the students in the RAIK and MAS classes.

5.4.1 RAIK Class

After performing the maximal pairs clustering on the RAIK class for each of the tracked attributes, we have the following activity placements:

Student ID	Edit Count (2)	Edit Length (3)	Edit Quality (3)	Cmt Count (2)	Cmt Length (2)	Cmt Quality (3)	Rates Count (2)	Tags Count (2)	Net Score
1	L	H	H	H	H	H	H	L	+4
2	L	M	L	H	L	L	H	H	-1
3	L	H	H	L	H	M	H	L	+1
4	L	H	H	L	H	M	L	L	-1
5	L	M	M	H	L	H	L	L	-2
6	H	L	H	H	L	H	H	H	+4
7	H	L	H	L	L	H	H	L	0
8	H	L	H	H	L	M	H	L	+1
9	L	L	L	L	L	L	L	L	-8
10	H	L	M	H	L	H	L	L	-1
11	H	L	M	L	L	L	H	L	-3

12	L	L	L	L	L	L	H	L	-6
13	H	L	M	H	H	M	H	H	+4
14	L	M	M	H	L	M	H	L	-1
15	L	M	M	L	L	H	L	L	-4

Table 5.27: RAIK student activity level for each tracked attribute. The number following each attribute, i.e. (2), indicates the number of clusters used to categorize the students. See Table 5.25 for the score values mapped to each activity level.

Clustering the students on their net scores gives us the following clusters:

RAIK Cluster 0		
Student ID	**Net Score**	**Collaboration Phase Score**
1	+4	100
6	+4	95
13	+4	95
Average	**+4**	**96**
RAIK Cluster 1		
Student ID	**Net Score**	**Collaboration Phase Score**
2	-1	85
3	+1	95
4	-1	100
5	-2	100
7	0	85
8	+1	*
9	-8	85
10	-1	90
11	-3	85
12	-6	50
14	-1	90
15	-4	*
Average	**-2.08**	**86.50**

Table 5.28: Clusters and members for the RAIK class, based on minimalist vs. overachiever net score. Collaboration phase score is also listed for comparison. Asterisks denote users who did not consent to the use of assignment grade for this analysis.

5.4.1.1 Observations

The two clusters formed by the Maximal Pairs algorithm appear to consist of the extreme overachievers for the class (Cluster 0) and the remaining students (Cluster 1). Due to this basis for grouping the students, there is a relatively lopsided distribution between the two, with 20% of the students placed in Cluster 0 and 80% placed in Cluster 1.

We are interested in whether the students' net scores correlate to the grades they received during the collaboration phase of the wiki assignment, and the grades for consenting students are thus listed alongside their net scores in Table 5.27. Excluding the non-consenting users from the calculation, the correlation coefficient between the net score and collaboration phase score is +0.529. It should also be noted that the average grade of the overachievers' cluster is higher than that of the other cluster.

Additionally, we are also interested in whether the quantity vs. quality tradeoff identified in previous sections is still present in these results. In this analysis, we identify tradeoffs (or negative correlation) between two attributes to be a situation where one attribute is ranked high (H) and the other is ranked low (L). Consequently, if either are mid ranked (M), then it does not count. We also identify positive correlation between two attributes to be a situation when the rank for the two attributes match, such as when both are H, both are M, or both are L.

The following were observed in the categorizations for edits and comments specifically:

- Edits
    - A tradeoff between count and length exists for 9 out of 15 students. There are 2 students for which count and length are positively correlated. However, it should be noted that these students with positive correlations are extreme minimalists, which is relatively rare. The relation between count and length is thus distinctly a tradeoff.

  - For edit quality vs. count, 3 out of 15 students exhibit a tradeoff between the two attributes. However, the two appear to be positively correlated (i.e., quality level matches count activity level) for 6 out of 15 students. No pattern between the positive/negative correlation and net scores appears to exist. Consequently, there does not appear to be a distinct relation between edit quality and count.
  - For edit quality vs. length, 8 out of 15 students exhibit a positive correlation (i.e. length level matches quality level) between edit quality and length. For 3 out of 15 students, there is a tradeoff between these categories instead. There does not appear to be a distinct pattern between positive/negative correlation and student net score.
- Comments
  - The tradeoff between comment count and length is observed in 8 out of 15 students. The remaining 7 students appear to have a positive correlation instead. While this would typically be considered as being an indistinct relation due to nearly equal numbers on opposing ideas, a correlating pattern was found: *students on the extreme ends of the spectrum* (e.g., greater than +4 or lower than -3) *exhibited the positive correlation, whereas the students in the middling range exhibit the tradeoff*.
  - For comment quality vs. count, only 3 out of 15 students exhibit a tradeoff in the two attributes. A positive correlation is observed for 6 of the 15 students. There does not appear to be a distinct relation between the two attributes.

- For comment quality vs. length, 5 of the 15 students exhibit matching ranks between the two attributes. However, a tradeoff is also exhibited in 5 of the other students. There does not appear to be a distinct relation between the two attributes.

5.4.1.2 Justifications

Why is there a positive correlation between Net Scores and Grades?

Since the net scores are based upon the same criteria for assignment grades, there is a positive correlation between the two. In particular, the grading by the instructor places particular emphasis on counts (i.e. comments and edits across 3 pages, for a minimum of three edits and three comments) and content quality.

Quantity vs. Quality

As previously highlighted, there is a tradeoff between the contribution count and contribution length for edits and comments, and the tradeoff is more immediately apparent for edits. The tradeoff on edits can be explained per previous discussion in sub-chapter 5.2.3.2.2. More interestingly, the tradeoff between comment count and length appears to be dependent on the student’s net score: the extreme minimalists and overachievers have a positive correlation between comment count and length, whereas the students in between the extremes exhibit the tradeoff. Intuitively, this can be justified as extreme overachievers having the motivation to make many lengthy comments, whereas extreme minimalists generally lack the drive to bother.

However, there is a lack of a tradeoff between length and content quality for both edits and comments in general. This may be attributed to the two not being particularly correlated, at least for this class. Intuitively, using more words is associated with more-detailed explanations, which consequently leads to the idea that a longer contribution is of higher quality. However this may be a misconception in actuality: valuable contributions can be made in relatively few words if the writer is concise, and contribution length can be bloated while adding little value if the writer is long-winded. This can similarly explain the lack of correlation between contribution count and content quality.

5.4.1.3 Implications

Usefulness of Two Clusters

The Maximal Pairs algorithm identified 2 as the optimal number for categorizing this class on the students' net scores. Upon closer inspection, there are three possible outcomes when clustering with 2 as the optimal number:

- One cluster containing extreme overachievers, one cluster containing everyone else
- One cluster containing extreme minimalists, one cluster containing everyone else
- One cluster containing overachievers (both extreme and slight), one cluster containing minimalists (both extreme and slight)

In this particular case, the clusters formed are one with extreme overachievers and one with the rest of the students. While this can be useful for identifying which students' works to highlight (average grade of 96), these clusters are not as useful for identifying minimalists for the instructor. Of the listed outcomes for 2 clusters, only the second would be particularly useful for this scenario, as “slight” minimalistic tendencies might not warrant particular attention from the instructor.

A possible avenue of future investigation is “forcing” the number of cluster results to 3. With it, it is possible for the clustering to group the students into extreme overachievers, extreme minimalists, and the remaining students in between. The results of such a clustering would be more valuable for application to the suggested scenarios in the section introduction.

Using Net Scores as Grade Predictors

One of the potential uses of the net score is to assist instructors in grading student activity in the wiki. However, the current calculation of the net score may not be sufficient for this task.

The range of collaboration phase grades for the majority of students is also relatively narrow, spanning from 85 to 100 points. Although the calculated correlation between the net score and grade is +0.529, one can see that some of the grades received may not align with expectations arising from net score, when using those two values as the benchmarks for overachieving and minimalist behavior. Thus, some modifications to the net score calculation will be needed before we can confidently use it as an indicator for grades.

One possible area for modification would be in the weighting of the individual tracked attributes. Currently, each category of tracked activity is equally weighted with one another, contributing up to +/- 1, i.e. 1/8th, of the total net score. By modifying the weights in this sum to reflect assignment expectations, the net score may better reflect the grades students should receive.

5.4.2 MAS Class

After performing the maximal pairs clustering on the RAIK class for each of the tracked attributes, we have the following activity placements:

Student ID	Edit Count (1)	Edit Length (2)	Edit Quality (3)	Cmt Count (2)	Cmt Length (2)	Cmt Quality (3)	Rates Count (2)	Tags Count (2)	Net Score
1	M	H	H	H	H	H	H	L	+5
2	M	L	M	H	H	H	H	H	+4
3	M	L	M	H	H	M	H	L	+1
4	M	L	L	L	H	H	L	L	-3
5	M	L	M	H	H	M	H	L	+1
6	M	H	M	L	H	H	H	H	+4
7	M	L	L	H	L	M	H	H	0
8	M	H	M	L	H	M	L	L	-1
9	M	L	H	H	H	H	H	L	+3
10	M	L	L	L	H	H	H	H	+1
11	M	H	H	L	H	H	L	L	+1
12	M	L	M	L	H	M	L	H	-1
13	M	L	L	L	L	L	L	L	-7
14	M	L	L	H	L	L	L	L	-5
15	M	H	M	H	H	H	H	H	+6
16	M	L	M	L	H	H	L	L	-2
17	M	H	H	L	H	H	L	L	+1

Table 5.29: MAS student activity level for each tracked attribute. The number following each attribute, i.e. (2), indicates the number of clusters used to categorize the students. See Table 5.25 for the score values mapped to each activity level.

Clustering the students on their net scores, we obtain the following clusters:

MAS Cluster 0		
Student ID	**Net Score**	**Collaboration Phase Score**
1	+5	135
2	+4	100

3	+1	95
5	+1	100
6	+4	100
7	0	85
9	+3	100
10	+1	35
11	+1	80
15	+6	105
17	+1	85
Average	**2.45**	**92.73**
MAS Cluster 1		
Student ID	**Net Score**	**Collaboration Phase Score**
4	-3	83
8	-1	90
12	-1	95
13	-7	0
14	-5	70
16	-2	75
Average	**-3.17**	**68.83**

Table 5.30: Clusters and members for the MAS class, based on minimalist vs. overachiever net score. Collaboration phase score is also listed for comparison.

5.4.2.1 Observations

The two clusters formed by the Maximal Pairs algorithm yields a cluster of students whose net scores are positive (Cluster 0), and a cluster of students whose net scores are negative (Cluster 1). This distribution of students is relatively more balanced compared to the RAIK class, with 64.7% (11 out of 17) belonging to Cluster 0 and 35.3% belonging to the other cluster.

As with the RAIK class, we are interested in whether the students' net scores correlate to the grades they received during the collaboration phase of the wiki assignment. The grades for the students are thus listed alongside their net scores in Table 5.30. The correlation coefficient between the net score and collaboration phase score is +0.724, which is higher than that of the RAIK class. It should also be noted that the average grade of the overachievers' cluster is higher than that of the other cluster, at 92.73 vs. 68.83. This still holds when the relative outlier scores for each cluster (35 and

135 for Cluster 0 and 0 for Cluster 1) is excluded, averaging at 94.44 for Cluster 0 and 82.6 for Cluster 1.

Additionally, we are also interested in whether the quantity vs. quality tradeoff identified in previous sections is still present in these results. The following were observed in the categorizations for edits and comments specifically:

- Edits
    - Since the MAS class lacks distinct separation in its edit counts to determine clusters for them, we cannot confirm whether a tradeoff or positive correlation between count and length exists for this class.
    - Similarly, the relation between edit counts and quality is not distinct, due to the lack of categorization on edit counts.
    - 8 of the 17 students exhibit a positive correlation between edit length and edit quality, and only 1 student exhibits a tradeoff. This appears to be indicative of a generally positive correlation between edit length and quality.
- Comments
    - There is a tradeoff between count and length in 10 out of 17 students. The remaining 7 exhibit a positive correlation between them instead. As seen in the RAIK class, the students on the extreme ends of the spectrum appear to exhibit the positive correlation whereas the ones towards the center of it generally exhibit the tradeoff.
    - A tradeoff between content quality and count is observed in 7 out of 17 students, while a positive correlation was observed in 5 of the 17 students.

There does not appear to be a pattern between the negative/positive correlation status and the students' net scores. The relation between content quality relative to count is thus not distinct.

- For 11 of 17 students, a positive correlation is observed between quality and length. No tradeoffs were observed between these for any of the students for this class. This appears to be indicative of a generally positive correlation between comment length and quality.

5.4.2.2 Justifications

Why is there a positive correlation between Net Scores and Grades?

Since the net scores are based upon the same criteria for assignment grades, there is a positive correlation between the two. In particular, the grading by the instructor places particular emphasis on counts (i.e. comments and edits across 3 pages, for a minimum of three edits and three comments) and content quality.

Quantity vs. Quality

As previously highlighted, there is a tradeoff between the edit count and edit length for the RAIK class. However, this tradeoff is not observed via this this analysis for edits. Since the Maximal Pairs algorithm identified the optimal number of clusters for categorizing on edit count to be 1, we cannot observe the count vs. length tradeoff via the high/mid/low categories.

On the other hand, a tradeoff between comment count and comment length similar to that of the RAIK class is observed. The students' net scores also play a role in the MAS class: the extreme minimalists and overachievers exhibit a positive correlation between the two attributes, and the students in between exhibit the tradeoff.

As with the RAIK class, there is a lack of a tradeoff between count and content quality for both edits and comments in general. Instead, a positive correlation is observed for edit and comment lengths and content quality instead. This follows the intuition that using more words is associated with more-detailed explanations, which consequently leads to the idea that a longer contribution is of higher quality.

5.4.2.3 Implications

Usefulness of Two Clusters

As with the RAIK class, the optimal number of clusters for the MAS class is 2. However, the clusters appear to be of a different outcome (#3 of the ones listed in sub-chapter 5.4.1.3) – that of grouping overachievers, both slight and extreme, into one cluster and minimalists into the other. This class may similarly benefit from forcing the number of clusters to 3, as suggested for the other class.

Please refer back to the discussion in sub-chapter 5.4.1.3 for additional discussion on this topic.

Net score as grade indicator.

There is a stronger correlation between net score and grade for this class (+0.724 for MAS vs. +0.529 for RAIK). However, it may be possible to strengthen the correlation even further. Please see the corresponding heading in section 5.4.1.3 for additional information.

5.4.3 Summary

In this sub-chapter, we introduced the notion of minimalist vs. overachiever activity within the wiki as well as our approach towards measuring where students fall within the spectrum. It differs from the previous categorization analysis in sub-chapter 5.2 in that the previous analysis treats the tracked attributes as a *holistic* entity, whereas the minimalist vs. overachiever analysis examines the attributes *individually*. By examining the tracked attributes independently, we can rank each student's activity level on each of the individual attributes. These individual ranks are then combined into a whole to determine a student's overall placement on the minimalist/overachiever spectrum.

The minimalist vs. overachiever analysis in this sub-chapter raises multiple discussion points:

Count vs. Length, Count vs. Quality, and Length vs. Quality are three distinct comparisons with different relations.

As seen in the results of the individual classes, it is easy to conflate and intuitively assume particular correlations between the three attributes of count, length, and quality used for edits and comments. The general intuitive belief is that the three share the following relations:

- Count and length are believed to be negatively correlated and thus have a tradeoff relationship. The reasoning, as described in 5.2.3.2.2, is that due to limited time and/or cognitive resources, students will either make few large (length-wise) contributions or many shorter ones.
- Count and quality are believed to be negatively correlated and thus have a tradeoff relationship. The reasoning for this belief is akin to the one between count and length, where one needs to sacrifice "quality" if "quantity" is desired, and vice-versa.
- Length and quality are believed to be positively correlated and thus have a directly proportional relationship. The reasoning is that using more words allows for more-detailed explanations, resulting in a higher-quality contribution.

To summarize our actual findings on how the three inter-relate:

- ***A tradeoff is observed between edit counts and edit lengths.***

In the RAIK class, this tradeoff is relatively apparent with 9 out of 15 students exhibiting a high-low tradeoff between edit count and edit length. While two instances of low-low positive correlation were found, they belonged to students scoring as extreme minimalists – relatively rare occurrences in both class's data. This generally supports the intuition that there is a tradeoff between the two.

However, we are unable to confirm or debunk this finding in the MAS class, due to all students being placed into one cluster for edit counts. Why did this unfavorable result surface for this attribute? Unfortunately, a number of factors can play a role into this happening, including but not limited to: the distribution of edit counts being too tightly packed to form consistent clusters, the distribution of edit counts is sub-optimal for use with the X-means algorithm, etc.

There is a possibility that the relation between these two edit attributes may also be influenced by the student's net score, as observed in the upcoming discussion between comment count and length.

- ***Both positive and negative correlations are observed between comment counts and comment lengths. The correlation type is dependent on the absolute value of the student's net score.***

For both the RAIK and MAS class, it was observed that **students on the extreme ends of the spectrum have positive correlation between the two attributes**. That is, true to their namesakes, overachievers are willing to spend the cognitive effort to make multiple lengthy comments, whereas minimalists are not motivated to perform up to the threshold where the tradeoff is visible. **The remaining students in between the extremes exhibit the expected tradeoff.**

Why is this observed for comments but not edits? While both comments and edits are indisputably the activities with the highest cognitive cost relative to the other tracked activities, we believe that between the two of them, **edits generally have the higher cognitive cost to perform** since the topics for edits are limited to what is relevant and useful to the topic**.** In conversation threads, students can ask questions, teach, bounce ideas, and are generally not as limited. **Consequently, it is more difficult to exhibit**

overachieving behavior through edits. As for minimalist behavior, the RAIK class hints at the possibility of extreme minimalists exhibiting positive correlation between the two attributes. There are already two such instances in the data. However, we cannot verify this in the MAS class due to the rankings for edit counts. Data from more classes/students will be needed to verify whether this tradeoff between counts and lengths is observed beyond comments.

- ***A positive correlation is observed between length and quality for both comments and edits.***

In the MAS class, it was observed that there is a positive correlation between content length and content quality for both edits and comments. This supports the intuitive belief listed at the start of the discussion. But why does this not apply to the RAIK class? As discussed in 5.4.1.2, this expectation for a positive correlation may be a misconception in actuality: valuable contributions can be made in relatively few words if the writer is concise. Similarly, contribution length can be bloated while adding little value if the writer is long-winded. A possible reconciliation of the contradictory results is that **the positive correlation is dependent upon the writing skill of the student or the general writing skill of the class**. That is, a skilled writer is able to "say more with fewer words," whereas a lesser-skilled one may need more words to convey the same quality of ideas.

Correlation of Net Score to grade: using the minimalist-overachiever metric to aid grading

As observed in both classes, there is relatively strong positive correlation (greater than +0.500) between the calculated net score and the grade that the student earned for the wiki collaboration. With such strong correlations that can be improved via weighting net score calculation according to assignment criteria, it seems feasible to use this metric to aid instructors in grading student performance in the collaborative wiki assignment. The metric is particularly adept at identifying the quantifiable work put in by students, and will thus be well-suited to correlating with quantity-related aspects of assignments.

There are two particular challenges to using this metric as an aid to grading. First is that the minimalist-overachiever metric is a measure of where the student stands *relative to the other students in the class*. That is, particular net scores do not correspond to specific grades, nor are they comparable to the net scores of students in other classes. For example, it was observed that some students with relatively extreme minimalistic tendencies still attain scores such as 70 and 85 in spite of having a relatively low net score. While this may not make sense on an absolute scale (e.g., on the assumption that "a net score of -8 should always correspond to a grade of 0"), the received grade can make sense on a relative scale, since the 70-85 is towards the lower end of the grades distributed. Thus, *when using the Net Score to assist in determining a grade, the instructor still needs to determine the relative score range for the class.*

Finally, there is currently no completely-automated approach that can determine or rank the quality of a contribution's contents, so this ranking may need to be performed manually. The edit and comment quality are arguably the most vital parts of a student's collaborative contributions, and thus the net score metric is not a total substitute for the instructor's work in grading**.**

The optimal number of clusters: 3 vs. k

A discussion point arising in the RAIK and MAS class data is the optimal number of clusters for the minimalist vs. overachiever categorization. For these classes, k was determined to be 2, but the usefulness of two-cluster results was questioned in 5.4.1.3. To reiterate the points raised, there are three possible cluster outcomes for $k = 2$:

1. One cluster containing extreme overachievers, one cluster containing everyone else
2. One cluster containing extreme minimalists, one cluster containing everyone else
3. One cluster containing overachievers (both extreme and slight), one cluster containing minimalists (both extreme and slight)

For recommendation, all three outcomes when *k*=2 have their flaws. Clusters 1 and 2 are only able to identify and consequently generate recommendations that are only useful to one class of users, the extreme minimalists or extreme overachievers. In both those cases, the cluster containing everyone else spans too large a range of user behaviors such that one recommendation to the group would not meet the needs of all of the users within it. **By forcing the optimal number of clusters to 3, it is possible for the cluster the students into a group of extreme overachievers, a group of extreme minimalists, and a group of the remaining students in between.**

5.5 Editing and Commenting Page Sets

In this section, we wish to investigate the correlation (or lack thereof) between the pages that a student edits and the pages that the same student comments upon. Examining this aspect can improve wiki recommendations by gauging the breadth of collaborative activity that a particular student will participate in for a particular page. For example, it can help determine whether the student will *also* contribute towards a discussion on the page after the system successfully guides her to make an edit on it. While ratings and tags can also be used in this analysis, we limit the focus to comments and edits due to their relatively higher cognitive cost of contribution.We propose the following metric to measure the degree to which this behavior is present within students:

$$\%\ \boldsymbol{Overlap} = \frac{|\boldsymbol{Comment} \cap \boldsymbol{Edit}|}{|\boldsymbol{Comment} \cup \boldsymbol{Edit}|}$$

That is, we calculate the proportion between the set of pages that the student both edited *and* commented upon and the total set of pages edited or commented.

Intuitively speaking, if students have limited resources, they would aim to reduce the total cognitive cost of doing the assignment by minimizing the number of unique pages edited and discussed. That is, we expect the % Overlap score to be relatively high for all students.

5.5.1 RAIK Class

The edit/comment page sets and their overlaps for each student in the RAIK class are as follows:

Student ID	Pages Commented On	Pages Edited	# Overlap	% Overlap	# Edit First	# Cmt First
1	31, 36, 46, 47, 48	45, 47, 48	2	0.333	2	0
2	40, 41, 43, 48	43, 46, 48, 50	1	0.333	1	1

3	43	40, 43, 46	1	0.333	1	0
4	49, 50	49, 50, 51	2	0.667	2	0
5	31, 37, 38, 43	38, 46, 48, 49, 50	1	0.125	1	0
6	37, 42, 46, 48, 51	34, 40, 48, 51	2	0.286	2	0
7	45, 48	34, 45, 58	2	0.667	2	0
8	31, 38, 41, 43	36, 41, 43, 45, 51	2	0.286	1	1
9	40	34, 38, 42, 47	0	0.000	0	0
10	34, 41, 42, 48	34, 41, 42	3	0.750	3	0
11	31, 45	36, 46, 50	0	0.000	0	0
12	-	51	0	0.000	0	0
13	31, 34, 36, 50	31, 36, 50	3	0.750	3	0
14	31, 34, 51	36, 40, 41, 43	0	0.000	0	0
15	46	36, 46, 51	1	0.333	0	1
Average	**2.8 pages**	**3.4 pages**	**1.33**	**0.324**	**1.200**	**0.133**

Table 5.31: RAIK unique pages commented/edited and the overlap between the two

Contrary to expectations, there does not appear to be a consistent trend in % Overlap within the class as a whole. However, as exemplified in Sections 5.2 and 5.3, student behavior and motivations are not necessarily homogeneous. There is a possibility that the user's favored activity type (i.e. active or passive actions) may influence this overlap in pages edited and pages commented upon. If we separate the users based on the categorization from Chapter 5.3.1, then we have the following:

RAIK Cluster 0 (favors "active" actions)						
Student ID	**Pages Commented**	**Pages Edited**	**# Overlap**	**% Overlap**	**# Edit First**	**# Cmt First**

	On					
1	31, 36, 46, 47, 48	45, 47, 48	2	0.333	2	0
4	49, 50	49, 50, 51	2	0.667	2	0
5	31, 37, 38, 43	38, 46, 48, 49, 50	1	0.125	1	0
7	45, 48	34, 45, 58	2	0.667	2	0
8	31, 38, 41, 43	36, 41, 43, 45, 51	2	0.286	1	1
9	40	34, 38, 42, 47	0	0.000	0	0
10	34, 41, 42, 48	34, 41, 42	3	0.750	3	0
11	31, 45	36, 46, 50	0	0.000	0	0
15	46	36, 46, 51	1	0.333	0	1
Average	**2.778 pages**	**3.556 pages**	**1.444**	**0.351**	**1.222**	**0.222**

Table 5.32: RAIK unique pages commented/edited and the overlap between the two, for students placed in Cluster 0

RAIK Cluster 1 ("balanced" between active and passive actions)						
Student ID	**Pages Commented On**	**Pages Edited**	**# Overlap**	**%Overlap**	**# Edit First**	**# Cmt First**
6	37, 42, 46, 48, 51	34, 40, 48, 51	2	0.286	2	0
13	31, 34, 36, 50	31, 36, 50	3	0.750	3	0
14	31, 34, 51	36, 40, 41, 43	0	0.000	0	0
Average	**4 pages**	**3.667 pages**	**1.667**	**0.345**	**1.667**	**0**

Table 5.33: RAIK unique pages commented/edited and the overlap between the two, for students placed In Cluster 1

RAIK Cluster 2 (favors "passive" actions)						
Student ID	**Pages Commented On**	**Pages Edited**	**# Overlap**	**% Overlap**	**# Edit First**	**# Cmt First**
2	40, 41, 43, 48	42, 46, 48, 50	1	0.333	1	1
3	43	40, 43, 46	1	0.333	1	0
12	-	51	0	0.000	0	0
Average	**1.667 pages**	**2.667 pages**	**0.667**	**0.222**	**0.667**	**0.333**

Table 5.34: RAIK unique pages commented/edited and the overlap between the two for students placed in Cluster 2

5.5.1.1 Observations

When looking at the RAIK class as a whole, the overall % Overlap average is 0.324. However, dividing the students based on their activity type clusters results in average % Overlap of:

- 0.351 for Active
- 0.345 for Balanced
- 0.222 for Passive

Comparing averages between the Active and Balanced profile students, the average number of pages edited, number of pages commented, and the overlaps (both number and percentage) between the two clusters are relatively comparable. However, students with passive action-biased profiles have notably lower averages on all columns relative to the other two profiles.

Interestingly, for 85% (17 out of 20) of the overlaps, the user edits the wiki page prior to commenting upon it.

5.5.1.2 Justification:

Why would students with profiles favoring "active" actions have a greater % Overlap? Why would students favoring "passive" actions have a lower one?

Students may find it easier to make their edit and comment contributions to the same page since the "commitment cost" (e.g. time to read and familiarize page content, think of meaningful remarks/contributions, etc.) of doing so is lower than the commitment cost of commenting and editing two different pages. This "two-for-one" in providing edits and comments to the same page can drive a student's activity profile towards the "Active" end of the active-passive actions spectrum, particularly if the student repeats this across multiple pages.

Why the preference for edits over comments? Why the preference for edit-first?

The RAIK class students appear to strongly prefer editing a wiki page before participating in a threaded discussion on it, as evidenced by 85% of the overlaps being edit-first. Possible reasons for this behavior may include:

- The students prefer to focus on edits first, since they make up 60% of the collaboration phase grade.
- Immediately after reading the wiki page to become familiar with the topic, it may be "easier" to make an edit than it is to participate in (or begin) a discussion thread.

5.5.1.3 Implications:

When a student favors active actions, the student's collaboration initiation rate on a page (in terms of likelihood of commenting or editing) increases with the inclusion of a comment mechanism.

While we lack a point of reference or control value for collaboration without a mechanism for wiki page discussion, we believe that the feature encourages particular students to collaborate more on a single page. Specifically, we posit that students in this class that have an "active" or "balanced" action profile are more likely to participate in or begin a discussion on the page after making an edit on it than students preferring "passive" actions. This implication follows from the previous justification for why there is a greater % Overlap for such students.

Comparison to the MAS class results will be needed to further support this.

5.5.2 MAS Class

We similarly calculate the % Overlap value for the MAS class, and the results of the calculation are in Table 5.29:

Student ID	Pages Commented On	Pages Edited	# Overlap	% Overlap	# Edit First	# Cmt First
1	194, 197, 204, 208	194, 196, 204, 208	3	0.600	2	1
2	189, 201, 203, 205	112, 188, 205, 208	1	0.143	1	0
3	112, 114, 198, 204	112, 114, 198	3	0.750	1	2
4	176, 199, 203	199, 203	2	0.667	2	0
5	183, 188, 194, 200	183, 194, 200	3	0.750	0	3
6	179, 193, 202	189, 194, 199, 202	1	0.167	1	0
7	179, 189, 200, 202	179, 200, 202	3	0.750	3	0
8	173, 189	173, 197, 201	1	0.250	1	0
9	180, 191, 204	180, 191, 204	3	1.000	2	1
10	179, 189, 194	-	0	0.000	0	0
11	189, 195, 203	112, 204	0	0.000	0	0
12	179, 203, 205	179, 194, 204	1	0.200	1	0
13	-	-	0	0.000	0	0

14	112, 183, 196, 200, 205	196, 200	2	0.400	0	2
15	154, 176, 192, 193, 196, 199, 205	154, 176, 188, 192, 193, 196, 199, 205, 208	7	0.778	4	3
16	208	208	1	1.000	1	0
17	195	112, 114, 203	0	0.000	0	0
Average	**3.2 pages**	**2.9 pages**	**1.82**	**0.438**	**1.118**	**0.706**

Table 5.35: MAS unique pages commented/edited and the overlap between the two

As with the RAIK class, there appears to be no consistent trend in % Overlap when looking at the MAS class as a whole. We thus similarly separate the users based on their clusters from Chapter 5.3.2, resulting in the following tables:

MAS Cluster 0 (favors “passive” actions)						
Student ID	**Pages Commented On**	**Pages Edited**	**# Overlap**	**% Overlap**	**# Edit First**	**# Cmt First**
2	189, 201, 203, 205	112, 188, 205, 208	1	0.143	1	0
10	179, 189, 194	-	0	0.000	0	0
13	-	-	0	0.000	0	0
Average	**1.5 pages**	**1.333 pages**	**0.333**	**0.048**	**0.333**	**0**

Table 36: MAS unique pages commented/edited and the overlap between the two, for students placed in Cluster 0

MAS Cluster 1 (favors “active” actions)						
Student ID	**Pages Commented On**	**Pages Edited**	**# Overlap**	**% Overlap**	**# Edit First**	**# Cmt First**
1	194, 197, 204, 208	194, 196, 204, 208	3	0.600	2	1
3	112, 114, 198, 204	112, 114, 198	3	0.750	1	2
4	176, 199, 203	199, 203	2	0.667	2	0
5	183, 188, 194, 200	183, 194, 200	3	0.750	0	3
6	179, 193, 202	189, 194, 199, 202	1	0.167	1	0
7	179, 189,	179, 200,	3	0.750	3	0

	200, 202	202				
8	173, 189	173, 197, 201	1	0.250	1	0
9	180, 191, 204	180, 191, 204	3	1.000	2	1
11	189, 195, 203	112, 204	0	0.000	0	0
12	179, 203, 205	179, 194, 204	1	0.200	1	0
14	112, 183, 196, 200, 205	196, 200	2	0.400	0	2
15	154, 176, 192, 193, 196, 199, 205	154, 176, 188, 192, 193, 196, 199, 205, 208	7	0.778	4	3
16	208	208	1	1.000	1	0
17	195	112, 114, 203	0	0.000	0	0
Average	**3.357 pages**	**3.214 pages**	**2.143**	**0.522**	**1.286**	**0.857**

Table 5.37: MAS unique pages commented/edited and the overlap between the two, for students placed in Cluster 1

5.5.2.1 Observations:

When examining the MAS class as a whole, the overall % Overlap average is 0.438. However, dividing the students by activity type cluster results in average % Overlaps of:

- 0.522 for Active
- 0.048 for Passive

This difference is relatively significant. As can be seen in the above tables, students with a profile biased towards “active” activities also have a notably higher average in all other columns.

Unlike the RAIK class, the MAS class has a relatively greater balance between edit-first and comment-first overlaps: only 61% of the overlaps (19 out of 31) stem from students commenting after editing.

5.5.2.2 Justification:

Why would students with profiles favoring "active" actions have a greater % Overlap? Why would students favoring "passive" actions have a lower one?

The justification for this can be similarly explained with the one provided in sub-chapter 5.5.1.2. That is, students may find it easier to make their edit and comment contributions to the same page since the "commitment" or "cognitive" cost of doing so is lower than the cost of commenting and editing two different pages.

Why preference for edits over comments? (Why the preference for edit-first?)

As with the RAIK class, the MAS class students generally edit wiki pages before commenting on them when both actions are performed on the same page since the majority of overlaps are edit-first. However, the proportion of overlaps where the edit was performed before the discussion participation is only 61%, compared to the RAIK class's 85%. Since the grading for this class is similar to that of the other one, the justification for this preference in sub-chapter 5.5.1.2 may also apply here.

5.5.2.3 Implications:

When a student favors active actions, the student's collaboration initiation rate on a page (in terms of likelihood of commenting or editing) increases with the inclusion of a comment mechanism.

Similar to the RAIK class, the data suggests that a comment mechanism enables students preferring “active” actions to have increased participation on the wiki pages they contribute towards. However, it isn’t as clear whether the student comments because she has edited the page or vice-versa, due to the 61%/39% split between edit-first and comment-first overlaps. See sub-chapter 5.5.1.3 for additional discussion.

Additional verification against more classes will be needed to further verify this implication.

5.5.3 Summary

This section examined the possible correlation between the set of pages edited and the set of pages commented upon for users. The biggest key finding across the results is that ***students with “active” (or “balanced”) activity profiles tend to have a greater % Overlap than those with “passive” activity profiles.***

As seen in both classes, the students categorized as favoring “active” or “balanced” actions have a greater % Overlap between their pages edited and commented than students favoring “passive” actions. Sub-chapter 5.5.1.2 and 5.5.2.2 justify this as arising from the lower cognitive cost of editing and discussing on a single page (one topic) rather than editing and discussing two separate pages (and two separate topics). This option is particularly appealing as it minimizes the effort required to edit and/or participate in discussions for three distinct pages, as discussed in sub-chapter 5.4. This finding is important in user modeling and recommendation in that **the value of a successful recommendation to students favoring “active” actions is increased, since such students will likely contribute to both edits and discussion on the recommended page**.

An interesting difference observed between the two classes is that the MAS class has a greater number (12 vs. 3) and proportion (39% vs. 15%) of comment-first overlaps, compared to the RAIK class. Solely examining the grading criteria for the collaboration phase for the two classes, there doesn't seem to be an apparent difference between them that would suggest such an effect. However, there is a notable difference in the individual contribution phase: *students in the MAS class are required to start three discussion threads on their own pages*. **It is possible that these conversation "kick-starters" are responsible for the increased comment-first overlaps observed in the MAS class by reducing the cost of participating in a discussion. That is, potential contributors no longer have to take on the cost of deciding upon an appropriate topic for open-ended discussion and creating the thread, if they choose not to.** This is another example of instructor guidelines influencing student behavior for the assignment, in a more-indirect way.

5.6 Student Cliques

Students may have particular peers with which they prefer to work with, for reasons varying from personal affinity and familiarity over the course of their education, to reputation with regards to intelligence and work ethic. A "clique" is a group of such students, preferring to collaborate with others within the clique than those outside it. For a wiki setting in particular, students with a strong preference towards their cliques will prefer to contribute towards (e.g. editing and commenting upon) pages that the other clique members have written or contributed towards.

In this section, we wish to examine whether editing cliques exist and play a role in the collaboration initiation rate for the two classes. While the interpersonal relations between users are not known or feasible to deduce from the tracked attributes, we can look for the appearance of cliques regardless of their reasons for forming. We are particularly interested in the presence of “strong” cliques – cliques that occur frequently across multiple wiki pages. They can be leveraged in recommendation by: 1) notifying cliques when one of its members participates on a wiki page (e.g., reinforcing cliques), or 2) biasing recommendations towards students outside of the target user’s cliques (e.g., weakening cliques).

We determine editing cliques among the students by performing the following steps:

1. Determine the unique contributors for each wiki page for the class.
2. Let *i* = 2.
3. For each possible grouping of *i* students in the class:
    a. Count the number of wiki pages where that particular *i*-student grouping occurs.
    b. If the occurrence count is greater than 1, note the grouping as a possible clique.
4. Increment *i* and repeat Step 3 until there are no *i*-student groupings with more than one occurrence.

With the above process alone, strong editing cliques would be defined solely by a relatively large number of grouping occurrences. However, consider the following scenario: Student A has only edited three unique pages, and Student B has edited twelve unique pages. Let us suppose that the number of mutually-edited pages between them is three. Would the A-B grouping be considered a strong editing clique even though the number of occurrences arises from Student B editing a relatively large number of unique pages? Compare this to a scenario where Students A and C both edited three unique pages and the three pages edited are the same for both of them.

To account for this possibility of larger clique occurrences due to students editing a larger number of unique pages, we thus introduce the notion of “clique strength,” which adjusts a clique’s occurrence count relative to the average number of unique pages edited by the clique members.

$$Clique\ Strength = \frac{\#\ of\ clique\ occurrences}{Average\ \#\ of\ unique\ pages\ edited\ across\ clique\ members}$$

5.6.1 RAIK Class

As per the previously listed procedure, we first determine the unique contributors for each wiki page. Table 5.37 lists the IDs of the wiki pages along with the consenting contributors who have edited them.

Page ID	Contributor IDs
31	1, 5, 11, 13
34	6, 7, 9, 10, 13
36	5, 8, 11, 13, 14, 15
38	5, 8, 9
40	3, 6, 14, 15
41	2, 8, 10, 14
42	7, 9, 10
43	2, 3, 4, 8, 14
45	1, 7, 8, 16
46	2, 3, 5, 6, 11, 15
47	1, 3, 9
48	1, 2, 5, 6, 7, 10

49	4, 5, 11
50	2, 4, 5, 11, 12, 13
51	4, 6, 8, 12, 14, 15

Table 5.38: Distinct contributors for each wiki page in the RAIK class

Using the above, we count the number of occurrences of each clique. Tables 5.38 and 5.39 list the cliques with more than two occurrences and their clique strengths. Note that the tables are for cliques of size 2 and 3 – there are no cliques of larger sizes with more than one occurrence.

Clique Members	# Clique Occurrences	Avg # of Unique Pages Edited	Clique Strength
1, 5	2	5.5	0.364
1, 7	2	4	0.500
2, 3	2	4.5	0.444
2, 4	2	4.5	0.444
2, 5	3	6	0.500
2, 6	2	5	0.400
2, 8	2	5.5	0.364
2, 10	2	4.5	0.444
2, 11	2	5	0.400
2, 14	2	5	0.400
3, 6	2	4.5	0.444
3, 14	2	4.5	0.444
3, 15	2	4	0.500
4, 5	2	5.5	0.364
4, 8	2	5	0.400
4, 11	2	4.5	0.444
4, 12	2	3	0.667
4, 14	2	4.5	0.444
5, 6	2	6	0.333
5, 8	2	6.5	0.308
5, 11	5	6	0.833
5, 13	3	5.5	0.545
5, 15	2	5.5	0.364
6, 7	2	4.5	0.444
6, 10	2	4.5	0.444
6, 14	2	5	0.400
6, 15	3	4.5	0.667
7, 9	2	4	0.500
7, 10	3	4	0.750
8, 14	4	5.5	0.727
8, 15	2	5	0.400

9, 10	2	4	0.500
11, 13	3	4.5	0.667
11, 15	2	4.5	0.444
14, 15	3	4.5	0.667

Table 5.39: RAIK clique occurrences and clique strengths for groups of size 2 with 2+ occurrences

Clique Members	Number of Times Occurred	Avg # of Unique Pages Edited	Clique Strength
2, 5, 6	2	5.67	0.353
2, 5, 11	2	5.67	0.353
2, 8, 14	2	5.33	0.375
3, 6, 15	2	4.33	0.462
4, 5, 11	2	5.33	0.375
4, 8, 14	2	5	0.400
5, 11, 13	2	5.33	0.375
5, 11, 15	2	5.33	0.375
6, 7, 10	2	4.33	0.462
6, 14, 15	2	4.67	0.428
7, 9, 10	2	4	0.500
8, 14, 15	2	5	0.400

Table 5.40: RAIK clique occurrences and clique strengths for groups of size 3 with 2+ occurrences

5.6.1.1 Observations

The following was observed from the RAIK class clique analysis. The largest editing cliques found on more than one wiki page is of size 3, with the largest number of occurrences being exactly 2 for all such cliques. Surprisingly, there are a relatively large number of cliques occurring on at least two pages: 35 for cliques of size 2, and 12 for cliques of size 3. In spite of the numbers, few of these cliques are “strong” (i.e., have a clique strength larger than 0.500). For size-2 cliques, 8 out of the 35 are considered “strong” (0.229), and none of the size-3 cliques qualify with this criteria.

5.6.1.2 Justifications

Why are there a relatively large number of cliques with 2+ occurrences?

The relatively large number of cliques may arise from the following factors:

- The number of pages available for the class to edit is approximately equal to the number of consenting users.

- The minimum number of unique pages edited for the assignment is 4: 1 from a student's primary contribution + 3 from the collaboration phase requirements. This is 25% of the total pages for the class.

With relatively few total pages in the pool for the entire class, the probability of clique occurrences on at least two pages is relatively larger than if there were more pages in the pool.

Why are there few strong editing cliques relative to the many potential cliques?

It is possible that the current threshold for "strong" cliques may be set too high relative to the average number of unique pages edited for the class.

5.6.1.3 Implications

Clique identification may have limited usefulness within the constraints of the particular assignment specifications.

This may be particularly true for the RAIK class, due to: 1) the small number of required unique pages for contributions (4), and 2) the small number of total editable pages for the class (15). That is, although cliques can be identified as "strong" due to the relative nature of its strength calculation, they may not be significant due to the scale of the short-term assignment. For example, although a clique strength of 0.666 is relatively "strong," it can be achieved by two users having two mutually edited pages when each have edited only three unique pages. Compare this to two users with 200 mutually edited pages when each edited 300 unique pages throughout their membership on the wiki – the users in the 200/300 scenario is considered to be “more” of a clique than the ones in the 2/3 scenario.

The strength formula will need to be revised to account for absolute number of occurrences and the number of users/pages.

5.6.2 MAS Class

Repeating the procedure on the MAS class, we first determine the unique contributors for each wiki page. Table 5.40 lists the wiki pages that the MAS class’s consenting users contributed towards.

Page ID	Contributor IDs
112	2, 3, 11, 17
114	3, 17
154	9, 15
173	1, 8
174	1
176	15
179	5, 7, 12
180	9
183	5, 14
185	8
188	2, 15
189	6, 7
191	9
192	15
193	3, 15
194	1, 5, 6, 12
195	16
196	1, 14, 15
197	8
198	3, 6
200	5, 7, 14
201	4, 8
202	6, 7, 10
203	4, 11, 17
204	1, 9, 11, 12, 13
205	2, 15, 17
208	1, 2, 12, 15, 16

Table 5.41: Distinct contributors for each wiki page in the MAS class

We then repeat the occurrence counts and clique strength calculations with this data. Table 5.41 lists the cliques of size 2 with two or more occurrences. As can be inferred from the lack of additional tables, there were no cliques of sizes 3 or larger that meet the 2+ occurrences criteria.

Clique Members	# of Clique Occurrences	Avg # of Unique Pages Edited	Clique Strength $\left(\frac{\#Occurrences}{Avg.\#UniquePages}\right)$
1, 12	3	5	.600
1, 15	2	7.5	.267
2, 15	3	6.5	.462
2, 17	2	4	.500
3, 17	2	4	.500
5, 7	2	4	.500
5, 12	2	4	.500
5, 14	2	3.5	.571
6, 7	2	4.5	.444
11, 17	2	3.5	.571

Table 5.42: Potential cliques of size 2 for the MAS class

5.6.2.1 *Observations*

The following were observed in the clique analysis results for the MAS class. The largest clique size for the MAS class is smaller than that of the RAIK class at a size of 2. Ten different size-2 cliques occurred on at least two wiki pages. This is relatively interesting since there are considerably fewer cliques in spite of there being more consenting students in the MAS class data than the RAIK class data (albeit a lower percentage). Although there are relatively fewer such cliques identified, three of the ten are considered “strong” cliques (0.300).

5.6.2.2 *Justification*

Why are there a smaller number of identified cliques for the MAS class?

This may arise from the following factors of the MAS class. First, there is a larger number of students in the MAS class than the RAIK class (29 vs. 16).

Consequently, there are a larger number of pages available for students to contribute towards, since each wiki page in the pool is created by one student in the class. Since the minimum number of unique pages edited stays at approximately 4 (1 primary contribution page + 3 different pages in collaboration phase) regardless of class size, this results in the dilution of student participation across more pages. Combined with a lower percentage of the students in the MAS class filling out the consent form for data usage, this creates the appearance of the wiki pages having a sparse set of contributors, and consequently, smaller and fewer cliques.

Why are there few strong editing cliques relative to the number of potential cliques?

As mentioned in the RAIK class POJI, the current threshold for “strong” cliques may be set too high relative to the average number of unique pages edited.

5.6.2.3 Implications

A larger class size dilutes the efforts of the student body and limits the largest possible clique size that can form.

This is demonstrated by comparing the clique results of MAS and RAIK – although MAS has approximately the same number of consenting users, there are considerably more cliques (i.e. more unique sets of students with more 2+ occurrences) found in the RAIK class. Consent issues also contribute towards this. See sub-chapter for 5.6.2.2 for additional discussion.

Clique identification may have limited usefulness within the constraints of the particular assignment specifications.

As with the RAIK class, using cliques may have limited usefulness within the constraints of the particular assignment specifications for the MAS class, due to the small number of required unique pages for contributions (4). See sub-chapter 5.6.1.3 for additional discussion on this.

5.6.3 Summary

This section covered the concept and analysis of editing cliques among the students. To summarize, the benefits of investigating the identification of cliques include: 1) improving user modeling by determining the degree to which cliques influence a user's collaborative actions, and 2) leveraging clique information in recommendations to bias results for or against clique members, depending on instructor goals. There were two key findings in this analysis:

Possible clique sizes and frequency of clique occurrences are dependent on the number of students in the class, the number of pages in the wiki, and the minimum number of unique pages edited/commented required by the assignment.

Due to the size differences of the classes, the RAIK and MAS classes highlight the effects of varying the number of students in the class and the number of pages in the wiki. Sections 5.6.1.2 and 5.6.2.2 discuss these in greater detail. Table 5.42 summarizes the outcome of modifying each factor.

Factor	**When Increased…**	**When Decreased…**

# of students	More cliques and larger cliques due to more students being available.	Fewer cliques, smaller cliques due to fewer available students to form cliques with.
# of wiki pages	Fewer cliques, smaller cliques, fewer clique occurrences due to spreading students out across more pages.	More cliques, larger cliques, more clique occurrences due to students choosing from a smaller pool of pages.
Minimum # of unique pages to edit	More cliques, larger cliques, more clique occurrences due to students working on more pages from a same-sized pool of pages.	Fewer cliques, smaller cliques, fewer clique occurrences due to students possibly choosing fewer pages to work on.

Table 5.43: Effects of various class/instructor-controlled factors on cliques

While the last item in the table, minimum number of unique pages to edit, is not explicitly covered in individual class results, we deduce the effects with the assumption that students will strive to meet the minimum specified by the instructor. When the minimum is increased, students will edit a greater portion of the page pool for the wiki. Consequently, this can lead to more cliques being identified, particular cliques occurring more frequently, and/or larger cliques in general. Conversely, decreasing the minimum may result in reduced participation on the wiki, and consequently lead to fewer identified cliques, fewer clique occurrences, and smaller clique sizes.

It should be noted that it may be possible for the effects from altering multiple factors simultaneously to result in a net offset of zero. For example, the specifications for this particular assignment call for an increase in the number of wiki pages with an increase in the number of students, since each are responsible for being the primary contributor to their own unique page. We predict that clique identification between two classes with different user and page populations but comparable student-page ratios will be relatively similar.

Relevant factors in the identification of "strong" cliques: number of clique occurrences, relative number of unique pages edited among clique members, absolute number of pages edited among clique members, number of users and pages in wiki.

In the opening to this section (5.6), we defined a "strong" clique as a function of:

1) The number of unique pages that the editing clique occurred on ("clique occurrences"), and…
2) The average number of unique pages edited by the members of the editing clique (a "relative" number of unique pages edited).

The latter condition was introduced to separate users who simply have a large editing page set from those who deliberately seek out particular users to collaborate with. In sections 5.6.1.3 and 5.6.2.3, we questioned whether this purely relative strength calculation is appropriate for this short-term assignment.

There may be other factors that need to be considered for the strength calculation. First, the absolute number of pages edited may need to be considered, as the 2/3 vs. 200/300 example in section 5.6.1.3 highlighted. Second, the number of users and pages in the class/wiki may also be a factor in clique strength. Referring back to Table 5.42, there are scenarios where it would be harder to form cliques. It would then logically follow that cliques formed in such an environment may possibly be stronger than those formed in more-ideal environments.

5.7 Chapter Summary

In this chapter, we performed the post-hoc analyses of real usage data, collected from students in two separate classes carrying out a collaborative writing assignment within the Biofinity Intelligent Wiki. Table 5.43 lists the specific user actions that we focus our analyses upon, along with the specific attributes pertaining to each.

Action	Attribute(s)
Edit	Number/count, word length, quality
Comment/Discussion	Number/count, word length, quality
Rate	Number/count
Tag	Number/count

Table 5.44: User actions and associated attributes focused upon in post-hoc analyses

In our first analysis, we clustered the students in each class based on these "raw" attributes by using our X-means based "Maximal Pairs" clustering algorithm (detailed in 5.2.1). In the following analyses, we clustered the students upon two composite metrics: active vs. passive profile percentage (5.3) and minimalist vs. overachiever score (5.4). The former calculates the weighted proportion of edits and comments over the total actions performed by a user, whereas the latter calculates a user's overall activity level relative to peers in the class. Finally, we analyze the users with respect to overlaps in their editing and commenting page sets (5.5) and possible editing cliques between them (5.6).

Each analysis provided us with a series of insights regarding our metrics, user behavior, and instructor influence within the wiki. Table 5.44 lists the highlights of each sub-chapter, along with its analogous equivalent for virtual collaboration outside of the classroom (when needed).

Section	Classroom Finding	Virtual Collaboration Equivalent
5.2	Student behaviors appear to be motivated by factors such as the evaluation criteria of the wiki assignment, “quantity vs. quality”, and the relative “ease” of making a particular contribution vs. others. While the instructor can influence student behavior via assignment requirements and evaluation criteria, student behavior can also be influenced by the behavior of their peers.	User behaviors are motivated by factors such as community contribution expectations and guidelines, the “quantity vs. quality” tradeoff, and the relative “ease” of making a particular contribution. While moderators can influence user behavior with rules and other expectations, user behavior can also be influenced by the behavior of their peers.
5.2	Clusters found via [holistic attributes clustering] can be leveraged to guide recommendation and profiling of students.	Clusters found via holistic attributes clustering can be leveraged to guide recommendation and profiling of users.
5.3	A positive correlation exists between the number of total active actions and total passive actions. The concept of “making an edit/comment” needs to be separated from the concept of “making a high quality edit/comment.”	(Same as column 1)
5.3	Tradeoff (or lack thereof) between the total number of unweighted collaborative actions and % active actions. Although total active actions and total passive actions are positively correlated, active actions still cannot be performed as quickly as passive actions.	(Same as column 1)
5.4	Count vs. Length, Count vs. Quality, and Length vs. Quality are three distinct comparisons with different relations. • A tradeoff is observed between edit counts and edit lengths. • Both positive and negative correlations are observed between comment counts and comment lengths. The correlation type is dependent on the absolute value of the student’s net score. • A positive correlation is observed between length and quality for both comments and edits. • No correlation is observed between count and quality for both comments and edits.	(Same as column 1)

5.4	The optimal number of clusters: 3 vs. k. For recommendation, all three outcomes when k=2 have their flaws. By forcing the optimal number of clusters to 3, it is possible for the cluster the students into a group of extreme overachievers, a group of extreme minimalists, and a group of the remaining students in between.	The optimal number of clusters: 3 vs. k. For recommendation, all three outcomes when k=2 have their flaws. By forcing the optimal number of clusters to 3, it is possible for the cluster the users into a group of extreme overachievers, a group of extreme minimalists, and a group of the remaining students in between.
5.5	Students with "active" (or "balanced") activity profiles tend to have a greater % Overlap than those with "passive" activity profiles. The value of a successful recommendation to students favoring "active" actions is increased, since such students will likely contribute to both edits and discussion on the recommended page.	Users with "active" (or "balanced") activity profiles tend to have a greater % Overlap than those with "passive" activity profiles. The value of a successful recommendation to users favoring "active" actions is greater than those favoring "passive" actions, since such users will likely contribute to both edits and discussion on recommended pages.
5.5	It is possible that these conversation "kick-starters" are responsible for the increased comment-first overlaps observed in the MAS class by reducing the cost of participating in a discussion. That is, potential contributors no longer have to take on the cost of deciding upon an appropriate topic for open-ended discussion and creating the thread, if they choose not to.	Conversation "kick-starters" can increase comment-first overlaps by reducing the cognitive cost of participating in a discussion. The cost of starting a discussion (i.e., deciding upon an appropriate topic for open-ended discussion) is now optional for potential contributors.
5.6	Possible clique sizes and frequency of clique occurrences are dependent on the number of students in the class, the number of pages in the wiki, and the minimum number of unique pages edited/commented required by the assignment.	Possible clique sizes and frequency of clique occurrences are dependent on the number of users on the wiki, the number of pages in the wiki, and the minimum number of unique pages edited/commented requested by the wiki moderator.
5.6	Relevant factors in the identification of "strong" cliques include: number of clique occurrences, relative number of unique pages edited among clique members, absolute number of pages edited among clique members, number of users and pages in wiki.	(Same as column 1)

Table 5.45: Primary findings within the various Chapter 5 analyses

Chapter 6: Conclusion

In this chapter, we summarize the purpose of the thesis in Chapter 6.1, briefly cover the accomplishments and results found in Chapter 6.2, and close out the thesis by highlighting the potential avenues for future work in Chapter 6.3.

6.1 Summary of Purpose

In this thesis, we highlighted and motivated the importance of virtual collaboration and the benefits of improving it (Chapter 1). We focused our scope on collaboration within wikis in particular, since it is a medium that is still seeing widespread use to this day. While multiple tools have been developed to support and improve collaboration in such a setting (e.g., Annoki platform by Tansey and Stroulia and the Socs application by Atzenbeck and Hicks), there were relatively few *intelligent* ones beyond the SuggestBot of Cosley et al. (Chapter 2). We thus identified recommendation-based intelligent support for improving wiki collaboration as the target niche for our work.

The primary goal of the thesis is to investigate and provide the foundation for future implementation of recommendation systems for a wiki environment. The contributions our work provides include: 1) wiki-based user and data models and a proposed recommendation algorithm leveraging those models, 2) a design for and implementation of an intelligent wiki that allows for the addition of social and intelligent features, and 3) insights to user behavior, moderator influence, and model efficacy that can be leveraged in future work.

6.2 Summary of Achievements and Results

There are two major achievements from the work performed for the thesis. One is the design and implementation of our own intelligent wiki. Rather than modifying an existing wiki, we implemented the majority of it from the ground up using common technologies including Java, Javascript, HTML, Glassfish, and the Google Web Toolkit. In addition to providing us with a great deal of flexibility for future development and freeing us from the confines of more-restrictive software licenses, this enables us to design the wiki to include social/Web 2.0 features (e.g., page tagging, page ratings, intra-wiki and social network sharing, thread-based discussions) and intelligent features (e.g., page and user modeling, user tracking via an agent-based framework, recommendation framework) from the start. For additional implementation details, please refer back to Chapter 4.

Another major achievement is the insights discovered in the analysis of the wiki usage data which span topics such as user behavior and contribution strategies, the degree that moderator influence and guidelines affect users, relations between tracked attributes, and the applicability and relevance of our new metrics. Our investigation was carried out in the following manner: 1) deploy the Biofinity Intelligent Wiki to multiple settings, 2) collect usage data as the various users in each setting carry out their tasks on it, and 3) perform post-hoc analysis on the data once a setting-specific milestone is reached or the purposes for using the wikis were fulfilled. Ultimately, the two data sets used in this thesis originated from a classroom setting where the wiki is used for a collaborative writing assignment. Our analyses of the usage data provided us with the following generalized findings:

- User behaviors are motivated by factors such as community contribution expectations and guidelines, the “quantity vs. quality” tradeoff, and the relative “ease” of making a particular contribution. While moderators can influence user behavior with rules and other expectations, user behavior can also be influenced by the behavior of their peers.
- Clusters found via holistic attributes clustering can be leveraged to guide recommendation and profiling of users.
- A positive correlation exists between the number of total active actions and total passive actions. The concept of “making an edit/comment” needs to be separated from the concept of “making a high quality edit/comment.”
- Tradeoff (or lack thereof) between the total number of unweighted collaborative actions and % active actions. Although total active actions and total passive actions are positively correlated, active actions still cannot be performed as quickly as passive actions.
- Count vs. Length, Count vs. Quality, and Length vs. Quality are three distinct comparisons with different relations.
    - A tradeoff is observed between edit counts and edit lengths.
    - Both positive and negative correlations are observed between comment counts and comment lengths. The correlation type is dependent on the absolute value of the student’s net score.
    - A positive correlation is observed between length and quality for both comments and edits.

  - No correlation is observed between count and quality for both comments and edits.

- The optimal number of clusters: 3 vs. k. For recommendation, all three outcomes when k=2 have their flaws. By forcing the optimal number of clusters to 3, it is possible for the cluster the users into a group of extreme overachievers, a group of extreme minimalists, and a group of the remaining students in between.
- Users with "active" (or "balanced") activity profiles tend to have a greater % Overlap than those with "passive" activity profiles. The value of a successful recommendation to users favoring "active" actions is greater than those favoring "passive" actions, since such users will likely contribute to both edits and discussion on recommended pages.
- Conversation "kick-starters" can increase comment-first overlaps by reducing the cognitive cost of participating in a discussion. The cost of starting a discussion (i.e., deciding upon an appropriate topic for open-ended discussion) is now optional for potential contributors.
- Possible clique sizes and frequency of clique occurrences are dependent on the number of users on the wiki, the number of pages in the wiki, and the minimum number of unique pages edited/commented requested by the wiki moderator.
- Relevant factors in the identification of "strong" cliques include: number of clique occurrences, relative number of unique pages edited among clique members, absolute number of pages edited among clique members, number of users and pages in wiki.

With the work presented in this thesis – a proposed user and data model, a proposed recommendation algorithm, and insights to user behavior and moderator influence over it – we established a strong foundation for improving virtual wiki collaboration via intelligent support. However, this is only a "first step": there is still much work to be done such as revising the model and recommendation algorithm per our findings and deploying and evaluating the wiki under different scenarios. Chapter 6.3 delves more into the future work to be accomplished to extend this work.

6.3 Future Work

In this subsection, we outline multiple avenues of possible future work for the ideas developed or introduced in this thesis.

6.3.1 Testing Against Additional Data Sets

As previously multiple times throughout the results section, our analyses were performed on a relatively small number of data sets and users. Specifically, we only had permission to examine 32 students across two classes. It is not surprising that we are in need of additional data sets to further support (or possibly refute) our findings from the post-hoc data analysis.

Data sets that would be interesting to examine include: 1) classroom wiki assignments structured and graded differently from the assignments used for the MAS and RAIK classes, and 2) wiki usage for non-classroom purposes, i.e., a user-base motivated by the group's goals rather than a grade. With the first data set, we can confirm the generality of the RAIK and MAS class implications while still constrained to a classroom setting. With the second, we can confirm the generality of the findings to the broader scope of all wiki collaboration.

6.3.2 Live Analyses and Usage of Model

The analyses carried out in the results chapter were all carried out post-hoc, and while this approach facilitates the discovery of insights and patterns, our findings do not take into account the various obstacles and challenges that arise during the live use of the user and data models. For example, the “Cold Start” problem may be of particular concern in a classroom setting due to:

- The relatively short duration of the assignment. A class’s duration is roughly 18 weeks long, and the time allotted for the collaborative writing assignment will only be a fraction of that time.
- Due to the above, students typically will not have an extensive activity history from which to generate recommendations from, or even identify the student’s strategy, behavior archetype, etc.

This particular obstacle may be mitigated to some degree by leveraging student data from previous classes/assignments. However, this will only serve to reduce the amount of information that is needed from the student and will not eliminate the need for it entirely. Additional investigation and preparation will be needed before live usage of the model (and eventually, the recommendation algorithm) can be possible.

6.3.3 Additional Intelligent Features

In addition to page recommendations, there are multiple other intelligent features that may be valuable to add to the Biofinity Intelligent Wiki. Possibilities include:

- User recommendation – suggesting specific users for the target user to collaborate with. May be based on mutual interests, users’ expertise, “friend-of-a-friend” connections, etc.

- Activity level-based alerting – notifying moderators when users with extreme overachieving/minimalistic scores and users with relatively extreme “active” or “passive” action profiles are identified. Users can also be notified when their own performance appears to be lacking or exceeding the norm.

References

Adomavicius G., Tuzhilin A. (2005). Toward the Next Generation of Recommender Systems: a Survey of the State-of-the-Art and Possible Extensions. In IEEE Transactions on Knowledge and Data Engineering, Volume 17, Number 6, pp. 734-749.

Anderson, T., Gunawardena, C., Lowe, C. (1997). Analysis Of A Global Online Debate And The Development Of An Interaction Analysis Model For Examining Social Construction Of Knowledge In Computer Conferencing. Journal of Educational Computing Research, 397-431.

Atzenbeck C., Hicks D. (2008). Socs: increasing social and group awareness for Wikis by example of Wikipedia. In Proceedings of the 4th International Symposium on Wikis, pp. 9:1-9:10.

Burke R., Felfernig A., Goker M. (2011). Recommender Systems: An Overview. In AI Magazine, Volume 32, Number 3, pp. 13-18.

Chen J., Geyer W., Dugan C., Muller M., Guy I. (2009). "Make New Friends, but Keep the Old" – Recommending People on Social Networking Sites. In Proceedings of the 27th International Conference on Human Factors in Computing Systems, pp. 201-210.

Chen J.R., Wolfe S.R., Wragg S.D. (2000). A Distributed Multi-Agent System for Collaborative Information Management and Sharing. In Proceedings of the 9th International Conference on Information and Knowledge Management, pp.382-388.

Chen J., Geyer W., Dugan C., Muller M., Guy I. (2009)."Make New Friends, but Keep the Old" - Recommending People on Social Networking Sites. Proceedings of the 27th International Conference on Human Factors in Computing Systems, (pp. 201-210).

Cosley, D., Frankowski, D., Terveen, L., Riedl, J. (2007). SuggestBot: Using Intelligent Task Routing to Help People Find Work in Wikipedia. Proceedings of the 12th International Conference on Intelligent User Interfaces, (pp. 32-41).

Demartini G. (2007). Finding Experts using Wikipedia. In Proceedings of the Workshop on Finding Experts on the Web with Semantics at ISWC/ASWC2007, pp. 33-41.

Durao F., Dolog P. (2009). Analysis of Tag-Based Recommendation Performance for a Semantic Wiki. In 4thWorkshop on Semantic Wikis in conjunction with the 6th Annual European Semantic Web Conference.

Forte A., Bruckman A. (2007).Constructing text: Wiki as a toolkit for (collaborative?) learning. In Proceedings of the 2007 International Symposium on Wikis, pp. 31-41.

Gokhale A. (1995). Collaborative Learning Enhances Critical Thinking. In Journal of Technology Education, Volume 7, No. 1, pp. 22-30.

Griffiths N. (2006). Enhancing Peer-to-Peer Collaboration using Trust. In Expert Systems with Applications, Volume 31, Issue 4, pp. 849-858.

Guy I., Ronen I., Wilcox E. (2009). Do You Know? Recommending People to Invite into Your Social Network. In Proceedings of the 13th International Conference on Intelligent User Interfaces, pp. 77-86.

Han E.H.S, Karypis G. (2005). Feature-Based Recommendation System. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pp. 446-452.

Hara N., Solomon P., Sonnenwald D.H., Kim S-L. (2003). An Emerging View of Scientific Collaboration: Scientists' Perspectives on Collaboration and Factors that Impact Collaboration. In Journal of the American Society for Information Science and Technology, Volume 54, Issue 10, pp. 952-965.

Herring S. (2004).Slouching Toward the Ordinary: Current Trends in Computer-Mediated Communication. In New Media & Society, Volume 6, pp. 26-36.

Hu M., Lim E-P., Sun A., Lauw H., Vuong B-Q. (2007). Measuring article quality in Wikipedia: models and evaluation. In Proceedings of the 16th ACM Conference on Information and Knowledge Management, pp. 243-252.

Jermann P., Soller A., Muehlenbrock M. (2001). From Mirroring to Guiding: a Review of State of the Art Technology for Supporting Collaborative Learning. In Proceedings of the First European Conference on Computer-Supported Collaborative Learning, pp. 324-331.

Karoly L., Panis C. (2005). The 21st Century at Work: Forces Shaping the Future Workforce and Workplace in the United States.

Katz J.S., Martin B.R. (1997). What is Research Collaboration? In Research Policy, Volume 26, Issue 1, pp. 1-18.

Kester L., van Rosmalen P., Sloep P., Brouns F., Koné M., Koper R. (2007). Matchmaking in Learning Networks: Bringing Learners Together for Knowledge Sharing. In Interactive Learning Environments, pp. 117-126.

Khosravifar B., Bentahar J., Moazin A., Thiran P. (2010). On the Reputation of Agent-Based Web Services. In Proceedings of the 24th AAAI Conference on Artificial Intelligence, pp. 1352-1357.

Konstan J., Miller B., Maltz D., Herlocker J., Gordon L., Riedl J. (1997). GroupLens: Applying Collaborative Filtering to Usenet News. Communications of the ACM, 40(3), pp. 77-87.

Kraut R., Egido C., Galegher J. (1988). Patterns of Contact and Communication in Scientific Research Collaboration. In Proceedings of the 1988 ACM Conference on Computer-Supported Cooperative Work, pp. 1-12.

Li L., Wang Y. (2010). Subjective Trust Inference in Composite Services. In Proceedings of the 24th AAAI Conference on Artificial Intelligence, pp. 1377-1384.

Lih A. (2004). Wikipedia as participatory journalism: Reliable sources? Metrics for evaluating collaborative media as a news resource. In Proceedings of the 5th International Symposium on Online Journalism.

Limerick D., Cunningham B. (1993). Collaborative Individualism and the End of the Corporate Citizen.

Linden G., Smith B., York J. (2003). Amazon.com Recommendations: Item-Item Collaborative Filtering. IEEE Internet Computing, 7(1), pp. 76-80.

Liu G., Wang Y., Orgun M. (2010). Optimal Social Trust Path in Complex Social Networks. In Proceedings of the 24th AAAI Conference on Artificial Intelligence, pp. 1391-1398.

Martin F., Donaldson J., Ashenfelter A., Torrens M., Hangarter R. (2011). The Big Promise of Recommender Systems. In AI Magazine, Volume 32, Number 3, pp. 19-27.

Mattessich P.W., Monsey B. (2001). Collaboration: What Makes it Work. A Review of Research Literature on Factors Influencing Successful Collaboration, pp. 15-17.

McDonald D.W. (2003). Recommending collaboration with Social Networks: A Comparative Evaluation. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 593-600.

McDonald D.W., Ackerman M. (2000). Expertise Recommender: a Flexible Recommender System and Architecture. In Proceedings of the 2000 ACM Conference on Computer-Supported Cooperative Work, pp. 231-240.

Meier A., Spada H., Rummel N. (2007). A rating scheme for assessing the quality of computer-supported collaboration processes. In International Journal of Computer-Supported Collaborative Learning, Volume 2, Number 1, pp. 63.68.

Melville P., Mooney R., Nagarajon R. (2002). Content-boosted Collaborative Filtering for Improved Recommendations. In Proceedings of the 18th National Conference on Artificial Intelligence, pp. 187-192.

O'Reilly T. (2004). What is Web 2.0: Design patterns and business models for the next generation of software.

Ocker R., Yaverbaum G. (1999). Asynchronous Computer-mediated Communication versus Face-to-face Collaboration: Results on Student Learning, Quality and Satisfaction. In Group Decision and Negotiation, Volume 8, Issue 5, pp.427-440.

Paek T., Gamon M., Counts S., Chickering D.M., Dhesi A. (2010). Predicting the Importance of Newsfeed Posts and Social Network Friends. In Proceedings of the 24th AAAI Conference on Artificial Intelligence, pp. 1419-1424.

Papagelis M., Plexousakis D. (2005). Qualitative Analysis of User-based and Item-based Prediction Algorithms for Recommendation Agents. In Engineering Applications of Artificial Intelligence, Volume 18, Number 7, pp. 781-789.

Preece J., Shneiderman B. (2009). The Reader-to-Leader Framework: Motivating Technology-Mediated Social Participation. In AIS Transactions on Human-Computer Interaction, Volume 1, Issue 1, pp. 13-32.

Sabater J., Sierra C. (2001). Regret: a Reputation Model for Gregarious Societies. In Proceedings of the Fourth Workshop on Deception, Fraud, and Trust in Agent Societies, pp. 61-69.

Sabater J., Sierra C. (2002). Reputation and Social Network Analysis in Multi-agent Systems. In Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1, pp. 475-482.

Schaffert S. (2006). IkeWiki: A Semantic Wiki for Collaborative Knowledge Management. In 15th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, pp. 388-393.

Scott S., Henninger S., Jameson M.L., Moriyama E., Soh L-K., Harris S., Ocampo F., Simpson R. (2008). NSF Proposal for the Semantic Cyberinfrastructure for Information Discovery.

Shepitsen A., Gemmell J., Mobasher B., Burke R. (2008). Personalized Recommendation in Social Tagging Systems using Hierarchical Clustering. Proceedings of the 2008 ACM Conference on Recommender Systems, pp. 259-266.

Song M., Lee W., Kim J. (2010). Extraction and Visualization of Implicit Social Relations on Social Networking Sites. In Proceedings of the 24th AAAI Conference on Artificial Intelligence, pp. 1425-1430.

Sriurai W., Meesad P., Haruechaiyasak C. (2009). Recommending Related Articles in Wikipedia via a Topic-Based Model. In Proceedings of the 9th International Conference on Innovative Internet Community Systems, pp. 194-203.

Suh B., Chi E., Kittur A., Pendleton B. (2008). Lifting the Veil: Improving Accountability and Social Transparency in Wikipedia with WikiDashboard. In Proceeding of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems, pp. 1037-1040.

Tansey, B., Stroulia E. (2010). Annoki: a MediaWiki-based collaboration platform. In Proceedings of the 1st Workshop on Web 2.0 for Software Engineering , pp. 31-36.

Vassileva J., McCalla G., Greer J. (2003). Multi-Agent Multi-User Modeling in I-Help. User Modeling and User-Adapted Interaction, 13(1-2), pp. 179-210.

Vivacqua A., Moreno M., de Souza J. (2003). Profiling and Matchmaking Strategies in Support of Opportunistic Collaboration. In Lecture Notes in Computer Science, Volume 2888, pp. 162-177.

Vivacqua A., Moreno M., de Souza J. (2006). Using Agents to Detect Opportunities for Collaboration. In Lecture Notes in Computer Science, Volume 3865, pp. 244-253.

Wang Y., Vassileva J. (). Trust and Reputation Model in Peer-to-Peer Networks. In Proceedings of the 3rd International Conference on Peer-to-Peer Computing, pp. 150-157.

Stahl G., Koschmann T., Suthers D. (2006). Computer-supported collaborative learning: An historical perspective. In Cambridge handbook of the learning sciences, pp. 409-426.

Warkentin M., Sayeed L., Hightower R. (1997). Virtual Teams versus Face-to-Face Teams: An Exploratory Study of a Web-based Conference System. In Decision Sciences, Volume 28, Number 4, pp. 975-996.

Wessner M., Pfister H-R. (2001). Group Formation in Computer-Supported Collaborative Learning. In Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work, pp. 24-31.

Appendix A: Supplementary Algorithms

A.1 Weighted Harmonic Mean

A weighted harmonic mean of attribute *k* is defined as follows:

$$WHM(k) = \frac{\sum_{j=1}^{n} \alpha_j}{\sum_{j=1}^{n} \frac{\alpha_j}{k_j}} = \frac{1}{\sum_{j=1}^{n} \frac{\alpha_j}{k_j}} \text{ where:}$$

- $k_j$is the value of attribute *k* during period *j*
- *n* is the number of time periods, where each period is characterized as being within a certain temporal distance from the current date
- α_j is the weight for period *j*, and $\sum_{j=1}^{n} \alpha_j = 1$
- *j* is the index of the time period evaluated. The corresponding weights and period lengths for each *j* is:

j	**Period**	α_j
1	within six months from the current timestamp	0.40
2	between six to 12 months from the current timestamp	0.30
3	between 12 to 18 months from the current timestamp	0.20
4	older than 18 months	0.10

Table A.1: Time periods and corresponding weights used in Weighted Harmonic Mean calculations

A.2 User Interests

A user *u*’s interests are computed in the following manner:

$$Interests(u) = \sum_{p \in P_u} \sum_{action \in Actions_{u,p}} \beta_{action} * Tags(p)$$

Where:

- P_u is the set of all pages that the user *u* has interacted with
- $Actions_{u,p}$ is the set of actions that the user *u* has performed on *p*
- β_{action} is the weight of the action performed

- *Tags(p)* returns the binary tag vector of page *p*

The weights for each of the user actions are preliminarily assigned in Table A.2 (below), based upon the relative amount of interest needed to perform them.

Action	β_{action}
Create	2.25
Edit	2.25
Discuss	2.0
Recommend	1.5
Rate Up	1
View	1

Table A.2: Weighted counts used for calculating user interests.

Appendix B: RAIK Clusters – Four Clusters

To obtain the four-cluster results, the end condition for our algorithm is changed to the moment the target number of clusters is reached. That is, the pseudocode is now:

```
• While number of unique clusters remaining in C is greater than k:
    o For each cluster A in C:
        ▪ For each cluster B in C:
            • If A != B AND HasStrongMaximalPair( A, B ) == TRUE
                o Merge( A, B )
                o If |C| == k, return  // new step bolded for emphasis
```

When multiple cluster merges are available for the iteration, the one with the larger occurrence count is prioritized.

The new memberships for the four-cluster results of the RAIK class are as follows:

Cluster ID	Members (Student ID)
Cluster 0	0, 2, 3, 8, 11, 14
Cluster 1	1
Cluster 2	4, 6, 10, 13
Cluster 3	5, 7, 9, 12

Table B.1: Cluster membership for RAIK four-cluster results

The following tables list the attributes for each cluster member:

Cluster 0							
Student	RaikNumEdits	RaikEditLength	NumCmts	CmtLength	NumTags	NumRates	Dist. From Centroid
0	4	408.250	6	108.333	0	11	211.377
2	3	285.000	1	113.000	0	18	93.736
3	6	256.667	2	99.500	3	7	62.113
8	5	48.200	1	52.000	0	5	153.172
11	3	57.667	0	0.000	0	16	160.856

14	3	141.000	1	81.000	0	0	59.489
Avg	4.000	199.464	1.833	75.639	0.500	9.500	123.457
Std Dev	1.265	141.835	2.137	43.227	1.225	6.834	-

Table B,2: Attribute details for RAIK cluster 0

Cluster 1							
Student	**RaikNumEdits**	**RaikEditLength**	**NumCmts**	**CmtLength**	**NumTags**	**NumRates**	**Dist. From Centroid**
1	5	106.400	4	33.000	34	16	0
Avg	-	-	-	-	-	-	-
Std Dev	-	-	-	-	-	-	-

Table B.3: Attribute details for RAIK cluster 1

Cluster 2							
Student	**RaikNumEdits**	**RaikEditLength**	**NumCmts**	**CmtLength**	**NumTags**	**NumRates**	**Dist. From Centroid**
4	7	83.714	4	69.250	10	6	13.625
6	9	54.222	2	73.000	0	17	21.351
10	9	40.667	2	54.500	5	12	35.041
13	5	115.400	3	69.667	3	15	42.187
Avg	7.500	73.501	2.750	66.604	4.500	12.500	28.051
Std Dev	1.915	33.214	0.957	8.242	4.203	4.796	-

Table B.4: Attribute details for RAIK cluster 2

Cluster 3							
Student	**RaikNumEdits**	**RaikEditLength**	**NumCmts**	**CmtLength**	**NumTags**	**NumRates**	**Dist. From Centroid**
5	8	66.25	6	86.5	16	29	13.687
7	11	76.73	5	79.6	8	17	19.305
9	11	41.73	4	70	11	8	27.571
12	13	57.62	4	114	25	14	28.711
Avg	10.750	60.580	4.750	87.525	15.000	17.000	22.318
Std Dev	2.062	14.800	0.957	18.902	7.439	8.832	-

Table B.5: Attribute details for RAIK cluster 3

Since the cluster assignments are fairly similar to that of the three-cluster results, we will highlight the differences with the previous results.

B.1 Observations

Regarding impressions from the results, the singleton cluster from the three-cluster results persists in these four-cluster results. For additional discussion pertaining to this cluster, refer back to the previous section. Particularly notable is the cluster “split” – that is, the New Clusters 2 and 3 appear to be a “split” of one of the clusters from the “old” three-cluster result. In other words, they are the two clusters that comprise Old Cluster 2.

With regards to new correlations between attributes and cluster assignments, while the intervals for each attribute of New Clusters 2 and 3 overlap with each other within one standard deviation from their means, the new clusters seem to split some attributes into “high” and “low” subgroups.

- *Number of edits* - Cluster 2 comprises the lower half of the range covered by the two clusters whereas Cluster 3 comprises the upper half.
- *Number of comments* - Cluster 2 comprises the lower half of the range covered by the two clusters whereas Cluster 3 comprises the upper half.
- *Comment length* - Cluster 2 comprises the lower half of the range covered by the two clusters whereas Cluster 3 comprises the upper half.
- *Number of tags* - Cluster 2 comprises the lower half of the range covered by the two clusters whereas Cluster 3 comprises the upper half.

B.2 New Justifications

We felt the following items needed particular justification.

- *Why the appearance of a "split"?*

Our procedure can be seen as a hierarchical clusterer, with the metric for merging clusters being the presence of strong maximal pairs. That is, at some particular point in the original algorithm execution, the number of clusters goes from four to three. Thus, to go from three clusters to four, execution is "halted" when four clusters are remaining, and one of the "old" clusters would appear to be "split" into two separate ones.

- *Why was the "Old" Cluster 2, i.e., Cluster 2 from the three-cluster results, "split" instead of the others?*

Old Cluster 1 is a singleton cluster and thus is not eligible to be "split". Also, Old Cluster 0 had a "stronger" maximal pair, i.e. larger co-occurrence count with their maximal pairs, between the clusters to be merged than the one between new clusters 2 and 3. It thus had higher "priority" for merging over these two.

- *Little benefit from four-cluster?*

For this particular data, the four-cluster results do not seem to significantly differ from the three-cluster results. This appearance of having little benefit may arise from:

1. The discrepancy between the target number of clusters and the actual number of clusters in the results being only one short, rather than being much smaller.

2. The relatively small standard deviations and relative lack of outliers in the “old” cluster before it was “split”.
3. The “new” clusters being a relatively “even” split that do not highlight/isolate outliers.