A Support Platform for Collaborative Workflow based on Seamless Repository

Yilang Wu

A DISSERTATION
SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN COMPUTER SCIENCE AND ENGINEERING

Graduate Department of Computer and Information Systems
The University of Aizu
2016

The Thesis titled

A Support Platform for Collaborative Workflow based on Seamless Repository

by

Yilang Wu

is reviewed and approved by:

Chief Referee

Professor

Zixue Cheng

Professor

Subhash Bhalla

Professor

Neil Y.(Yuwen) Yen

Professor

Peng Li

The University of Aizu

2016

This dissertation is dedicated to my wife and our new born baby.

Acknowledgments

I first would like to express my great appreciation to Dr. Zixue Cheng for his excellent advice and diligent efforts to guide my research through my Ph.D. career. Dr. Zixue Cheng is always available with his valuable suggestions and guidance throughout my Ph.D. study. I will never forget the revisions he made on each of my technical papers. Without his support from various aspects, I could not make it today.

I must unequivocally express my profound gratitude to all the dissertation committee members, Prof. Subhash Bhalla, Prof. Neil Y.(Yuwen) Yen, and Prof. Peng Li for many suggestive comments to improve the quality of this dissertation.

I also would like to thank all the Performance Evaluation Lab members, Dr. Junbo Wang, Dr. Lei Jing, Dr. Yinghui Zhou, Dr. Wangmin Chu, Mr. Sato Kouichi, Mr. Abe Daisuke, Mr. Sato Takeyuki, Mr. Yamada, who were always there in both my study and life. It was a great time to be with you these years. Various support from the University of Aizu was of great importance for me to pursue my Ph.D. studies. I especially would like to thank Mr. Tatsuki Kawaguchi, Miss. Hoshi, as well as the other staffs for their always warm help during these years.

Finally, my deep gratitude goes to my family. I thank my parents, as well as my wife and our new born baby for their love and support.

Abstract

Teamwork participants are always in demand of better working and collaboration support. The mobile cloud network has empowered the teamwork to be more pervasive by easing the spatial and temporal constrains. SNS and big data enable a verity of support tools or systems for supporting teamwork. However, the collaborative workflow gaps still commonly exist in working and collaboration among co-workers who are familiar with different support systems.

This dissertation aims at bridging the collaborative workflow gaps by providing support of seamless integration and knowledge correlating. The different but persistent personal preferences of using the support systems require new support of seamless integration. And also the different background knowledge and the different purposes of utilizing knowledge require the support of knowledge correlating. Therefore, a novel support platform is developed through a seamless integration of multiple support systems and knowledge correlating to solve the following issues.

- Seamless Integration using a three-layered architecture
  - Support of Sharing to reduce the gap of information.
  - Support of Interconnection to reduce the gap of communication.
  - Support of Visualization to reduce the gap of representation.
- Knowledge Correlating using the terms-frequency and chained links-ratio (TFCLR) measure
  - Support of Correlation measure to reduce gap of knowledge.

Comparing with other support systems, the seamless integration in this platform has better functionality in sharing, interconnection, and visualization. And comparing with other collaboration measure, the TFCLR measure achieves better performance in information coverage and usability, and also has tolerable performance in speed and feasibility.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Teamwork participants are always in demand of better working and collaboration support. The mobile cloud network has empowered the teamwork to be more pervasive by easing the spatial and temporal constrains. SNS and big data enable a verity of support tools or systems for supporting teamwork. However, there are gaps in working and collaboration among users/developers who are familiar with different support systems.

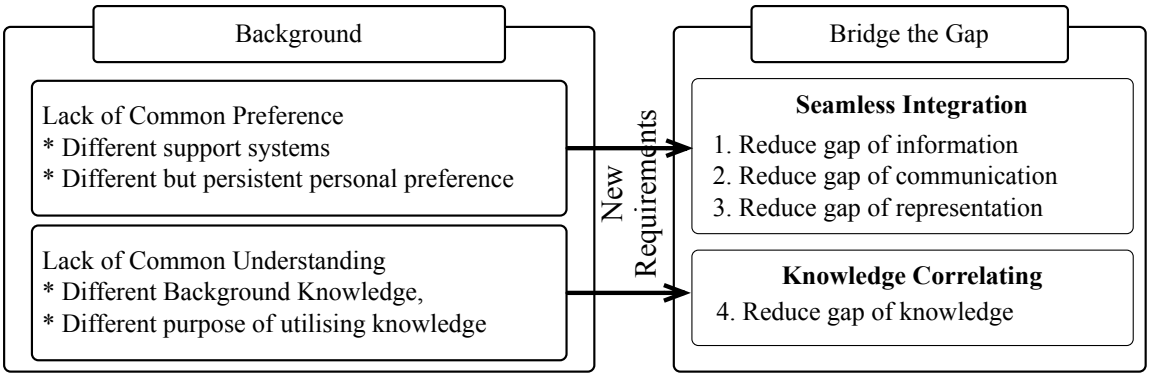


Figure 1.1. New Requirements to Bridge Gaps in Collaborative Workflow

Figure 1.1 illustrates the present situations in nowadays group working and collaboration: the lack of common preferences in development support systems and also the lack of common understanding. That results in the rising gaps in collaborative workflow, for example the gaps of information, communication, representation and knowledge. There are increasing new requirements to

bridge the gaps in collaborative workflow.

Figure 1.2 illustrates the difficulties for the teamwork participants to work and collaborate with each other. Each existing support system even though is not perfect, still has a number of group users for specific usage. Teamwork members have to switch between different support systems in order to utlize the data or functionality inside the systems or cooperate with other members. However, is not reasonable for each member to get familiar with all support systems. The different but persistent personal preference of using the support systems requires the new support of seamless integration. People from different domain fields are usually called up together to implement complex projects. It is usually hard to find or deliver the feasible experience and achievements for solving similar problems in another domain fields. Thus, the difference in background knowledge and the different purposes of utilizing knowledge require the support of knowledge correlating.

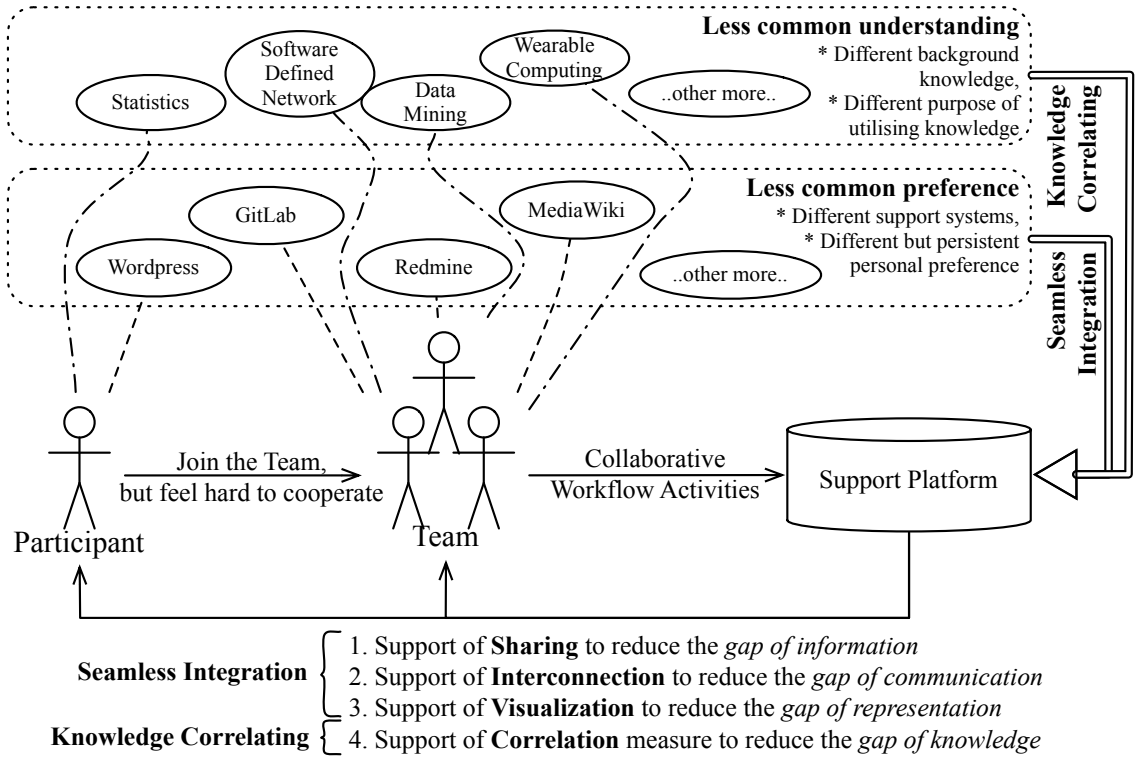


Figure 1.2. Support Platform for Collaborative Workflow based on Seamless Repository

1.2 Support Platform

This dissertation aims at bridging the gaps in collaborative workflow by providing support of seamless integration and knowledge correlating. A framework illustration of the support platform is given in Figure 1.2. The first 3 layers from the bottom show the components for seamless integration, and the 4th layer on the top shows the components for knowledge correlating. The services of the support platform are enumerated in Table 1.1 in regard to the framework of the support platform (see Figure 1.2).

The seamless integration is the first major issue studied in this dissertation; it is expected to systematically integrate the data and functionality of different support systems. There are three sub issues regarding the three layers for seamless integration, including the support of sharing, interconnection, and visualization.

- Issue 1.1: It is about how to reduce the gap of information; and an improved support of sharing is proposed. Regarding the critical collaborative activities, the Layer-1 provides a service of online portfolio to make it easier for users to know other participants more quickly; and also a service of tasks template to process the tasks more feasibly.

- Issue 1.2: It is about how to reduce the gap of communication; and an improved support of interconnection is proposed. The Layer-2 collects the raw data of collaborative workflow activities to provide 3 services. One is a service of web portal to help users to harness the skills; the second one is a service of communication channels to deliver needs more meaningfully; and the third one is a service of notification to warm up the workplace.

- Issue 1.3: It is about how to reduce the gap of representation; and an improved support of visualization is proposed. The Layer-3 collects the context data and relation data to provide a service of heat-map to give user a broad awareness of the whole team, and also a service of animation to give user a narrow awareness of individual members in teamwork.

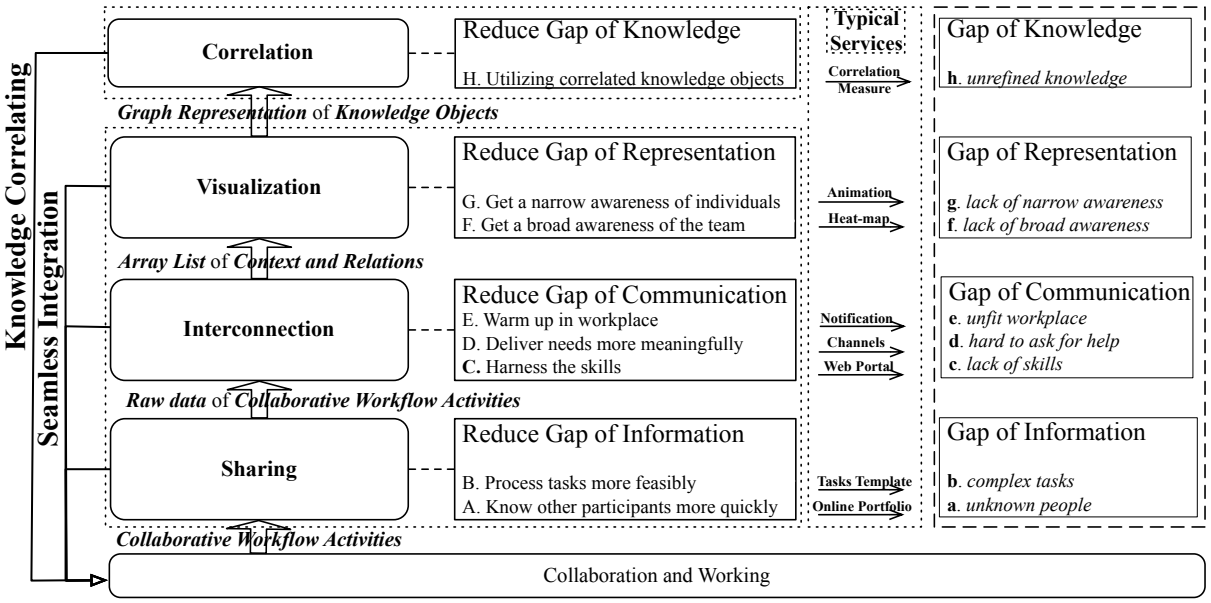


Figure 1.3. Framework of Support Platform

Table 1.1. Services of Support Platform

Gaps	Collaborative Workflow Barriers			Services for Support			Support
Gap of Information	a	*Heterogeneity*	*unknown people*	A	Online Portfolio	*know other participants more quickly*	Sharing
	b	*Workflow complexity*	*complex tasks*	B	Task Templates	*process tasks more feasibility*	
Gap of Communication	c	*Skills Gap*	*lack of skills*	C	Web Portal	*harness the skills*	Inter-connection
	d	*Poor Communication*	*hard to ask for help*	D	Communication Channels	*deliver needs more meaningfully*	
	e	*Workplace Conflicts*	*unfit workplace*	E	Notifications	*warm up in workplace*	
Gap of Representation	f	*Teamwork disruption*	*lack of broad awareness*	F	Heat-map	*get a broad awareness of the team*	Visualization
	g	*Less attractive experience*	*lack of narrow awareness*	G	Animation	*get a narrow awareness of individuals*	
Gap of Knowledge	h	*Knowledge avalanche*	*unrefined knowledge*	H	Correlation Measure	*utilizing correlated knowledge objects*	Correlation

The knowledge correlating is the second major issue studied in this dissertation; it is expected

to recommend users the correlated knowledge for supporting their real-time development, including the useful experience and achievements that are shared in the integrated support systems.

- Issue 2: It is about how to bridge the gap of knowledge; and an improved support of correlation measure is proposed. Although the support systems are integrated through the seamless integration, the knowledge object still staying unrefined, the Layer-4 (on the top of Figure fig:scenario-framework) provides a service of correlation measure to utilize the correlated knowledge objects.

Figure 1.2 shows the framework of the support platform. The seamless integration serves seamless Sharing, Interconnection and Visualization, by providing the solutions from "A" to "G" to reduce the barriers from "a" to "g".

- Layer-1: Support of Sharing
  - A. Portfolio: Know other participants more quickly by learning from their shared portfolio in repository, and vice versa.
  - B. Templates: Process tasks more feasibly by breaking them down onto feasible templates, and match with execution plan in repository.
- Layer-2: Support of Interconnection
  - C. Web Portal: Harness the skills by cooperating with others in executing the predefined tasks in repository, and exploring over Web portal.
  - D. Channels: Deliver needs more meaningfully by contacting in preferable channels, and with reference to predefined pieces in repository.
  - E. Notification: Warm up in workplace by setting up repository preference, and notifying activities in repository to achieve interests.
- Layer-3: Support of Visualization
  - F. Heat-map: Get a broad sense of team by reviewing the historical teamwork performance visualized by repository.
  - G. Animation: Get a narrow sense of individual participant by reviewing

- Layer-4: Support of Correlation
  - Correlation Measure: Utilize correlated knowledge objects in repository

In the framework, layer-1 (support of sharing) collects the collaborative workflow activities and provides the raw data of the collaborative workflow activities to the Interconnection layer; layer-2 (support of interconnection) receives that, and provides the array list of context and relations (extracted from the raw data) to the layer-3 (support of visualization); and layer-3 receives that and provide the graph representation of the modeled knowledge objects to the Correlation layer.

1.3 Originality and Contributions

To tackle the challenging issues aforementioned, this dissertation is firstly compared with other related work to specify the research domain and effort; then four scenarios are introduced respectively to clarify the four issues (see Section 1.2). The originality and contributions of this dissertation concentrate on seamless integration and knowledge correlation, and they are summarized in the following items. The structure of this dissertation is further illustrated in Figure 1.4.

- The related work is described in Chapter 2. In this chapter compares the seamless integrated system with other support systems to show its advantages in terms of improved support of sharing, interconnection and visualization. It also compares the proposed correlation measure with other methods to show its advantages especially the high information coverage.

- Chapter 3 introduces the seamless integration. Three scenarios are given to show the necessity of improving the supports of sharing, interconnection, and visualization to bridge the gaps of information, communication, and representation respectively. And a three-layered framework is specified to show the approaches for seamless integration of multiple support systems in to a seamless repository. And the information security and system scalability design is also considered to guarantee the quality of service after seamless integration.

- Chapter 4 introduces the knowledge correlating. A graph model is presented to organize the knowledge objects from different support systems. And a correlation measure based on terms-frequency and chained links-ratio (TFCLR) is proposed to quantify the conceptual and

relational correlation among the knowledge objects. Then the system specification and technical implementation are also given to measure the correlations among knowledge objects by analyzing the raw data of collaborative workflow activities.

- Chapter 5 demonstrates the support platform for collaborative workflow through successful case studies. To bridge the gaps of information, communication, representation, and knowledge, the implemented services include portfolio, workflow templates, web portal, communication channels, notification, teamwork involvement heat-map and animation, and correlation measure to knowledge objects.

- Chapter 6 summarizes the achievements of the support platform, and discusses the future improvement of the seamless integration and knowledge correlating.

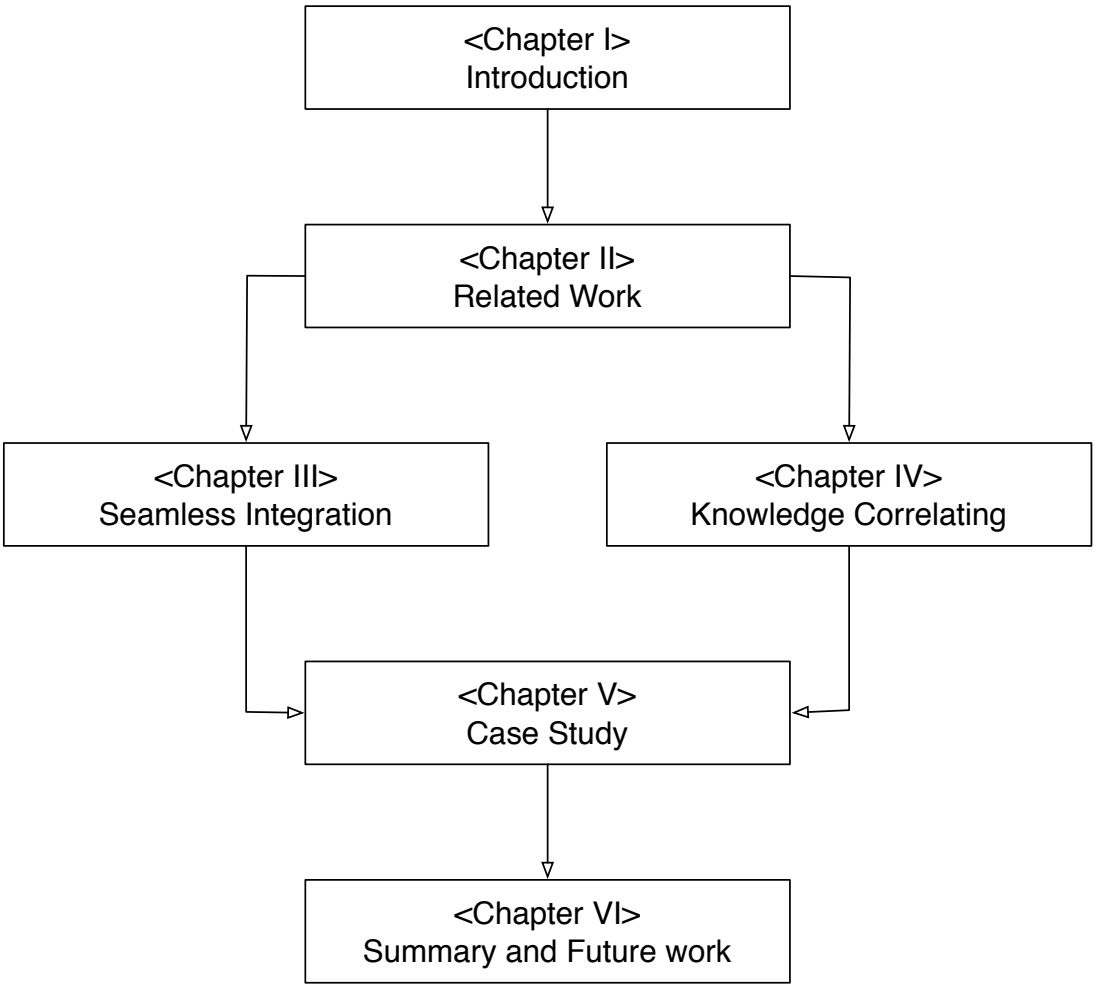


Figure 1.4. Structure and Contents of this Dissertation

Chapter 2

Related Work

Many business activities and scientific disciplines are now data and information driven, new business value and scientific knowledge are often gained by the co-workers putting together data resource and operation flows. In recent years, a number of collaboration and working oriented researches have been studied for the purpose of supporting teamwork in commercial and scientific domains. Some researches focus on the descriptive language for optimal workflow design to facilitate workflow activities, such as Kepler [1] which implements workflow description model to record the information about a workflow run. Some researches focus on the coordination of human activities to facilitate collaboration, such as PetriNet based scheduling [2] to optimize the time-based coordination.

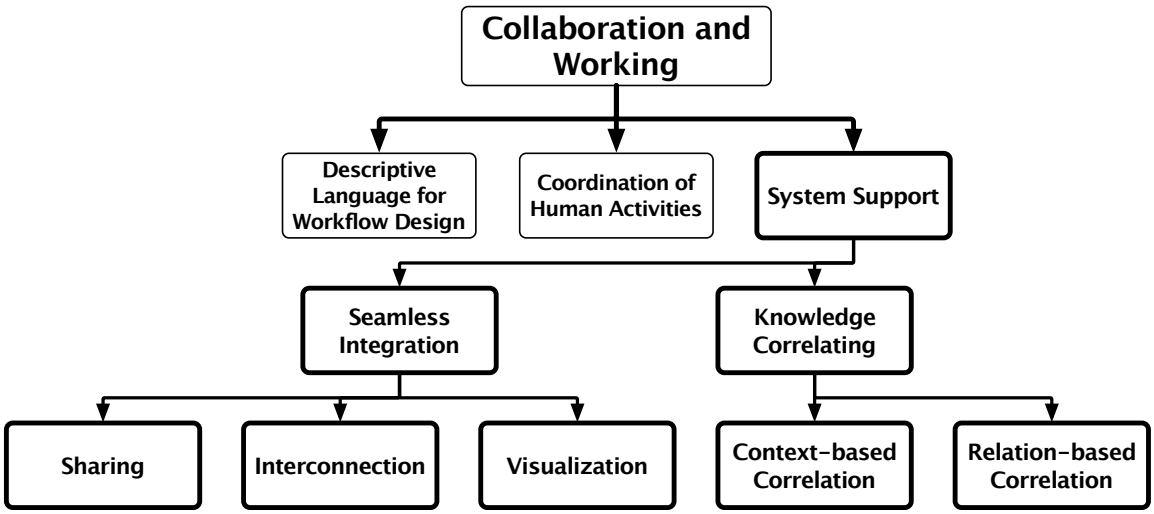


Figure 2.1. Relations with Other Related Work

As the business and scientific projects become more collaborative and more detailedly segmented, there is a compelling need of system support. Therefore, this dissertation focus on the research of system support (see Figure 2.1) for collaborative workflow by proposing a new support platform. There are two major challenging issues for this study, one is the seamless integration of different support systems, and the other is the knowledge correlating to utilize the correlated knowledge objects.

2.1 Related Work of Seamless Integration

Teamwork [3] has permeated all aspects of working around the world, which can greatly enhance motivation and efficiency. With the advent of mobile cloud technology, such as cloud storage [4], indexing [5], syndication [6], and the Internet of Things [7], teamwork enters the new paradigm of pervasiveness [8], becoming adaptable anytime and anywhere even in dynamic virtual organizations. The mobile cloud-based collaborative workflow (MCCW) empowers teamwork by using cloud resources to manage communication, working documents, and flows. It is precedent in today's IT industry, which has shown the effectiveness of mobile cloud-based collaborative workflow in the following aspects:

- The mobile cloud helps to break down and formalize a collection of parallel and sequential tasks that rely on communication and coordination. Through a mobile cloud environment, team members can collaborate with each other by chaining each others' workflow. The collaboration work can be planned as optimized workflow, and then be divided and assigned to different participants. The workflow can further be broken into small pieces of feasible work, being stored in the cloud and made available for access through wired or mobile devices.

- Compared to existing workflows restricted by location and time, teamwork participants in MCCW have less temporal or spatial constraints to complete minor tasks and coordinate with each other to finish the whole workflow.

The mobile cloud has brought great positive impact to teamwork collaboration more pervasively with less temporal and spatial limitations. However, current existing platforms are still insufficient against barriers such as heterogeneity, [9] poor communication, [10] skills gap, [11] workplace con-

flicts, [12] workflow complexity, [13] team disruption, [14] and a less immersive experience [15]. Such barriers enhance the importance of teamwork transparency [16], which means more user-friendliness in pervasive computing, and herein more seamless collaboration without barriers. In this dissertation, we define those barriers are types of collaborative workflow gaps. Such gaps bring new challenges to share the collaborative workflow more comprehensively, to interconnect the collaborative workflow more smoothly, and to represent the collaborative workflow more perceptibly.

2.1.1 Collaborative Workflow Barriers

Teamwork suffers from collaborative workflow barriers in various aspects. There are also new challenges when teamwork enters the pervasive paradigm due to collaborative workflow barriers. Figure 2.2 shows the commonly existing collaborative workflow barriers in team work. For example, when Alice and Bob, who are interest in different domains and having different preferences in support systems, start collaborating and working in a team, they will face many barriers as outlined in the following items.

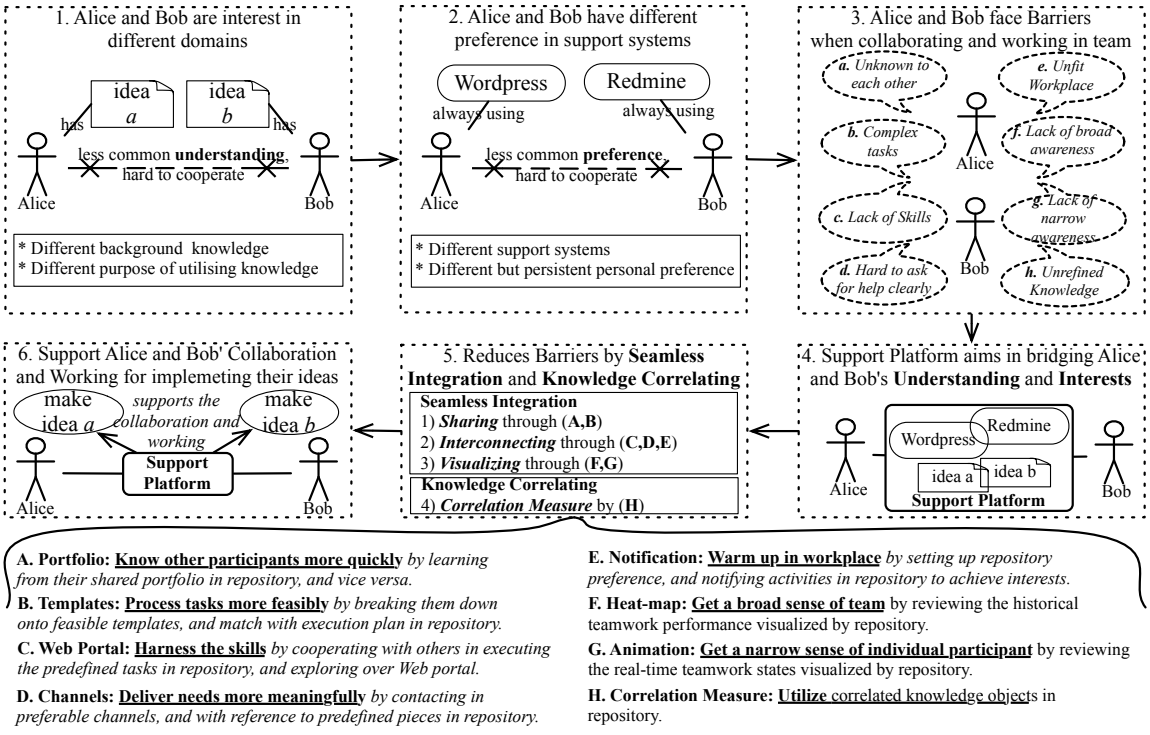


Figure 2.2. Commonly Existing Collaborative Workflow Barriers and Possible Services to Support

- An earlier paper [9] states that key barriers to collaboration in global software development

include geographic, temporal, cultural and linguistic distances. Such *heterogeneity* barriers make it difficult for teamwork participants to smoothly work together or be integrated synthetically.

- Communication [10] in terms of virtual or face-to-face communication is one of the most important teamwork activties. *Communication* barrier may cause misunderstanding or isolation, that will finally result in poor efficiency and effectiveness in teamwork.

- Skills gap [11] in teamwork, such as a lack of technical competency required for specific teamwork tasks, can result in an unbalanced workload between skilled and unskilled participants, which will decrease the feasibility of teamwork.

- Workplace conflicts, which are related with environmental factors, relationship conflicts, and personal dissatisfaction, take time for participants to eliminate or adjust to. Too many conflicts may result in very poor teamwork performance [12].

- Workflows [17] consist of an orchestrated and repeatable pattern of business activity, and high complexity in the workflow process may result in bad understandability and more errors, defects, and exceptions [13]. This excessive complexity should be avoided [18].

- Cohesion is the important resistance of the group to disruptive forces [14] in order to keep organizational alignment. Teamwork will lose its sustainable energy due to the barrier of *teamwork disruption*.

- Pervasive teamwork without an immersive collaboration environment [19] is still insufficient to establish the sense of working together as one integrity. The barrier of a less immersive experience may result in the loss of situation awareness among participants.

In the era of the mobile cloud, collaborative workflow barriers are in higher, more challenging levels. Therefore, to break these barriers, it is necessary to improve the collaborative workflow support with better sharing, interconnection, and visualization to bridge the gap of information, communication, and representation respectively.

2.1.2 Collaborative Workflow Repositories

The collaborative workflow [20] aims at serving synergistic efficiency gains to its constituents by removing collaboration barriers. There are various open and closed-source tools available including workflow software and social software. These tools and data resources integrated together are considered to be teamwork repositories. The repositories are classified into into two types: one targets workflow activities, which hold the activity body information of dedicated workflow such as issue tracing, revision control, content management; the other targets collaborative activities, which hold communication contents about teamwork, such as e-mail and notification.

Table 2.1. List and Comparison of Existing Workflow Repositories

Repositories for Workflow Activities			Comparison		
Workflow		**R**epositories	(1)	(2)	(3)
Issue Tracking		Manages and maintains lists of issues needed by teamwork, such as Redmine [21, 22], JIRA [23]	◯	△	×
Revision Control	Distributed	Git [24] for Source Code Revision, such as GitLab [25], GitHub [26], BitBucket [27]	◯	△	×
	Centered	Collaborative documentation [28] such like Wiki [29], including MediaWiki [30], PukiWiki [31]	◯	△	×
Content Management		Supports the collection, managing, and publishing of information in any form or medium such as Joomla, Drupal, and Wordpress [32]	◯	△	×
Data Archiving		Acquisition, preparation, preservation, and dissemination of data, such as vsFTPD and SAMBA [33]	△	△	×
System Virtualisation		The act of creating a virtual system, including hardware, OS, storage, network resources, such as VirtualBox, VMWare	◯	△	△
Description		◯: high level; △: middle level; ×: poor level. (1): Sharing; (2) Interconnection; (3): Visualization;			

Table 2.2. List and Comparison of Existing Collaborative Repositories

Repositories for Supporting Collaborative Activities		Comparison		
Collaboration	**R**epositories	(1)	(2)	(3)
Proximity Tracking	To track onsite meetup between teamwork participants, using GPS, BLE Proximity sensor technology [34], and so on	△	△	△
Mail Transferring	Mailing and notification via mail transferring such as using Postfix	△	△	△
Instant Messaging	Messaging and notification through team communication tool, such as Slack [35] which supports both IRC [36] and XMPP [37]	△	◯	△
Description	◯: high level; △: middle level; ×: poor level. (1): Sharing; (2) Interconnection; (3): Visualization;			

Based on the classification, the commonly used repositories for collaborative workflow are listed in Table 2.1 and 2.2. There are still systematic and social barriers among participants, due to duplicated functionality, poor systematic integrity, and unfriendly user experience. The support of sharing in workflow support systems is better than that of collaboration software, but its support of interconnection is relatively poor than that of the later one. And both have insufficiency support to visualize the collaborative workflow.

More specifically, data archiving tools like FTP and SAMBA [33] provide basic functions for index archives in a tree-view structure and share it on the server side; they do not record information regarding who updates the archives. Issue tracking tools like Redmine [21, 22], or JIRA [23] have very detailed functionality of workflow tracking and working roles definition, but are unable to publish workflow archives elegantly. Content management tools like Wordpress [32] are good at categorizing archives and have a number of aesthetic themes, but can not simply manage workflow archives such as source code. Distributed revision control tools like GitLab [25], GitHub [26], BitBucket [27], and centralized revision control tools like MediaWiki [30], PukiWiki [31] and Wikipedia have limited functions of categorizing contents, and only provide information as a guidance of how to reuse the workflow archives without the help of an IDE to process archive entries such as source code. System virtualization tools such as VirtualBox can initialize a virtual OS (operating system) which encapsulate the IDE to process workflow archives, but they lack server side

support to share virtual OS partitions among participants. Besides workflow software, collaborative software based on mail transferring tools (such as Postfix) and notification tools (such as Slack) have to be integrated with other workflow software for a more efficient communication. As for the proximity tracking tools for onsite collaboration, GPS-based mobile apps have poor performance for indoor behavior tracking, while BLE-based proximity [34] is more suitable in comparison [38].

In summation, the general disadvantages of the existing collaboration workflow repositories are:

- Each of the existing sub repositories only targets a specific purpose or workflow, and lacks the functionalities to support collaborative teamwork.
- Users cannot simply search data from one single entry, thus creating a lack of a broad overview to collaborative teamwork.

Therefore, there is high demand in advancing the collaborative workflow enlivenment with better support of sharing, interconnection, and visualization of the collaborative workflow data and functionality.

Table 2.3. Comparison of Seamless Integrated Platform with Other Support Systems

Support for Collaborative Workflow Activities			Comparison		
Category	**S**upported Activities	**R**epository Information	(1)	(2)	(3)
This Platform	Collaboration Activities, and Workflow Activities	Seamless integration of collaborative workflow repositories by serving the *Portfolio*, *Templates*, *Channels*, *Web Portal*, *Heat-map*, and *Animation*.	◯	◯	◯
Other Systems	Collaborative Activities (Mainly)	Repositories that mainly support collaborative activities, such as *proximity tracking*, *mail transferring*, *instant messaging*.	△	△	△
	Workflow Activities (Mainly)	Repositories that mainly support workflow activities, such as *issue tracking*, *revision control*, *content development*, and so on.	◯	△	△
Description		◯: high level; △: middle level; ×: poor level. (1): Sharing; (2) Interconnection; (3): Visualization.			

The proposed platform in this dissertation is seamlessly integrated, better in sharing, interconnection, and visualization than other systems. Furthermore, this platform has the support of knowledge correlating, which is lack in other related systems. As shown in Table 2.3, so far there are two

kinds of support systems. One provides collaboration support only based on collaborative repositories; another provides workflow support only based on workflow repositories.

Regarding Table 2.3, the proposed support platform has the advantages in sharing, interconnection, and visualization, because it keeps not only the functionality of other repositories, but also provides additional services, such as 1) portfolio, workflow templates to share collaborative workflow more comprehensively, 2) channels, notification and Web portal to interconnect the collaborative more smoothly, 3) heat-map and also animation to represent the collaborative workflow more perceptibly. Other support systems though are advanced in supporting specific activities (for example issue tracking), still have insufficient synthetic performance in sharing, interconnection and visualization.

2.2 Related Work of Knowledge Correlating

2.2.1 Increasing Needs of Development Support

The development process is getting more intensive The progress of project development depends on both innovation and experience. As more organisations seek to gain competitive advantage through timely deployment of services and products that meet and exceed customer needs and expectations, developers are under increasing pressure to develop new or enhanced implementations quickly [39]. Facing such challenges, an organization's developers, researchers, and practitioners, should reuse the knowledge and achievements of other areas and disciplines [40]. Thus, the development support of resuing the correlated knowledge and achievements is gaining importance to save human resource costs in solving common problems.

The development process is not easy to understand The development requirement plays an important role throughout the development process, describing and clarifying the correlated properties of the application domain, selecting system specifications, evaluating design alternatives, and validating design [41]. However, it takes a very complex development process to make solutions dedicated for the requirement, either by inventing new solutions or reusing the existing ones. Misinterpretation of the requirement or mismatching with the solutions usually cause the failure of project development. Fortunately, development activities nowadays generate a number of develop-

ment data. Such data holds rich information about the design, making it possible to have deeper investigations and a better understanding of the requirements. Therefore, it is worthwhile to analyze development activities to improve the matching between the solutions and the requirements.

2.2.2 Knowledge Correlating for Development Support

2.2.2.1 The development activities result in large amounts of data

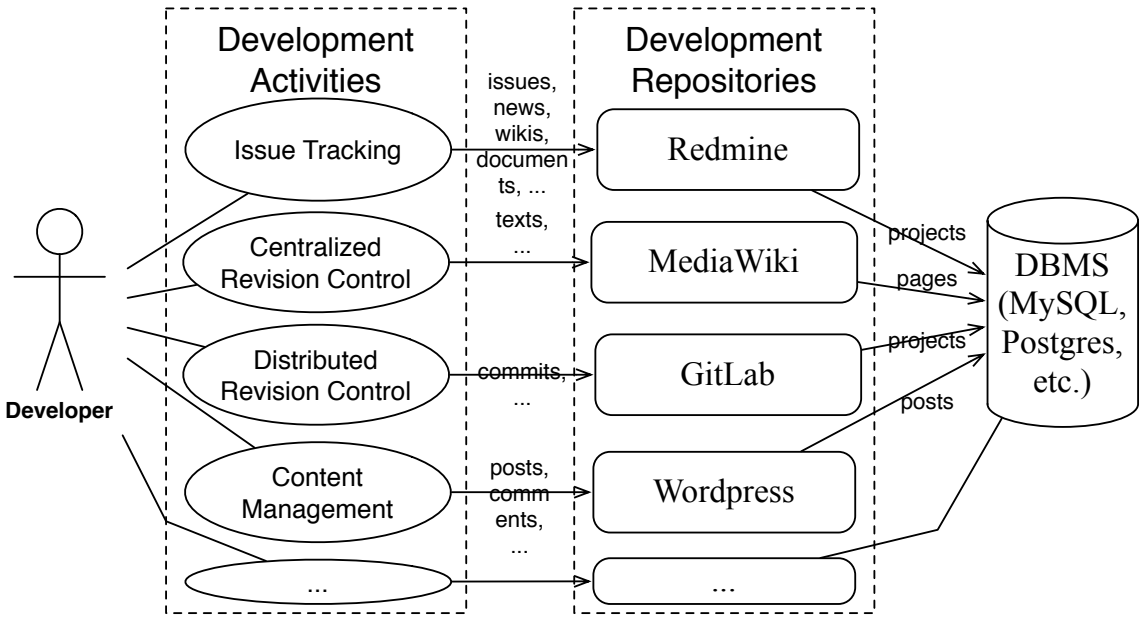


Figure 2.3. Development Activities result in Large Amount of Data

Internet-based collaboration and working tools provide convenient environment for project development while recording development activities and knowledge objects in various formats. The development process is also better formalized by the tools; for example, the software development life cycle [42] splits the software development into distinct phases (or stages) containing activities with the intent of better planning and management. In Figure 2.3, we introduce some critical development activities that have been widely supported by support systems. The performance of some open-sourced support systems is already as good as that of the commercial ones. More importantly, they could be freely redeployed and customized in local servers, working as private repositories.

- Issue tracking helps organizations manage issue reporting, assignment, tracking, resolution, and archiving. A study [43] on the social nature of issue tracking shows that in spite of teams

being collocated, which affords frequent and face-to-face communication, the issue tracker is still used as a fundamental communication channel. Typical tools, such as *Redmine* [21, 44] which is open-sourced, can store the data of issue tracking activities and development projects in private Web servers as private development repositories.

- Revision (version) control [24] records the changes to a file or set of files over time so that specific versions can be recalled later. Examples of revision features include reverting files back to a previous state, comparing changes over time, investigating detail of the historical changes.
  - As a centralized revision control system, the Wiki [45] provides a web-based evolving knowledge repository where users are encouraged to make additions by adding new documents or working on existing ones. The *MediaWiki* [30] is a typical open source tool, which can store development related text, files and pages in private Web servers.
  - On the other hand, as a distributed revision control system, series of Git [46] tools support the collaborative software development through tracking, branching, merging, and managing code revisions. And the *GitLab* [25] is a typical open-sourced one, which can store the project based revision control activities in private Web servers.
- The content management system (e.g. web blogs), provides a set of processes and technologies that support collecting, managing and publishing information in free form or medium. The *Wordpress* [32] is a typical open-sourced one that can store the text, images, videos in private Web server.

2.2.2.2 Knowledge Correlating for Development Support

As mentioned in the scenario in Figure 4.1, knowledge gap commonly exists among developers, requiring the development support of utilizing correlated knowledge objects which are stored in different development systems. Such support is related to a system's recommendation [47], which recommends knowledge based on the item assortment, user preferences, and recommendation criteria. Such recommendation systems are applied in scenarios where alternative approaches can not be simply applied, such as content-based filtering [48], which identifies the common characteristics

of contents, and collaborative filtering [49], which relies on users' heuristic ratings. The major challenges for the utilizing correlated knowledge are insufficient information coverage, including the lack of deep insights to the correlations among the knowledge objects that are under development, and lack of a broad view to the correlations among knowledge objects stored in different information silos.

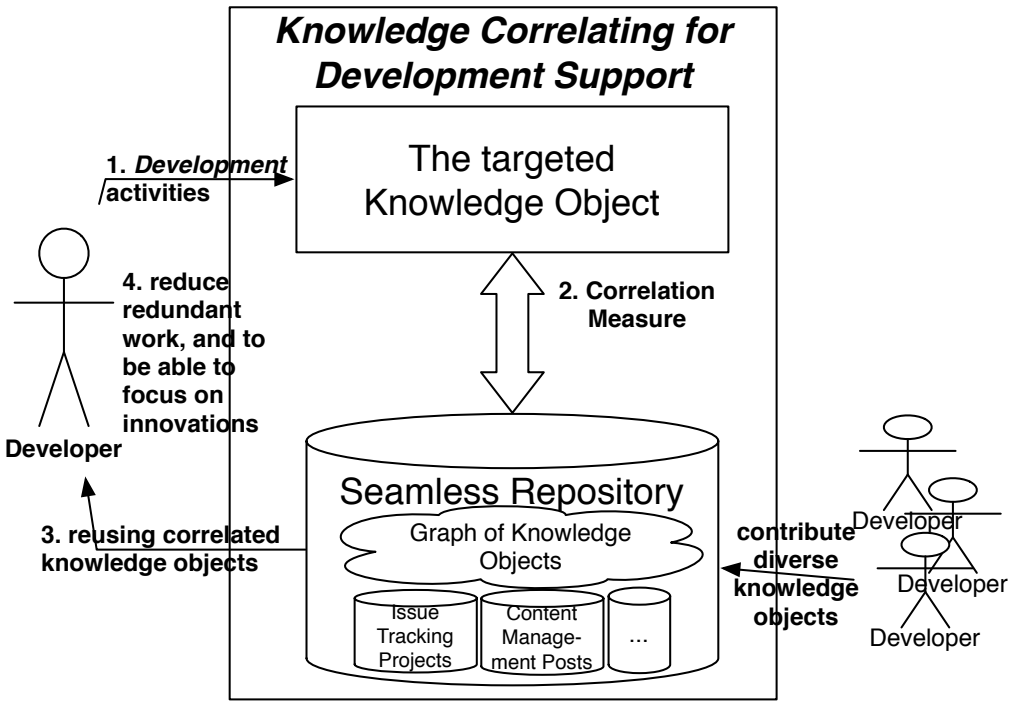


Figure 2.4. Knowledge Correlating for Development Support

By considering the challenges and scenarios, Figure 2.4 shows the use case of the activity awareness for development support, which identifies the developers' needs by analyzing the development activity data. The seamless repository integrates multiple support systems that are preferred by different developers, such as *Wordpress* and *Redmine*, and provide a graph of knowledge objects. By comparing the identified needs of the the items in the knowledge graph, correlated knowledge objects are selected to provide support for development.

2.2.3 Related Correlation Measures

Measuring the correlation among knowledge pieces is a major problem in knowledge management. There are basically two approaches with respect to the correlation measure; one is the contextual correlation regarding to specific domain knowledge, and the other is the relational correlation

regarding to the relations defined by people using their heuristic experience.

Assuming that the context could be represented by linguistic symbols (the terms), *tfidf* [50] is a widely-used model to represent the context of knowledge pieces as vectors, contextual correlation could be measured by computing the similarity among those vectors. This approach can measure the correlation among the knowledge pieces, which have more frequent common terms. However, it outputs different measuring results with respect to the dictionaries of different knowledge domains. It faces additional issues such as content privacy, possibly insufficient contents of single knowledge pieces, and difficulty in extracting the terms regarding to dictionary.

Again, assuming that the relations among knowledge pieces could be represented by the links appended by developers according to their heuristic experience, the neighbourhood or graph based models measure the correlation by computing the nearest neighbourhood through the links. This approach is independent from the knowledge domains, but it also faces problem such as possibly insufficient links of single knowledge pieces, and difficulty in extracting the links.

Table 2.4. Comparison of Related Correlation Measures

Correlation Measure	Contextual Correlation (using terms-frequency)	Relational Correlation (using links-ratio)	Integrated Correlation (using both terms-frequency and links-ratio)
Coverage	Low	Middle	High
Complexity	Middle	Middle	Low (But not too slow)
Feasibility	Middle	High	Middle
Usability	Low	Middle	High

The proposed measure using both terms-frequency and links-ratio is an integration of contextual and relational correlation. The comparison of related correlation measures is summarized in Table 2.4. According to the experimental results in Table 5.2, the proposed correlation measure is much better in information coverage, and with tolerable performance in computational cost. Furthermore, according to the layered system architecture in Figure 4.6, this correlation measure is also feasible to be implemented. Last but not least, according to the useful scenarios mentioned in Figure 4.1, this correlation measure is useful in supporting the utilization of knowledge objects in the seamlessly integrated repository.

Chapter 3

Seamless Integration

The seamless integration aims at improving the supports of sharing, interconnection, and visualization to bridge the gaps of information, communication, and representation in collaborative workflow. Three scenarios are introduced in Section 3.1 to explain how to bridge the three gaps by using the services provided by the seamlessly integrated repository. And the modeling and system architecture for the seamless integration is given in Section 3.2 and Section 3.3.

- ***Improved Support of Sharing to bridge the Gap of Information***: To overcome barriers of heterogeneity and workflow complexity, we introduce the advancing functionality of existing collaborative workflow tools to improve participants' capability and to ensure paperless pervasive teamwork among participants coming from different backgrounds. A local cloud service is constructed over TCP/IP networking secured by a firewall, and deploy a series of collaborative workflow support systems are integrated in a unified interface containing categorized and nested entries to sub repositories. And based on the integrated sub repositories, a portfolio service (see the demonstration in Figure 5.1) is introduced to make it easier for co-workers to know each other more quickly; and a service of workflow templates (see the demonstration in Figure 5.2) is also introduced to ease the workflow complexity.

- ***Improved Support of Interconnection to bridge the Gap of Communication***: To overcome the barriers of workplace conflicts, poor communication, and workflow complexity, we interconnect the data and functionality of collaborative workflow to break over isolated working environments, bring participants closer, and bridge the workflow's various gaps. A unified

search entry is provided on the web portal (see Figure 5.3) towards the whole data in the sub repositories. We also utilize the RESTFul Application Programming Interface (API) and Open Database Connectivity (ODBC) to integrate data-centric services with syndication, transferring, and notification functionality. Participants are better notified by transferred mails and instant messages about real-time updates from the whole team (see the demonstration in Figure 5.4).

- ***Improved Support of Visualization to bridge the Gap of Representation***: To overcome barriers of teamwork disruption and a less immersive experience, we provide participants more deep and attractive insights to their own collaborative workflow. A animation of teamwork involvement in Figure 5.5 brings users immersive experience in teamwork. Furthermore, analyzed collaborative workflow patterns, such as a daily and hourly heatmap as shown in Figure 5.6, bring participants a better understanding of the status of the whole teamwork.

3.1 Scenarios in Seamless Integration

3.1.1 Scenario of Bridging the Gap of Information

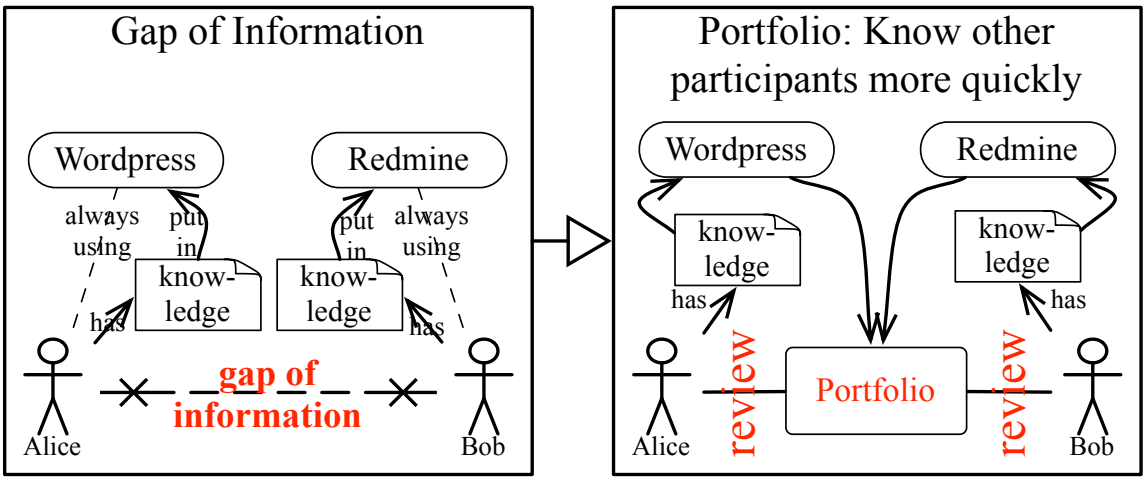


Figure 3.1. Scenario showing the Support of Sharing to Bridge the Gap of Information (taking the Portfolio Service as an example)

The support of sharing aims at reducing the gap of information, and the scenario of using portfolio to improve the support of sharing is described in Figure 3.1. Suppose that Alice is always

using Wordpress as support system, and she shares a lot of her knowledge there, and Bob puts a lot of his know-how in Redmine, which he always uses. There is a gap of information if Alice and Bob are both not familiar with each other's support systems. The portfolio automatically retrieves the shared information from the multiple support systems, and helps Alice and Bob know each other more quickly by reviewing those integrated information, such as participated projects, historical activity logs.

3.1.2 Scenario of Bridging the Gap of Communication

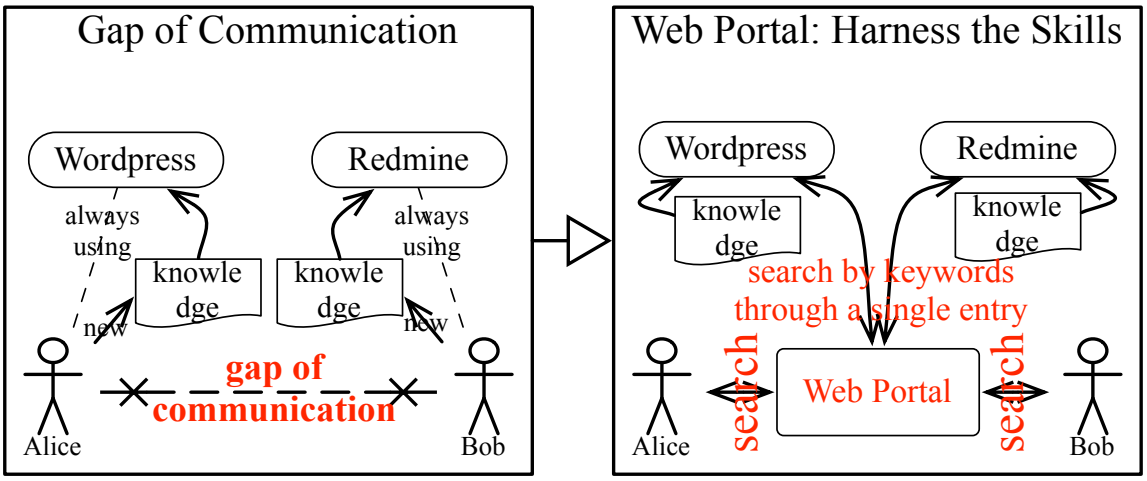


Figure 3.2. Scenario showing the Support of Interconnection to Bridge the Gap of Communication (taking the Web Portal Service as an example)

The support of visualization aims at reducing the gap of representation, and the scenario of using the heat-map service to improve the visualization is described in Figure 3.2. Suppose that Eve is the leader of Alice and Bob, she wants to check their weekly teamwork performance but has less interests in deeping into the details. The daily and hourly heat-map of Alice and Bob's teamwork involvement can give Eve a broad sense of the whole team. Never the less, with the heat-map, Eve can find out the teamwork patterns, or find out the abnormality of the teamwork.

3.1.3 Scenario of Bridging the Gap of Representation

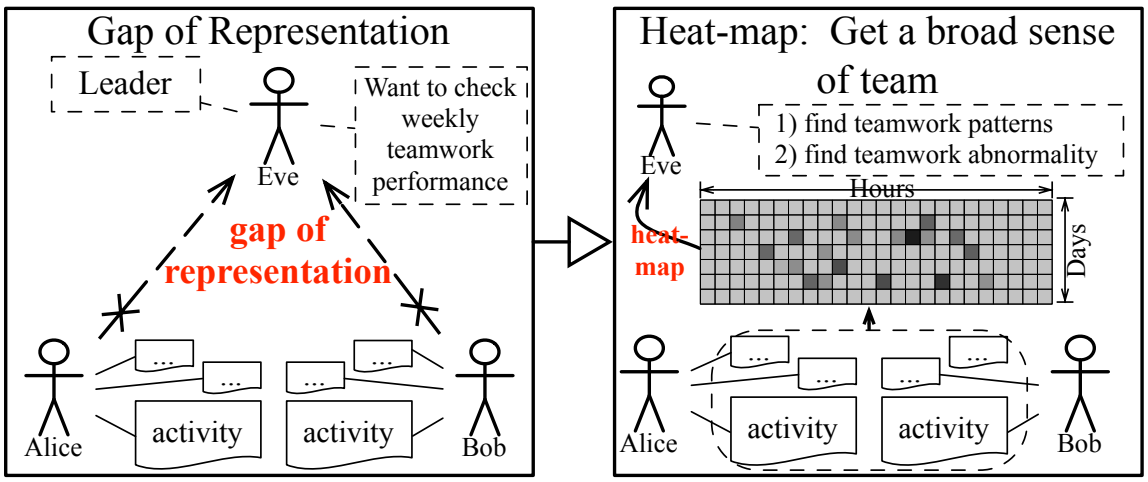


Figure 3.3. Scenario showing the Support of Visualization to Bridge the Gap of Representaiton (taking the Heat-map Service as an example)

The support of correlation aims at reducing the gap of knowledge, and the scenario of using the correlation measure to improve the knowledge correlation is described in Figure 3.3. Suppose Alice and Bob are both experienced in very different domain knowledge, they may feel hard to understand or utilize each other's knowledge or experience. The correlation measure correlates their knowledge objects based on a larger set of knowledge repository, so that they better utilize each other's correlated knowledge objects by passing through the knowledge graph in Figure 3.3.

3.2 Modeling of Seamless Integration

3.2.1 Information Model of Collaborative Workflow Activity

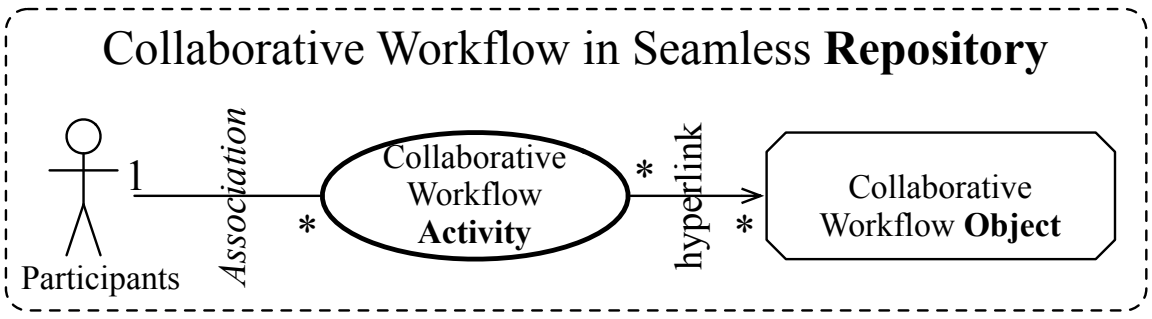


Figure 3.4. Information Model for Seamless Integration

Figure 3.4 illustrates the relation among five main factors that are involved in collaborative workflow (*CW*).

- ***Participants*** have membership and roles in teamwork, collaborating with each other in a series of activities through the collaborative workflow software.
- ***Collaborative Workflow Activities*** are granular activity records generated by participants when they are working and collaborating with each other.
  - Workflow Activities (*WA*) are supported and specialized by workflow software. Participants frequently embed referential links among the *WA* or communicating contents to chain the workflow steps into a more meaningful, more logical, and larger workflow to deal with complex tasks.
  - *Collaborative Activities (CA)* are events involving teamwork-oriented collaboration including on-site collaboration such as a meetup in office, and online collaboration such as transferring mails and instant messages among participants or between participants and systems. For efficient information delivery, participants frequently embed referential links into the communication contents, referring to the existing workflow or collaborative activities.
- ***Collaborative Workflow Objects*** are the achievements from accumulated *CW* activities. Some examples could be ready-to-publish products or work-in-progress projects or archives.
- ***Association*** refers to undirected relations between the participants and *CA* activities. One *CA* record can be associated with only one participant, while the one participant can have multiple *CA* activities.
- ***Hyperlink*** is denoted as a user-created reference during collaborative workflow, such as an embedded URL (uniform resource locator), a workflow object ID, or a symbolic tag that points to an object on the Web. Since the seamless repository is built using Web technology, any contents (e.g. *CW* activities, *CW* achievements) in the repository can be considered as *CW* objects on the Web. The hyperlinks build referential relations to those *CW* objects. Also, one *CW* activity can have multiple hyperlinks pointing to one or more *CW* objects. Hyperlinks can be automatically detected by URL parsing functions, and based on detected

hyperlinks with which we build the graph to understand the contribution weight (also known as teamwork involvement as defined in Section 3.2.3) of users.

Series of workflow and collaborative activities constitute the collaborative workflow, more detailed relations are given in Section 3.2.2 (see, Figure 3.5).

The key breakthrough point in overcoming various barriers in the collaborative workflow is to measure and enhance the teamwork involvement, the more active the teamwork is, the higher potentials to achieve the success in projects. To understand the information structure of the collaborative workflow, we generalize the activity data model over different collaborative workflow repositories in Section 3.2.2. And to evaluate the activeness of teamwork participation, we formulate the teamwork involvement as a statistical measure to the collaborative workflow activity data in Section 3.2.3.

3.2.2 Data Model of Collaborative Workflow Activity

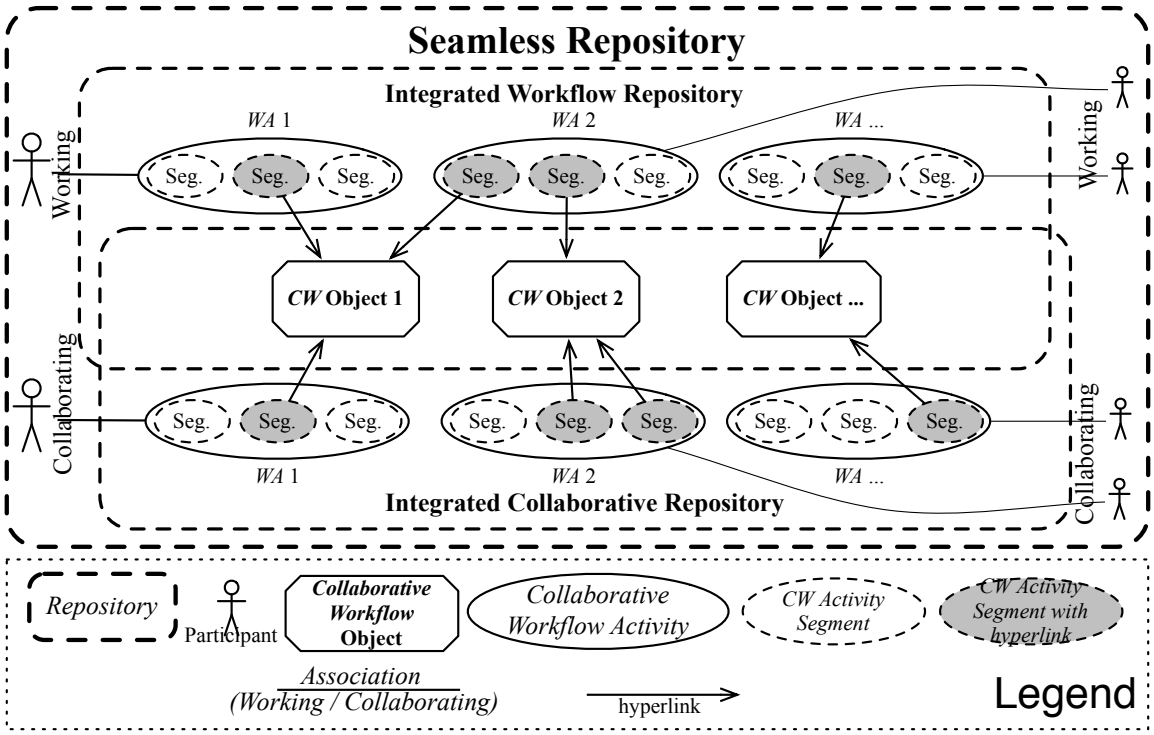


Figure 3.5. Data Model of Collaborative Workflow Activity

Based on the five major factors (participants, associations, *CW* activities, *CW* objects, and hyperlinks) in the information model as shown in Figure 3.4, a detailed information structure of collaborative workflow is illustrated in Figure 3.5. Multiple teamwork participants, who are working

and collaborating together in teamwork, generate multiple *CW* activities. The *CW* activity data is well-structured, and data schema samples are given in Figure 3.6. There are hyperlinks (e.g. URLs) and *CW* object (e.g. object IDs or symbolic tags) in the raw data of *CW* activities, building the referential relations from the *CW* activities to the *CW* objects. This dissertation mainly focus on the *CA* data within the seamlessly integrated repository, and pay less focus to the referred objects in other public clouds or repositories.

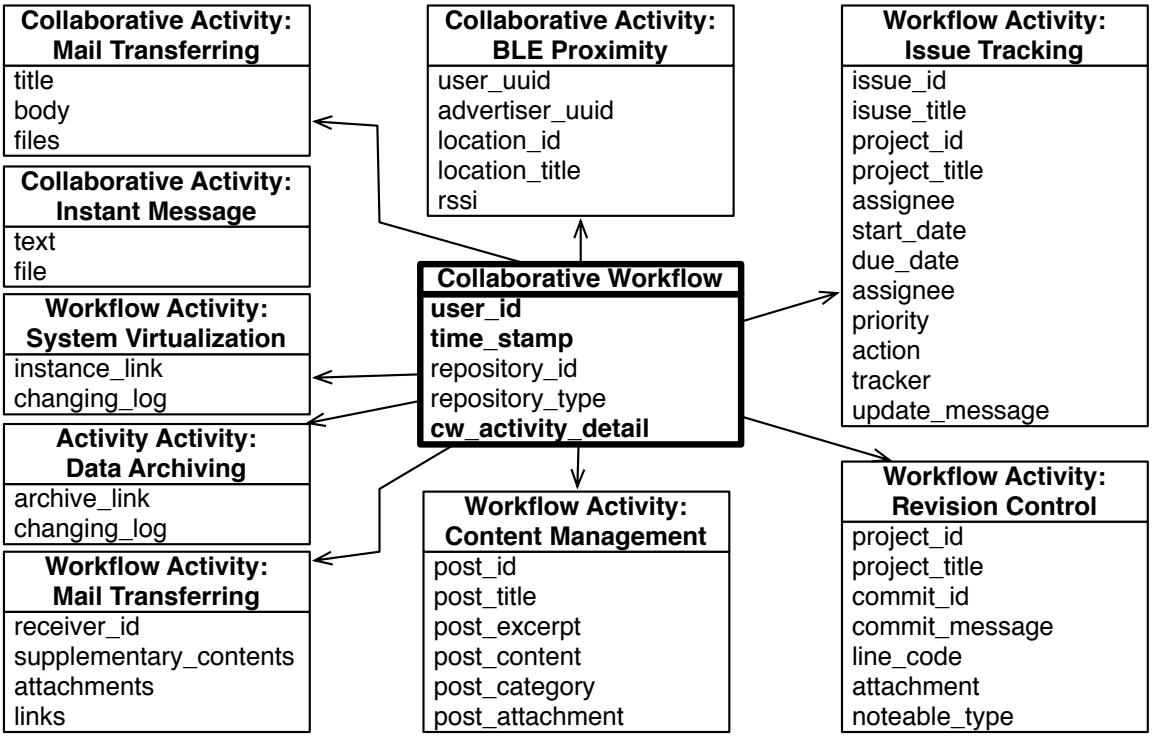


Figure 3.6. Integrated Data Schema of Collaborative Workflow Activity Data from Multiple Repositories

A data schema is generalized to integrate heterogeneous activities in Figure 3.6, including a general header which includes the *user id*, *time stamp*, *repository id*, *repository type*, and details of a *collaborative workflow activity* that point to the data schema of specific *CA* activities in other tables.

3.2.3 Teamwork Involvement Model

Since collaboration barriers in teamwork result in poor participation, the participants' involvement (the degree or intensiveness of participation activities in the workflow) is an important measure to the success and quality of a project. Here we try to formulate the involvement I in teamwork.

Given the collaborative workflow *CW* process over deployed repositories R during a period of time τ (starting from the initial time stamp τ_0), the expression for the five major information factors in a collaborative workflow repository are given as follows:

- Participants: $P = \{p_1, p_2, \ldots, p_h, \ldots\}$, and $|P| = H$,
- *CW* activities: $A = \{a_1, a_2, \ldots, a_k, \ldots\}$, and $|A| = K$.
- Associations: $Aso = \{aso_{(k,h)} | \forall p_h \in P \textit{ and } \forall a_k \in A\}$
- *CW* objects: $V = \{v_1, v_2, \ldots, v_n, \ldots\}$, and $|V| = N$
- Hyperlinks: $Lnk = \{lnk_{(k,n)} | \forall a_k \in A \textit{ and } \forall v_n \in V\}$.

Algorithm 1 Calculation of Teamwork Involvement

```
1:  Let P be the ID list of all participants
2:  Let Aso be the enumeration of participants' for all CW activities
3:  Let CWObjURIs be the enumeration of URIs of all CW Objects in Repository
4:  procedure EnumerateHyperLinks(CWObjURIs, A) ▷ Enumerate the Hyperlinks that point from CW activities to CW objects
5:      Lnk ← []                                          ▷ Initialize an empty list for storing the hyperlinks
6:      for a_k ∈ A do
7:          TmpCWLnk ← []          ▷ Initialize a temporal empty list for storing the hyperlinks from a single CW activity
8:          TmpDoc ← Extract the textual contents from a_k according to data schema (in Fig. 3.6).
9:          TmpURIs ← Decode the URIs in TmpDoc, including URLs, Object ID and symbolic tag of CW objects
10:         for TmpURI ∈ TmpURIs do
11:             for TmpCWObjURI ∈ CWObjURIs do
12:                 if TmpURI is pointing to the objects belonging to the object associated with TmpCWObjURI then
13:                     Append
14:                 end if
15:                 TmpCWLnk.append(TmpURI)
16:             end for
17:         end for
            Lnk.append(TmpCWLnk)
18:     end for
            return Lnk
19: end procedure
20: procedure CalculateTeamworkInvolvement(k h, Aso Lnk)   ▷ Calculate the Teamwork Involvement for Participants p_h
    in CW activity a_k
21:     TI_{k,h} ← 0                                      ▷ Participant p_h' Teamwork Involvement in CW activity a_k
22:     if h ∈ Aso[k] then
23:         TI_{k,h} ← count(Lnk[k])
24:     end if
            return TI_{k,h}
25: end procedure
26: Lnk ← EnumerateHyperLinks(CWObjURIs, A)                                  ▷ Calculate and store Lnk
27: CalculateTeamworkInvolvement(k h, Aso Lnk)
```

The teamwork involvement of one participant p_h though one activity a_k is calculated by the amount of hyper-links from a_k to all collaborative objects (v_n) (see, Eq. 3.1). And participants' teamwork involvement in *CW* objects is formalized in Eq. 3.2. And we provide the pseudo-code

for calculating the teamwork involvement in Algorithm 1.

$$I_{(p_h,a_k)} = \begin{cases} \sum\limits_{\substack{lnk_{(k,n')} \in Lnk \\ \forall v_n \in V}} |lnk_{(k,n')}| & , aso_{(h,k)} \in Aso \\ 0 & , aso_{(h,k)} \notin Aso \end{cases} \tag{3.1}$$

$$I_{(p_h,v_n)} = \sum_{a_k \text{is during } [\tau_0,\tau_0+\tau]} I_{(p_h,a_k)} \tag{3.2}$$

3.3 Three-layered System Architecture for Seamless Integration

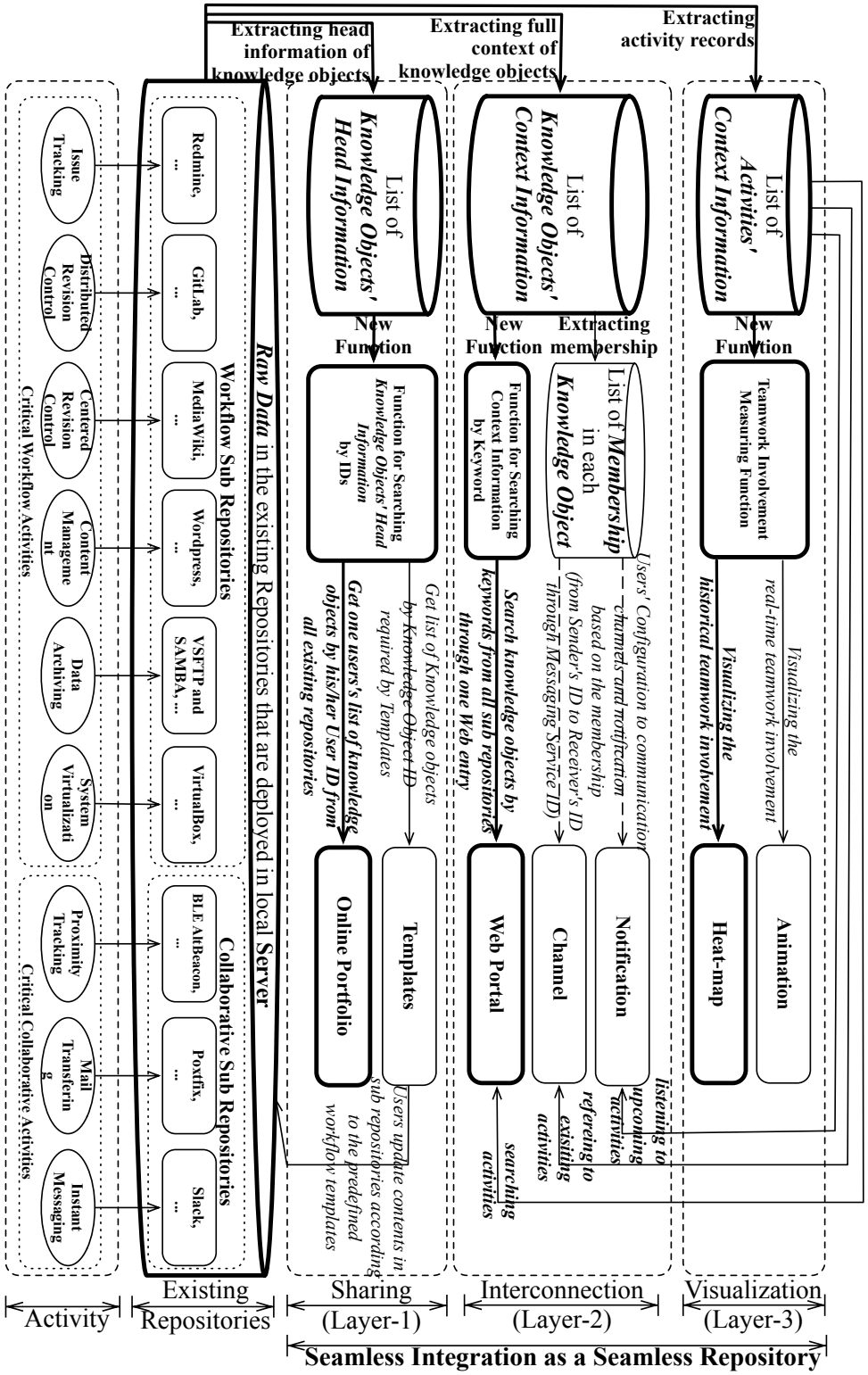


Figure 3.7. Seamless Repository for Pervasive Teamwork

Figure 3.7 shows the system architecture for seamless integration. The teamwork participants generate collaborative workflow activities and post into their preferred sub repositories, e.g. Wordpress, Redmine. The raw data in the existing sub repositories are then used to implement the support of sharing, interconnection, and visualizaiton by recalling the API of those multiple support systems. If the necessary API does not exist, or expires due to updates, our seamless integration will extract the list of knowledge objects (such as pages in Wordpress, projects in Redmine) from the raw data, and implement new functions to compensate the missing APIs.

In order to set up an workable seamless repository and also to be feasible for experiment, we target the *CW* in IT environment. We select 9 types of critical *CW* activities which are common in research and development tasks, and introduce them to the teamwork participants. And based on the *CW* activity data model and teamwork involvement measure (in Section 3.2), we design the system framework with 3 layers to support the 9 types of critical *CW* activities in Section 3.3.

1) Layer 1 aims at bridging the gap of information. This layer helps to ease the *CW* barriers of heterogeneity and workflow complexity. It consists of deployed and implemented repositories to support critical workflow activities and collaborative activities. And based on the sub repositories, there are services of portfolio and workflow templates.

2) Layer 2 aims at bridging the gap of communication. It consists of a web portal as a single entry to access and search the data and functionality to the sub repositories; and secondly a notification function to notify the participants the recent updates of *CW* activity updates that they concern through mail transferring or instant messaging. This layer helps to ease the barriers of skills gap, workplace conflict, and poor communication.

3) Layer 3 aims at bridging the gap of representation. It provides the teamwork involvement real-time animation and heat-map for an attractive and deep insights to the teamwork. This layer helps to ease the barriers of teamwork disruption and less immersive experience.

The three layers of the seamless repository serves improved support of sharing, interconnection, and visualization to bridge the gaps of information, communication, and representation respectively. From the vertical point of view to the framework, the higher layer takes advantage of the encapsulated service from the lower layer. And from the horizontal point of view, the layered service is tolerable with the failure of local components.

3.3.1 Implementation of Layer1: Support of Sharing

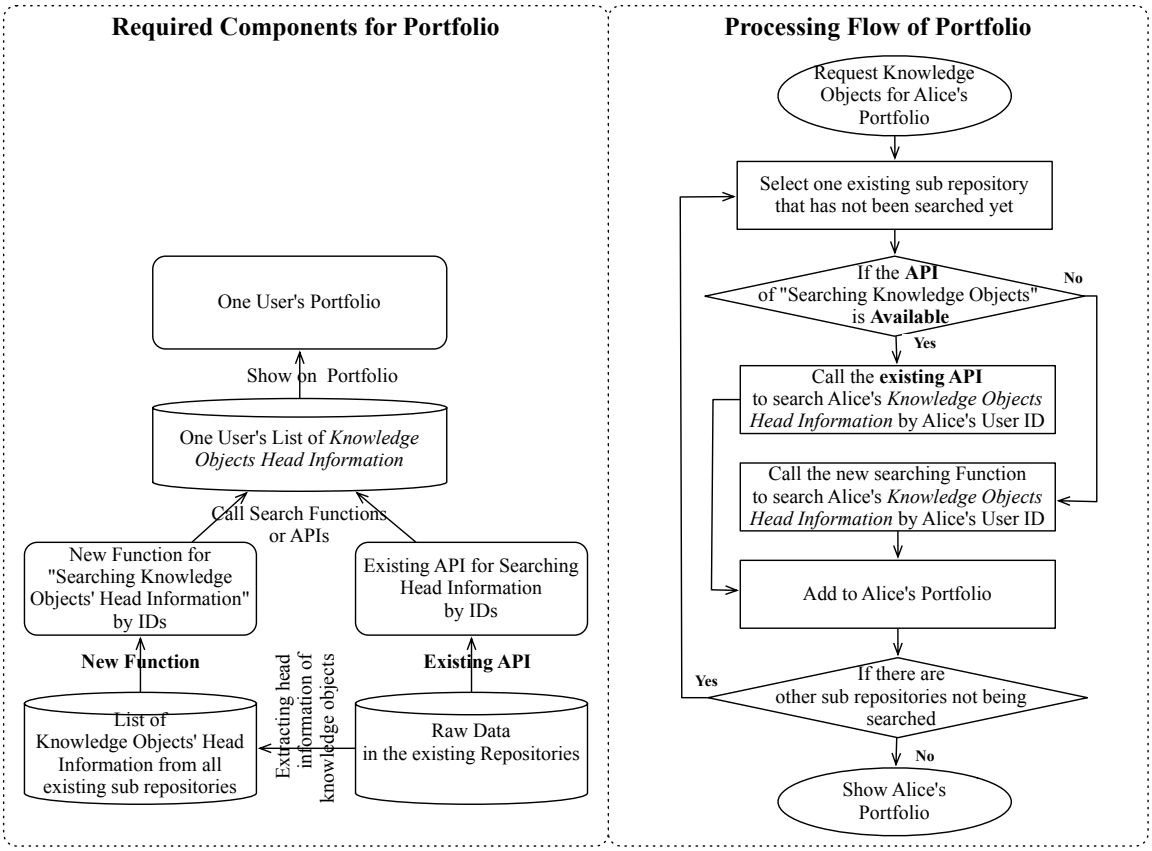


Figure 3.8. System Design of Portfolio Service

The major contribution in implementing the sharing is enabling the search by Object ID (e.g. User ID, Knowledge Object ID) to the multiple sub repositories. The system design of portfolio service is given in Figure 3.8, it shows the implementation of portfolio service, including the required components on the left side and processing flow on the right side. If there is no existing APIs for “searching knowledge objects’ head information by IDs" in the sub repositories, or the APIs expire in the latest version of the sub repositories, we implement the new search functions. Then we utilize the search APIs and new search functions to get a list of each user’s historical achievements, such as accomplished projects, from the multiple systems for building his/her online portfolio. A demonstration of portfolio service in Layer 1 is given in Figure 5.1 to reduce the gap of information.

3.3.2 Implementation of Layer 2: Support of Interconnection

The major contribution in implementing the interconnection is enabling the bridging the objects the multiple sub repositories by keywords, for example search by keywords. The system design of web portal service is given in 3.9, it shows the implementation of Web Portal service, including the required components one the left side and processing flow on the right side. If there is no existing API for “searching knowledge objects’ or activities’ context information by keywords", or the APIs expire in the latest version of the sub repositories, we implement the new search functions. Then we utilize the search APIs or search functions to find a list of knowledge objects or activities that match users’ given keywords from the multiple sub repositories, and show the list as the searching result in the Web portal. A demonstration of Web portal service in Layer 2 is given in Figure 5.3 to reduce the gap of communication.

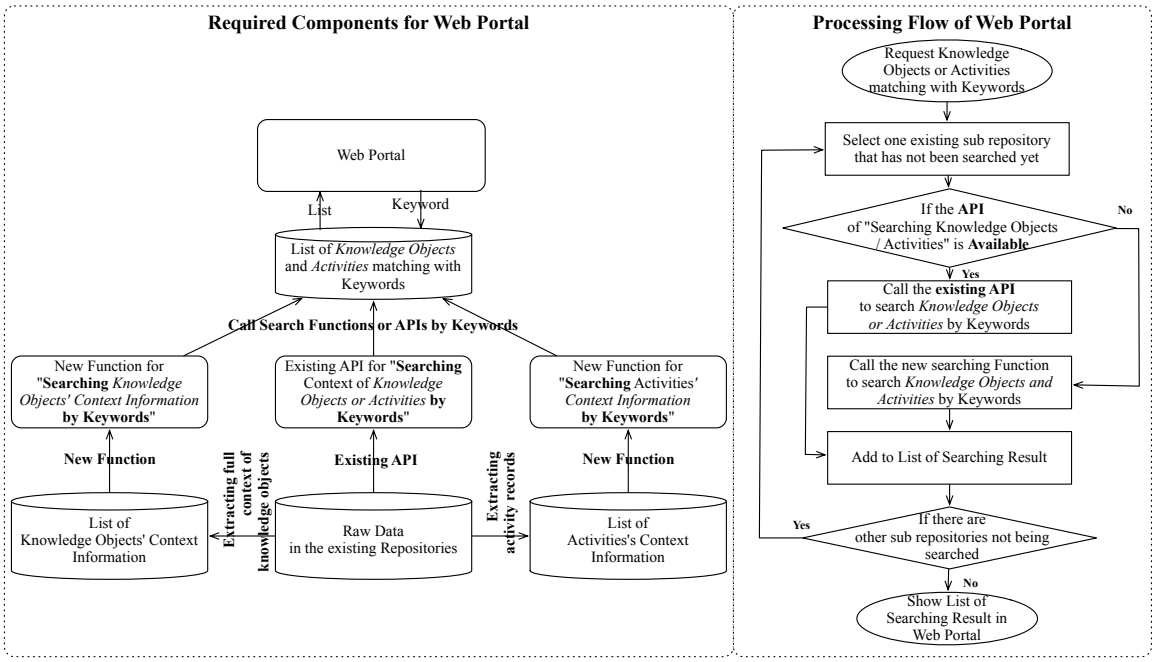


Figure 3.9. System Design of Web Portal Service

3.3.3 Implementation of Layer 3: Support of Visualization

The major contribution here is that a teamwork involvement measuring function is defined to assign the weight to each activity. And according to the weights, the platform visualizes the historical heat-map of collaborative workflow activities to give a broad sense of teamwork, and visualize the

real-time animation of collaborative workflow activities to give a narrow sense of each individuals real-time performance.

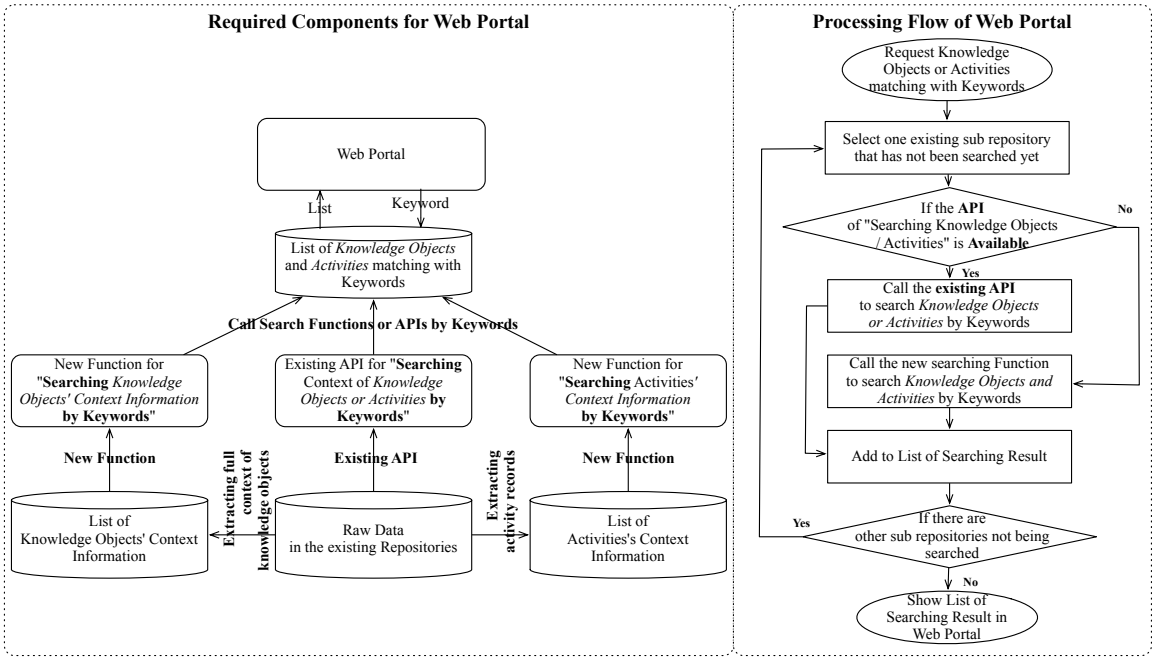


Figure 3.10. System Design of Heat-map Service

Figure 3.10 shows the system design of heat-map service. We first implement the teamwork involvement function to calculate the list of teamwork involvement of each activity; and then calculate the hourly heat-ness to plot the heat-map with 24 hours as horizontal axis and 7 days as vertical axis. A demonstration of heat-map service is illustrated in Figure 5.6 to give a broad sense about teamwork performance.

3.4 Integration of Existing Support Systems

3.4.1 Integration of Existing Support Systems for Critical Workflow Activities

Table 3.1 enumerates the software that we used to deploy and develop the the workflow repositories for the critical workflow activities, such as issue tracking. And the table also illustrates the data and functionality that each workflow repository provides.

- The *issue tracking* workflow in Redmine provides sub workflows such as progress status tracking through the Issue/Time tracking module, documentary reporting through the doc-

ument module, and system specification through the Wiki module. For example, with the Issue/Time tracking module, trouble ticketing activities involve creating, updating, and resolving issues. The Redmine (Version-2.4) runs on various OSes, and is developed using a Ruby on Rails (RoR) web framework and a MySQL data base.

- The *distributed revision control* workflow in GitLab supports Git versioning activities, such as adding, committing and pushing the change of source code. The GitLab (Version-7.0) runs over Linux variations, and requires Ruby (Version-2.0) and a database such as MySQL.

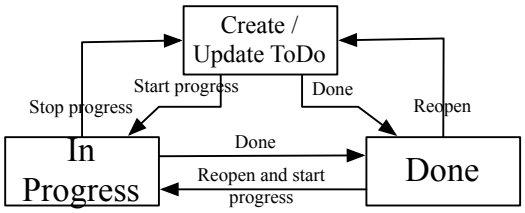
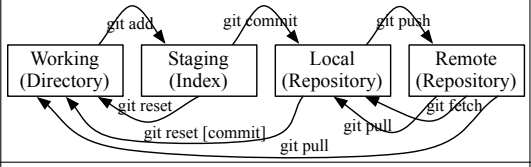
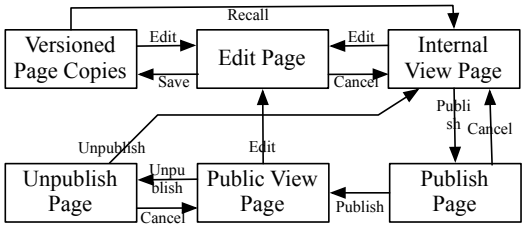
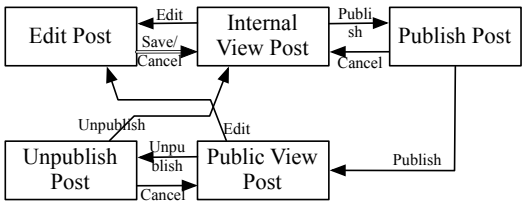
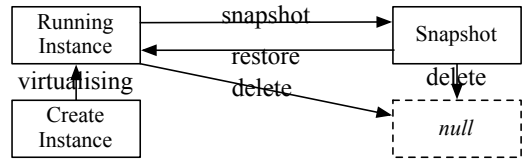
- The *centered revision control* workflow in MediaWiki supports Wiki documenting activities, such as adding, submitting, and updating Wiki pages. MediaWiki (Version-1.23) runs on various OSes, and a LAMP (Linux-Apache-MySQL-PHP) server environment is recommended.

- The *content management* workflow in Wordpress supports press release activities such as creating, updating, and publishing a blog. Wordpress (Version-4.0) can run over various OSes, and requires a server environment such as LAMP.

- The *data archiving* workflow in vsFTP and SAMBA server supports archiving activities such as uploading, moving, and downloading files. For our setup, vsFTP (Version-3.0) runs on Linux Ubuntu, serving as a simple secure File Transfer Protocol (FTP) service. Meanwhile, SAMBA (Version-4.2) runs on various OSes, and integrates the Server Message Block (SMB) and Common Internet File Systems (CIFS) protocols for file sharing.

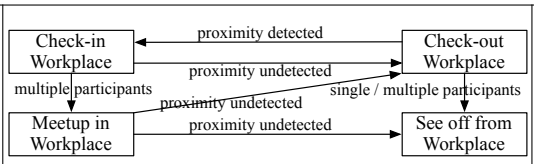
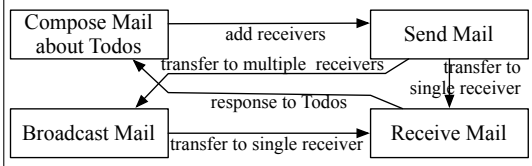
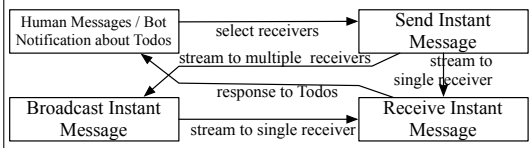
- The *system virtualization* workflow in VirtualBox supports virtual machine (VM) operating activities such as creating a virtual instance of an operating system, taking a snapshot of the instance, or cloning the instance. VirtualBox (Version-4.3) utilizes the x86 and AMD64/Intel64 processors, and supports various types OSes as both host and client. The virtualized OS can even be stored as disk image file, and can be archived and shared through an archive server.

Table 3.1. Integration of Existing Support Systems for Critical Workflow Repositories

Workflow	Software	Data Modules and Functionality	Repository
Issue Tracking	Redmine Version-2.4		http://[Local Host Server]/pm/p:80
Distributed Revision	GitLab Version-7.0		http://[Local Host Server]:8888
Centralized Revision	MediaWiki Version-1.23		http://[Local Host Server]/cm/wiki:80
Content Management	Wordpress Version-4.0		http://[Local Host Server]/cm/blog:80
Data Archiving	VSFTP Version-3.0, SAMBA Version-4.2		ftp://[Local Host Server]:21
System Virtualization	VirtualBox Version-4.3		Instances and snapshots can be stored on Data Archiving repository.

3.4.2 Integration of Existing Support Systems for Critical Collaborative Activities

Table 3.2. Integration of Existing Support Systems for Critical Collaborative Activities

Collaboration	Software	Data Modules and Functionality	Repository
Proximity Tracking	BLE AltBeacon Version-1.0		ftp://[Local Host Server]:3000 / 27017
Mail Transferring	Postfix Version-4.0, Squirrel-Mail Version-1.2		POP3 (Port 110), SMTP (Port 25), Webmail UI: http://[Local Host Server]/ squirrelmail:80
Instant Messaging	Slack		Integration with Slack service

- The *proximity tracking* based collaboration repository we deployed takes advantage of the BLE proximity [38] (see, Figure 3.11) is used to track the users onsite participation herein. Once user carrying on their smartphone (having BLE 4.0 and Internet enabled) enters the region of Beacon advertiser, the triggered proximity log is transmitted to the Proximity Log Server, then the Proximity Log Analysis server reports the status of user's onsite participation, including the indoor position and time of stay. The iOS and Android programming, NodeJS and MongoDB are used to construct this new service, and the AltBeacon [51] is used as the proximity protocol among BLE sensors.

- The *mail transferring* supports peer-to-peer or group based online communication. The deployed Postfix (Version-3.0) is a Linux based mail transferring agent, it supports the group mailing list which is useful for message broadcasting in teamwork. And the Squrrielmail (Version-1.2) provides the Web interface for teamwork participants to check the received mails and send emails to other receivers. The mail transferring system can be further inte-

grated with other repositories for sending notifications.

- The *instant messaging* also supports the agile online communication. The Slack is a popular instant messaging platform, it offers chat channels organized by topics, as well as private groups and direct messages. All contents in Slack are searchable. Most importantly, it can integrate many other services via easy configuration.

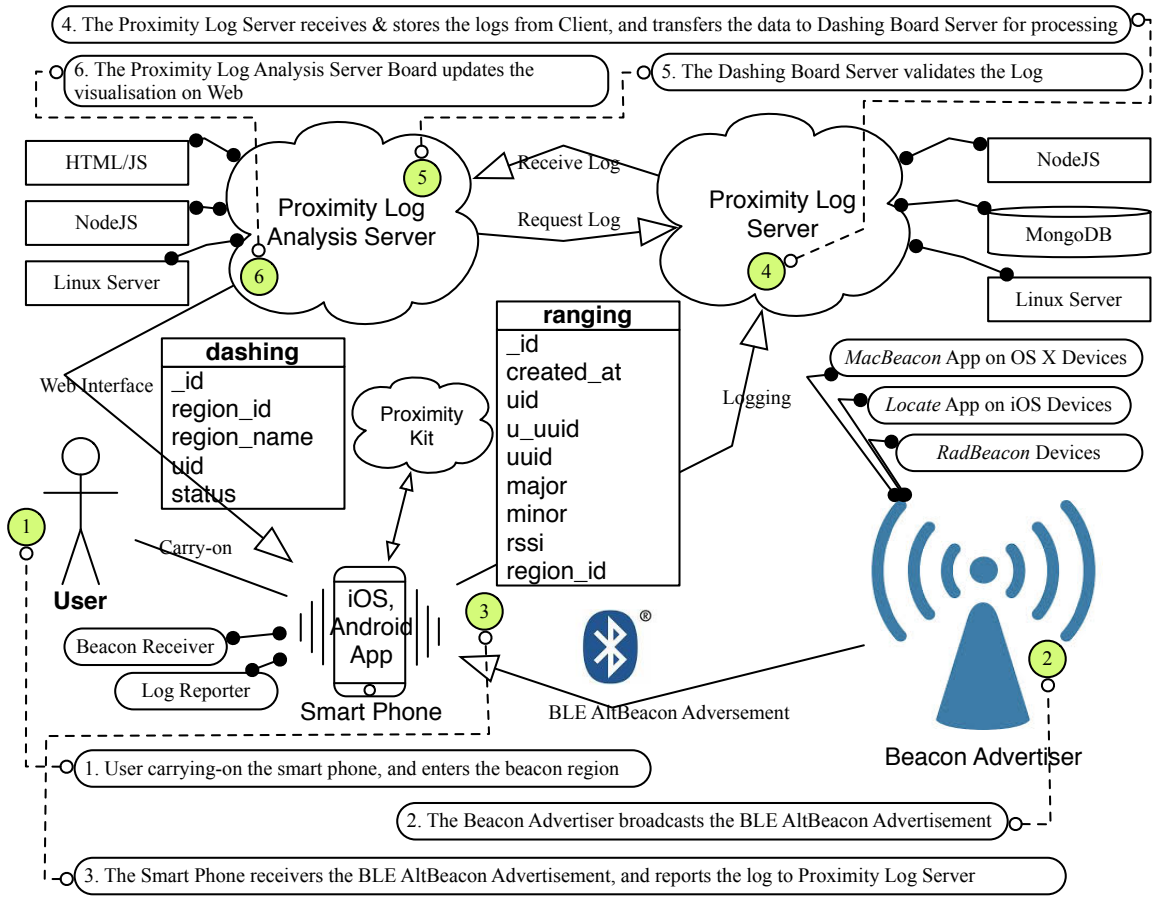


Figure 3.11. The Proximity Tracking of Onsite Collaborative Activities (e.g. Meetup and See off) based on BLE Proximity

The deployed and developed *CW* repositories bring the data and functionality pervasively available for participants. Besides, We also did effort to help the participants get familiar with operating the collaborative workflow in those repositories. Participants from different background can divide their work into pieces and each one contribute to different pieces through the cooperation on the repositories. The available data and functionality as well as the tasks division helps to simplify the complex workflow, and greatly enhance the feasibility of the tasks.

3.5 Quality of Service for Seamless Integration

Besides the aforementioned useful features in seamless integration, including supports of sharing, interconnection, and representation, the quality of service is also important, especially when it is meeting the security and scalability problems.

3.5.1 Information Security for Seamless Integration

The information security is critical for system integration. In order to make the platform workable, the integration requires the reading privileges to the database of existing support systems. There are four solutions outlined to guarantee the privacy and security.

- The platform saves the data in the local database, same as the existing support systems, which will not transfer the data outside the local network.
- The newly integrated head or context information is encrypted in the local new database, and only the administrator of the local server holds the key.
- In case of any emergency, the administrator can stop the platform service.
- The presented information to the local users will be filtered according to the their original privileges in the original support systems, which means that if the local user does not have the privilege to access the information, he or she still can not access through this platform.

3.5.2 System Scalability for Seamless Integration

The platform will face the system level scalability problem when a branding new support system needs to be integrated. We assume that the new support system is still using the commonly used database at the back end. The platform can do a general integration fundamentally from the database level if the administrator assigns the privileges to read the database of the existing support system. There is a data fusion function (such as shown in Figure 3.8, 3.9 , 3.10) in the platform that can automatically integrate the data from different sub repositories if the administrator can tell the data schema of the new support system. And there is formatted program interface in the platform to ease administrators' configuration to the data integration.

Chapter 4

Knowledge Correlating

As project development gets more intensive, there are increasing needs of development support by reusing shared knowledge objects, such as technical know-how and project achievements, which grow along with developers' activities through multiple support systems. However, there is a large gap of knowledge in providing such development support, because of developers' divergent background knowledge, as well as distinct personal preferences in using different support systems. To bridge the knowledge gap, the major challenge is to improve the information coverage in correlating the knowledge from different support systems. This challenge derives two issues: one is a deeper insight to the correlations among the knowledge objects that are developing and growing, and the other is a broad view to the correlations among knowledge objects that are stored in different support systems. In this chapter, The knowledge correlating for development support is proposed which identifies the development needs by analyzing developers' activities and finds the correlated knowledge objects from the integrated repository of multiple support systems. To this end, a term-frequency and chained links-ratio (TFCLR) based correlation measure to model the conceptual and relational correlation among the knowledge objects in the seamless repository. A graph representation among knowledge objects is output based on the correlation measure, and that helps to implement awareness service such as reusing the knowledge.

4.1 Scenario of Bridging the Gap of Knowledge

The project development process results in application or products, and meanwhile generates a large number of archive data along with development activities, such as computer programming, documenting, testing, and bug fixing. The historical development archives are rich in knowledge and experience information, and could be reused to solve similar problems in future. However, such archives are usually not well organized, making it difficult to reuse. Furthermore, commonly used search engines may have no privileges to access private or local development repositories.

As project development gets more intensive, the developers suffer in solving the increasing number of complex development problems with time constraints. Therefore, there is an increasing need of development support by utilizing existing knowledge objects, such as the technical know-how and project achievements, which are shared in the development repositories. However, because of developers' divergent background knowledge, and distinct personal preferences in using different support systems, there is a large gap of knowledge in providing the aforementioned development support.

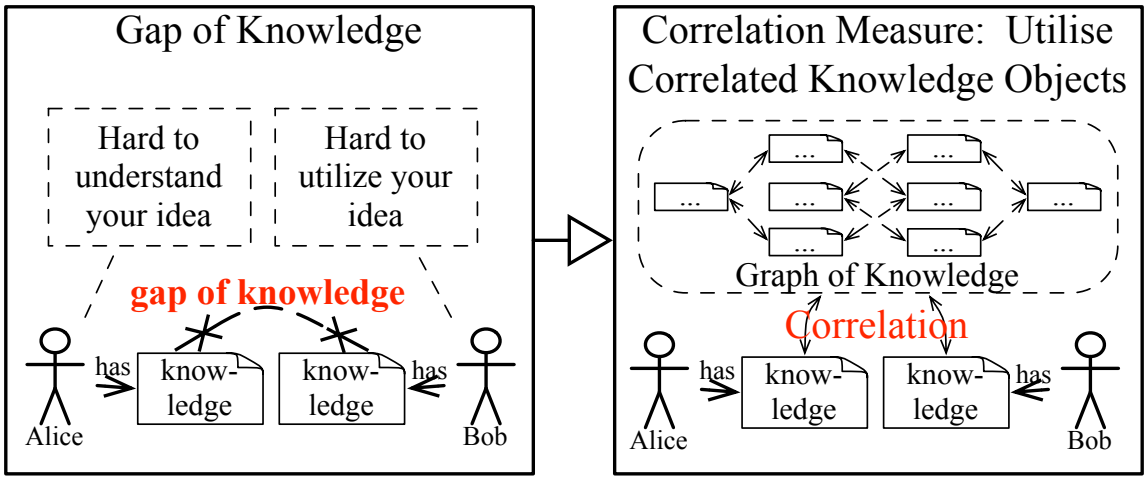


Figure 4.1. A Scenario of Reducing the Knowledge Gap

Figure 4.1 gives a scenario to show the necessity of reducing the knowledge gap. Suppose that Alice and Bob (in the left side of Figure 4.1) are both experienced in very different background knowledge, and have different preferences in using different support systems (e.g. *Wordpress*, *Redmine*), to which they apply their respective knowledge. Once they have a chance to work together in a development project, they feel that it is hard to understand or utilize each other's knowledge

or experience. Such a knowledge gap commonly exists among other developers, which calls for development support to bridge the knowledge gap by the unknown interrelationship among knowledge objects and covering the isolated information silos of each support system, as shown in the right side of Figure 4.1. Since Alice and Bob are already using support systems, their development activities are traceable, and hold dynamic information about their knowledge objects that are under developing and growing. Therefore, it is reasonable to identify the developers' needs by analyzing their development activity data and utilize the correlated knowledge objects from the integrated support systems.

Based on the scenario in Figure 4.1, the major challenge to bridge the knowledge gap is to improve the information coverage in correlating the knowledge from different support systems. This challenge derives two issues: a deeper insight to the correlations among the growing knowledge, and a broad view to the correlations among knowledge from multiple support systems.

To bridge the knowledge gap, we propose an knowledge correlating model for development support, which identifies the development needs by analyzing developers' activities, and finds the correlated knowledge objects from the integrated repository of multiple support systems. To over come the major challenge, we model the knowledge in graphs, and conceive an integrated correlation measure using terms-frequency and chained links-ratio (TFCLR) that are extracted from the activity data. In addition, we construct a seamless repository as an integrated development environment, which is able to collect and process a large amount of development activities and knowledge objects.

4.2 Modelling of Knowledge

In this section, we propose an integrated model to measure the correlations among knowledge objects in multiple development repositories. The software development process can be considered the evolution of *KO*s. For example, the process reflects in the growing domain concepts presented by *KO*s and the expanding relationships among *KO*s. In our web-based seamless repository (see Figure 2.3), there are large amounts of terms (such as words or tags) reflecting domain concepts and URL links constructing the relations. Therefore, we use a graph model (see Figure 4.2) to organize the *KO*s in development repositories, then use the teams-based model to measure the conceptual

correlation of *KO*s, and then use the links-based model to measure the relational correlation among the *KO*s.

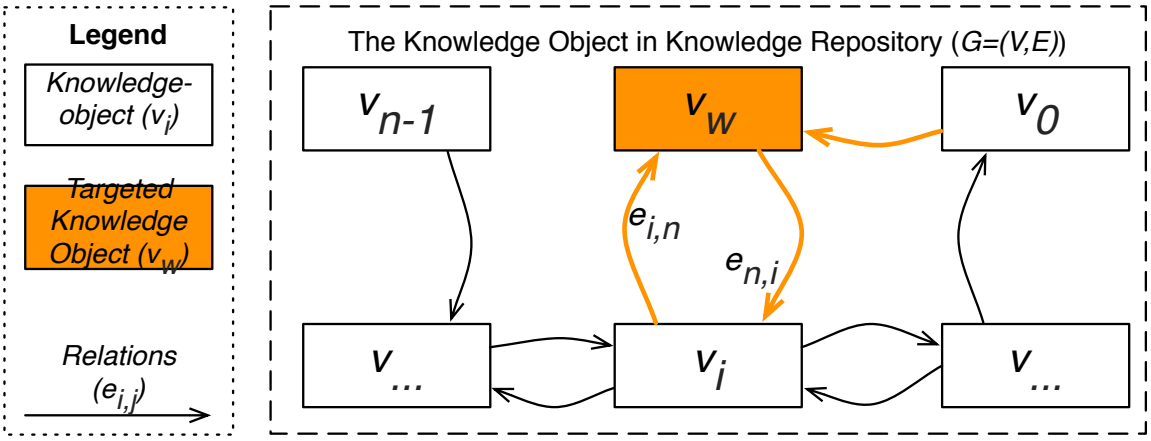


Figure 4.2. Graph Representation of Knowledge Objects

Table 4.1. Parameters for Graph of Knowledge Objects

Given: a graph of knowledge objects $G = (V, E)$, (where $V_{targeted}, V_{existing} \subset V$, $E_{targeted}, E_{existing} \subset E$, $\|V\| = N, \|E\| = M$)	
$V = \{v_i \mid 0 \leq i < N\}$, $V_{existing} \subset V$	The set of all *KO*s, including the existing ones
$E = \{e_{(i,j)} \mid \forall v_i, v_j \in V\}$	The set of directed edges among *KO*s
$V_{targeted} = \{v_w \mid v_w \in (V - V_{existing})\}$	The targeted *KO*s that one developer is currently working on
$E_{targeted} = \{e_{(w,u)} \mid v_w \in V_{targeted}$, and $v_u \in V_{existing}\}$	The directed edges among v_w and other existing v_u
Computation: correlation measurement via $R(V_{targeted}, E_{targeted})$	
$R^c(V_{targeted}, V_{existing}) =$ $\{R^c(w,u) \mid v_w \in V_{targeted}$, and $v_u \in V_{existing}\}$	Conceptual correlation between the targeted and the existing knowledge objects
$R^r(V_{targeted}, V_{existing}) =$ $\{R^r(w,u) \mid v_n \in V_{targeted}$, and $v_u \in V_{existing}\}$	Relational correlation between the targeted and the existing knowledge objects
Output: activity awareness to retrieve correlated *KO*s	
$top^{\alpha}(R(V_{targeted}, V_{existing}))$	Top α amount of correlated *KO*s

Given the directed-weighted Graph $G = (V, E)$, the vertices $v_i \in V$ denote the set of knowledge objects (*KO*) in the development repository, and the directed-weighted edges $e \in E$ among vertices denote the relations among *KO*s, and $V_{targeted} = \{v_w\}$ denote the set of new knowledge

objects, the target is to output an adaptive amount of correlated *KO*s from the existing ones $V_{existing}$ regarding to v_w. Table 4.1 enumerates the detailed parameters used for the graph model. We integrate the contextual correlations and relational correlation to build a common correlation among the knowledge objects, as described in following subsections.

4.2.1 Conceptual Correlation

The domain-concept allocation is different among the knowledge objects, since each correlation targets specific software development problems. The contextual correlation R^c measurement models the similarity (or distance) [52] between the symbolic description of two knowledge objects into a single numeric value. For example, during collaborative software development, there is either a shared dictionary of programming language or software engineering glossary among developers of each knowledge object respectively. Based on dictionary $\mathcal{T}$, each knowledge object is a mixture of linguistic symbols to extend and specify its domain-concept allocation, and such linguistic symbols are denoted as the terms ($\tau \in \mathcal{T}$) such as words or tags. Thus, each knowledge object is a document ($d_i \in D$, where $0 \leq i < N$) with mixed terms, and the terms used by the current knowledge objects construct the covered dictionary $\mathcal{T}_D \subset \mathcal{T}$, and $|\mathcal{T}_D| = L$.

The contextual correlation measure is simplified as terms-based similarity, and the major challenge is to extract and quantify the importance of those terms to *KO*s. In natural language and document processing, the *term frequency and inverse document frequency* ($tfidf$) is a numerical statistic which is intended to weight or reflect how important a term is to a document in a collection or corpus [50]; thus, it is used in Eq. (4.1) as a weighting factor to measure the significance of domain-concept. The *term frequency* (tf) is the count of a targeted term's occurrence in the selected documents, and is commonly used to proportionally represent the weight of a term in documents. Sometimes, the highly frequent terms will result in a large variance in the weight measurement to multiple terms. To reduce the variance for a better precision in measuring the weight of terms, we use the logarithmically scaled frequency (Eq. (4.2)), rather than using the raw frequency of terms directly. And the *inverse document frequency* (idf) is the specificity of a term that can be quantified as an inverse function of the number of documents in which it occurs (see Eq. (4.3)).

$$\textit{tfidf}_{(\tau,d)} = \textit{tf}_{(\tau,d)} \times \textit{idf}_{(\tau,D)}, \ \textit{where } d \in D \tag{4.1}$$

$$tf_{(\tau,d)} = \begin{cases} 1 + \log f_{\tau,d} & f_{\tau,d} > 0 \\ 0 & f_{\tau,d} = 0 \end{cases} \tag{4.2}$$

$$idf_{(\tau,D)} = \log \frac{|D|}{1 + |\{d \mid \tau \in d \text{ and } d \in D\}|} \tag{4.3}$$

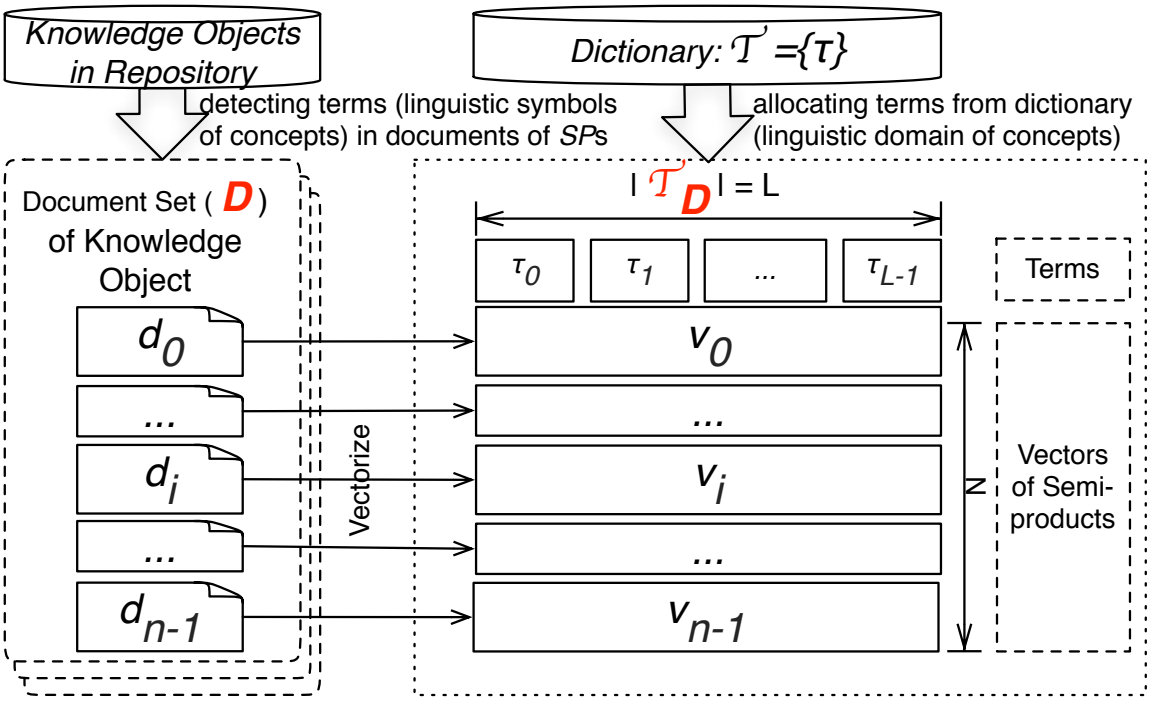


Figure 4.3. Probabilistic Representation of Domain Concept Allocation based on Terms (the linguistic symbols)

The *tfidf* value increases proportionally to the number of times a word appears in the document, and is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

Given the *tfidf* weighting factor, the knowledge objects could be normalised in a L-dimensions space model (see Figure 4.3), and each *KO* could then be formulated as a vector using Eq. (4.4), while the similarity measure can be achieved by computing any pair of *KO* vectors. There are various similarity measures outlined in previous works [52,53], which show that the performance of the Cosine similarity, Jaccard correlation, and Pearson's coefficient are very close and significantly better than the Euclidean distance measure. Therefore we choose the cosine similarity measure for

its lower computation costs, and the conceptual correlation using the terms-frequency is denoted as R^c_{tf} and outlined in Eq. (4.5).

$$\vec{v_i} = \langle \textit{tfidf}_{(\tau_0,d_i)}, \ldots, \textit{tfidf}_{(\tau_l,d_i)}, \ldots, \textit{tfidf}_{(\tau_{(L-1)},d_i)} \rangle, \quad \textit{where } d_i \in D,\ 1 \leq i \leq |D| = N \tag{4.4}$$

$$R^c_{tf}(i,j) = \textit{SIM}_{cos}(\vec{v_i}, \vec{v_j}) = \frac{\vec{v_i} \cdot \vec{v_j}}{|\vec{v_i}| \times |\vec{v_j}|} \tag{4.5}$$

As mentioned in Section 2.2.3, the contextual correlation may face a problem when there are insufficient contents of single knowledge pieces. It is also limited by the terms' dictionary of specific knowledge domains. Therefore, we further investigate the relational correlation (in Section 4.2.2) which is using the appended links according to developers' heuristic experience, and is also independent from knowledge domains.

4.2.2 Relational Correlation

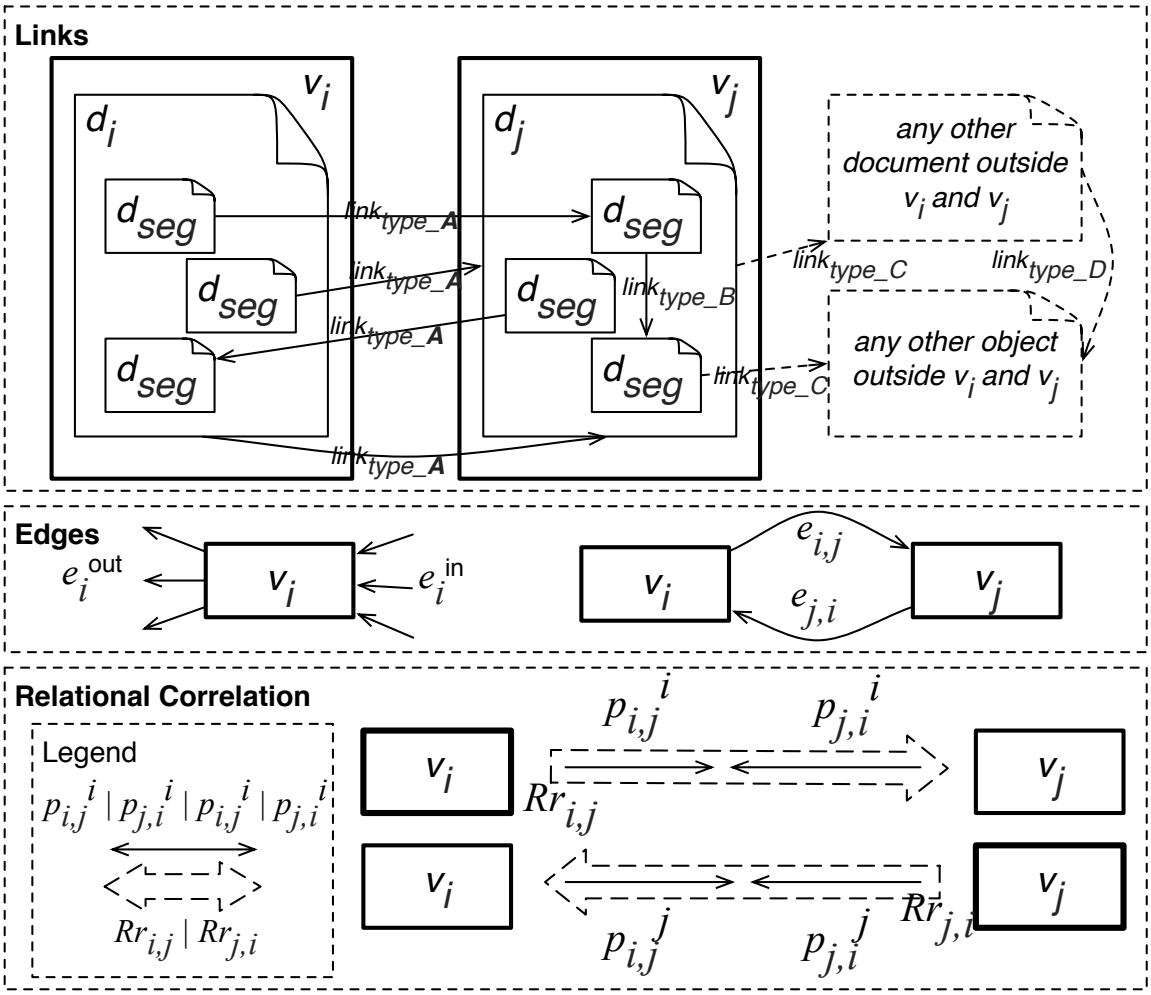


Figure 4.4. Model of *Links*, *Edges*, *Links-ratio* for Relational Correlation

Besides the conceptual correlation, the relational correlation is the probability that a developer may access other *KO* to learn about the achievements or experiences for his/her current developing knowledge objects. Since developers' travelling from one *KO* to another is mainly based on the relations (e.g. the URLs) among them; we assume that the rational correlation ($R^r(i,j)$ denotes the relation from v_i to v_j) is fit to the Bayesian network, which means that $R^r(i,j)$ is only affected by the relations of v_i and v_j neighbours. Figure 4.4 illustrates the relations among knowledge objects. By considering the knowledge objects (taking v_i and v_j for example) represented as a document (d_i and d_j), there are many document segments ($d_{seg} \in d_i$) included inside the knowledge objects.

Definition of Links There are 4 types of links that connect with the documents or document segments of the knowledge objects:

- Links of type A are the links between any two document segments that belong the two selected knowledge objects respectively: $link_{type\_A}(i,j) \in \{d_{seg_x} \rightarrow d_{seg_y} \mid \forall d_{seg_x} \in d_i \text{ and } \forall d_{seg_y} \in d_j\}$
- Links of type B are the links between any two document segments that both belong to one knowledge object: $link_{type\_C}(i,j) \in \{d_{seg_x} \rightarrow d_{seg_y} \mid \forall d_{seg_x} \in d_i \text{ and } \forall d_{seg_y} \notin d_j\}$
- Links of type C and D (as shown in Figure 4.4) are the links between any two document segments that at least one of the two does not belong to the two selected knowledge objects: $link_{type\_B}(i,j) \in \{d_{seg_x} \rightarrow d_{seg_y} \mid \exists d_{seg_x} \notin d_i \text{ or } \exists d_{seg_y} \notin d_j\}$.

The links of type A show the direct relations among any two given targeted knowledge objects (such as v_i and v_j), and the directed edge $e_{i,j} = \{link_{\text{type\_A}}(i,j)\}$ is represented as the set of links (type A only) between v_i and v_j.

Definition of Links-ratio As shown in the middle block of Figure 4.4, there are multiple edges connecting with one single knowledge object (taking v_i for example). We denote the $e_i^{in} = \{e_{j,i} \mid 0 \leq j < N\}$ as the set of all the edges directing into v_i, and also the $e_i^{out} = \{e_{i,j} \mid 0 \leq j < N\}$ as the set of all edges directing from v_i respectively. Thus, the links-ratio of edge $e_{i,j}$ over the outgoing edges of v_i is denoted $p_{i,j}^i = \frac{|e_{i,j}|}{|e_i^{out}|}$; and $p_{i,j}^i = 0$ when $|e_{i,j}| = 0$. Here in Eq. (4.6), we give the general expression for $p_{i,j}^i$, $p_{j,i}^i$, $p_{j,i}^j$, and $p_{i,j}^j$.

$$
\begin{aligned}
p_{i,j}^i &= \begin{cases} \frac{|e_{i,j}|}{|e_i^{out}|} & \text{if } |e_{i,j}| \neq 0 \\ 0 & \text{if } |e_{i,j}| = 0 \end{cases} \quad p_{j,i}^i = \begin{cases} \frac{|e_{j,i}|}{|e_i^{in}|} & \text{if } |e_{j,i}| \neq 0 \\ 0 & \text{if } |e_{j,i}| = 0 \end{cases} \\
p_{j,i}^j &= \begin{cases} \frac{|e_{j,i}|}{|e_j^{out}|} & \text{if } |e_{j,i}| \neq 0 \\ 0 & \text{if } |e_{j,i}| = 0 \end{cases} \quad p_{i,j}^j = \begin{cases} \frac{|e_{i,j}|}{|e_j^{in}|} & \text{if } |e_{i,j}| \neq 0 \\ 0 & \text{if } |e_{i,j}| = 0 \end{cases} \\
&\text{where } 0 \leq i,j < N \text{ AND } i \neq j
\end{aligned}
\tag{4.6}
$$

Relational Correlation using Neighbouring Links-ratio Based on the current observation to the knowledge objects, the URLs or URIs based links-ratio is the major factor of interrelationship,

and other factors like the undetectable references are the minorities. Since the edges here is represented by the links-ration, we assume that the in and out edges of v_i's are independent with those of v_j's except the their joint edges ($e_{i,j}$ and $e_{j,i}$). Therefore relational the non-correlation from v_i to v_j can be expressed as $(1 - p^i_{i,j}) \times (1 - p^i_{j,i})$. And then the relational correlation from v_i to v_j using the neighbouring links-ratio is $R^r_{ln}(i,j) = 1 - (1 - p^i_{i,j}) \times (1 - p^i_{j,i})$. The foot "$ln$" means using the neighbouring links-ratio, with "n" denoting neighbouring and "l" denoting links-ratio. And a general expression is given in Eq. (4.7).

$$R^r_{ln}(i,j) = \begin{cases} p^i_{i,j} + p^i_{j,i} - p^i_{i,j} \times p^i_{j,i} & \text{where } i \neq j \\ 1.0 & \text{where } i = j \end{cases} \tag{4.7}$$

The expression in Eq. (4.7) mainly measures the pairwise relational correlation between neighbouring vertices, and we name such a measurement the neighbouring relational correlation. The measure relational correlation between the pairwise vertices which are not neighbours to each other is further given in Section 4.2.3, and we name it the chained relational correlation.

4.2.3 Integrated Correlation

In order to implement knowledge correlating for development support, it is a major challenge to improve the significance of the correlation measure to real-time development activities. More specifically, there are two technical issues: first, to collect development activity data more comprehensively; and second, to improve the significance of correlation measurements with computational efficiency. In this section, we develop the seamless repository as an integrated development environment (IDE) to enable and record types of important development activities in Section 4.3.1. Then, we implement the TFCLR approach in Section 4.3.2 to improve the significance and correlation measure. The terms-frequency based method retrieves significant linguistic symbols as terms and specifies the linguistic domain of concepts that the developers focus on in specific engineering process. Also, the (neighbouring/chained) links-ratio based method computes the maximum product of the neighbouring relational correlation over the knowledge graph using an extended *Floyd-Warshall* algorithm.

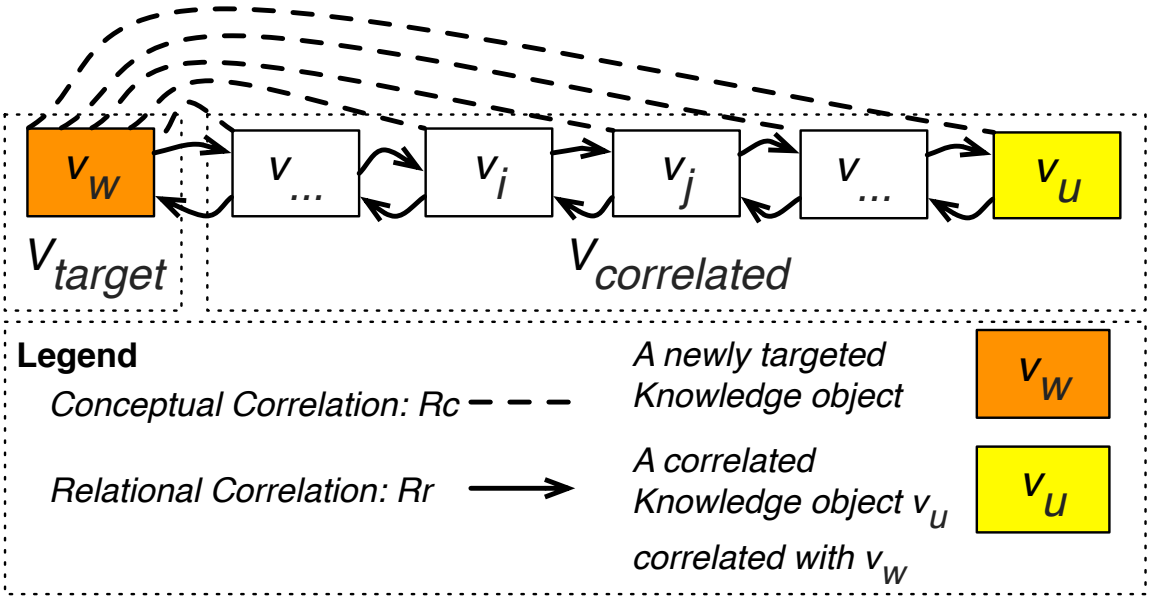


Figure 4.5. Integrated Contextual and Relational Correlation

Given the contextual (in Section 4.2.1) and relational (in Section 4.2.2) correlation, the use case of knowledge correlating for development support is modelled as finding a set of the most correlated knowledge objects regarding to the targeted knowledge objects that developers are working on. As shown in Figure 4.5, $v_w \in V_{targeted}$ is one of the new knowledge objects, and $v_u \in V_{existing}$ is one of the correlated knowledge objects regarding to v_w. However, there might be no directed edge ($e_{w,u} = \emptyset$) between v_w and v_u, but based on Bayesian theory the relational correlation is transitive along with the chained edges in Graph G.

Relational Correlation using Chained Links-ratio The relational correlation transferring through the chained edges from v_w to v_u is defined as the relational correlation using the chained links-ration (denoted as R^r_{lc}, see Eq. (4.8)), which equals the maximum joint probability of the neighbouring links-ratio based relational correlations by passing through vertices in a single continuous chain from v_w to v_u. Such a chain is known as the critical chain from v_w to v_u. The program implementation for chained relational correlation is further given in Section 4.3.2.3.

$$R^r_{lc}(w,u) = \max \prod_{e_{i,j}\neq\emptyset, i=w}^{j=u} R^r_{ln}(i,j) \tag{4.8}$$

Integrated Correlation Since developers' development activities of inputting the contents or appending the links can be considered as independent events, here we assume that the contextual and relational correlation are also independent with each other. An integrated correlation is given in Eq. (4.9) by incorporating the contextual and relational correlations.

$$\begin{gathered} R(w,u) = 1 - \Big(1 - R^c(w,u)\Big) \times \Big(1 - R^r(w,u)\Big), \\ R^c \text{ can be } R^c_{tf}, \text{ and } R^r(w,u) \text{ can be either } R^r_{ln} \text{ or } R^r_{lc}, \\ \text{where } v_w \in V_{targeted} \text{ and } v_u \in V_{existing} \end{gathered} \tag{4.9}$$

The integrated correlation provides a more comprehensive measure to compare targeted knowledge objects with the existing ones in repository. The awareness response (see Eq. (4.10)) for development support is defined as adaptive amount (equals α, developers can freely defined for a better user experience) of most correlated knowledge objects.

$$\begin{gathered} A = \{v_r \mid v_r \in top^{\alpha}_{R(w,u)}, v_w \in V_{\textit{targeted}}, v_u \in V_{\textit{existing}}\}, \\ \alpha \text{ is the adaptive amount of the most correlated } \textit{KO}\text{s} \end{gathered} \tag{4.10}$$

The proposed integrated contextual and relational correlation reorganize the achievements and knowledge as the knowledge objects in the development repository, and helps to retrieve the correlated knowledge objects to accelerate new development regarding to developers' undergoing activities.

4.3 Implementation

In order to implement knowledge correlating for development support, it is a major challenge to improve the significance of the correlation measure to real-time development activities. More specifically, there are two technical issues: first, to collect development activity data more comprehensively; and second, to improve the significance of correlation measurements with computational efficiency. In this section, we develop the seamless repository as an integrated development environment (IDE) to enable and record types of important development activities in Section 4.3.1. Then, we implement the TFCLR approach in Section 4.3.2 to improve the significance and correlation

measure. The terms-frequency based method retrieves significant linguistic symbols as terms and specifies the linguistic domain of concepts that the developers focus on in specific engineering process. Also, the (neighbouring/chained) links-ratio based method computes the maximum product of the neighbouring relational correlation over the knowledge graph using an extended *Floyd-Warshall* algorithm.

4.3.1 Seamless Repository for Knowledge Correlating

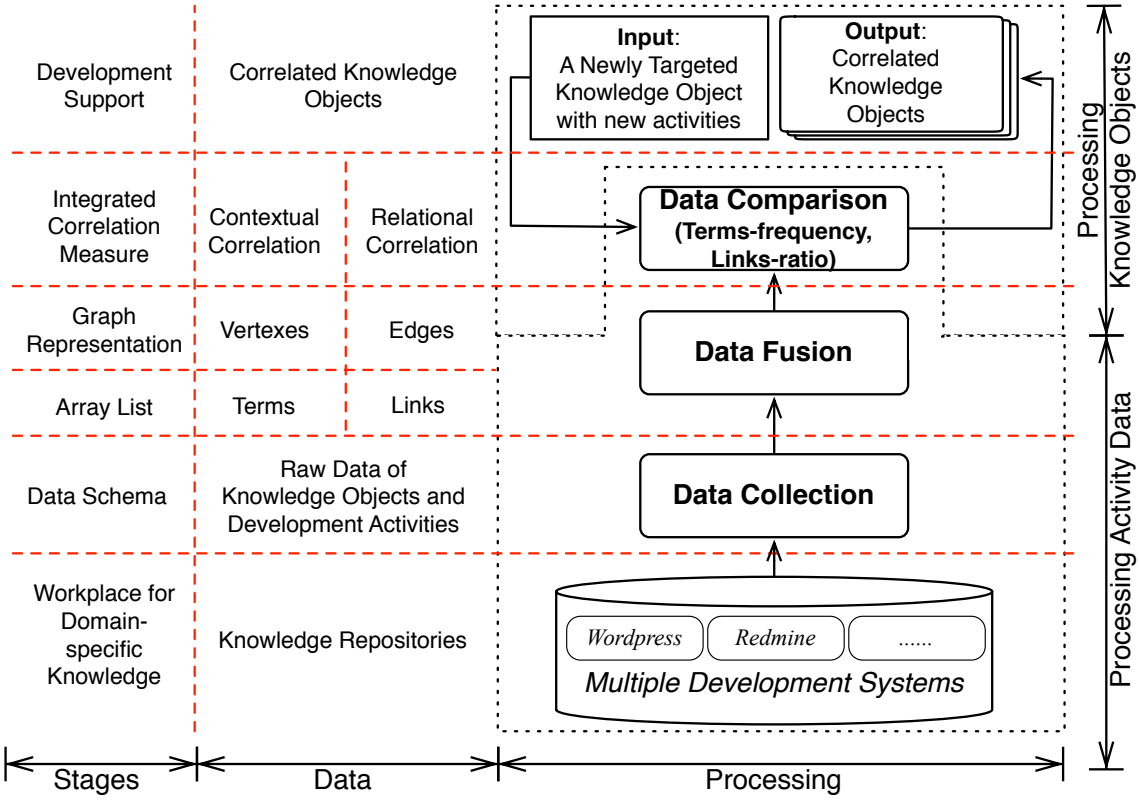


Figure 4.6. Collaborative Workflow Awareness for Development Support based on Seamless Repository

In order to support more types of development activities and record the data more comprehensively, the knowledge correlating is designed over the top of development environment, and require 3 major data processing components.

- The *data collection* component collects the raw data of development activities from the seamless development repository, and store the raw data into database for further analysis.
- The *data fusion* component re-constructs the development activity data into a graph based

representation of semi-products that the activities target by extracting the terms and links.

- The *data comparison* component computes the integrated correlation among semi-products, and returns the most correlated semi-products for the new semi-products.

The seamless repository provides developers an integrated development environment, and developers also contribute to enrich the knowledge and achievements. We denote such achievements and knowledge as knowledge objects (*KO*s), such as projects in Redmine, pages in MediaWiki, projects in GitLab, and posts in Wordpress. The accumulated knowledge objects construct the development repository. As the scale of the repository grows, developers will not be able to have full experience to all existing knowledge objects in repository. Thus, development support is to quantify the correlations among knowledge objects and find suitable ones to support current development.

Table 4.2. Growing Scale of Local Seamless Repository (Until Oct. 2015)

Repositories	Scale			
	Users	Groups	Knowledge Objects	Activities (Logs)
Redmine (Version-2.4)	58	6	118	693
GitLab (Version-7.0)	41	14	82	983
MediaWiki (Version-1.23)	—	—	213	970
Wordpress (Version-4.0)	32	4	47	603

Therefore, we implement a seamless repository (see Figure 4.6) as an IDE to provide functional supports for various types of software engineering development activities. For example, it integrates Internet-based teamwork tools, such as *Redmine*, *GitLab*, *MediaWiki*, and *Wordpress*, as sub repositories to support the critical development activities, such as issue tracking, distributed/centralized revision control, and content management. The seamless repository also records development activities in a structured format and stores data in rational database. Table 4.2 shows the scale of knowledge objects and development activities in a seamless repository that has grown over the past 12 months until Oct. 2015. This data set was contributed by a research team, mainly focused on the domain concept of information and communications technology (ICT), and was written in English.

4.3.1.1 Implementation of Data Collection Component

We first deploy the Internet-based teamwork tools on LAMP (Linux Apache MySQL PHP) and NodeJS server environment, and then develop the seamless repository to integrate those sub

repositories through an ODBC (Open Database Connectivity) API. Developers collaborate through the interface of those sub repositories as shown in Figure 2.3. The development activity data is collected in a structured format defined by the data schema of the seamless repository (see the Entity-Relationship model as given in Figure 4.7). It consists of three major entities in development process, including *users* (the developers), *activities*, and *knowledge objects*. The developers own the memberships in knowledge objects, and their development result in activities, devoting to the development progress of the knowledge objects. Therefore, the relationship tables of *development*, *contribution*, and *membership* relate the three major entities. Furthermore, the *knowledge objects* table is further related with the data schema of sub repositories, including tables of projects in *Redmine*, posts in *Wordpress*, pages in *MediaWiki*, and projects in *GitLab*.

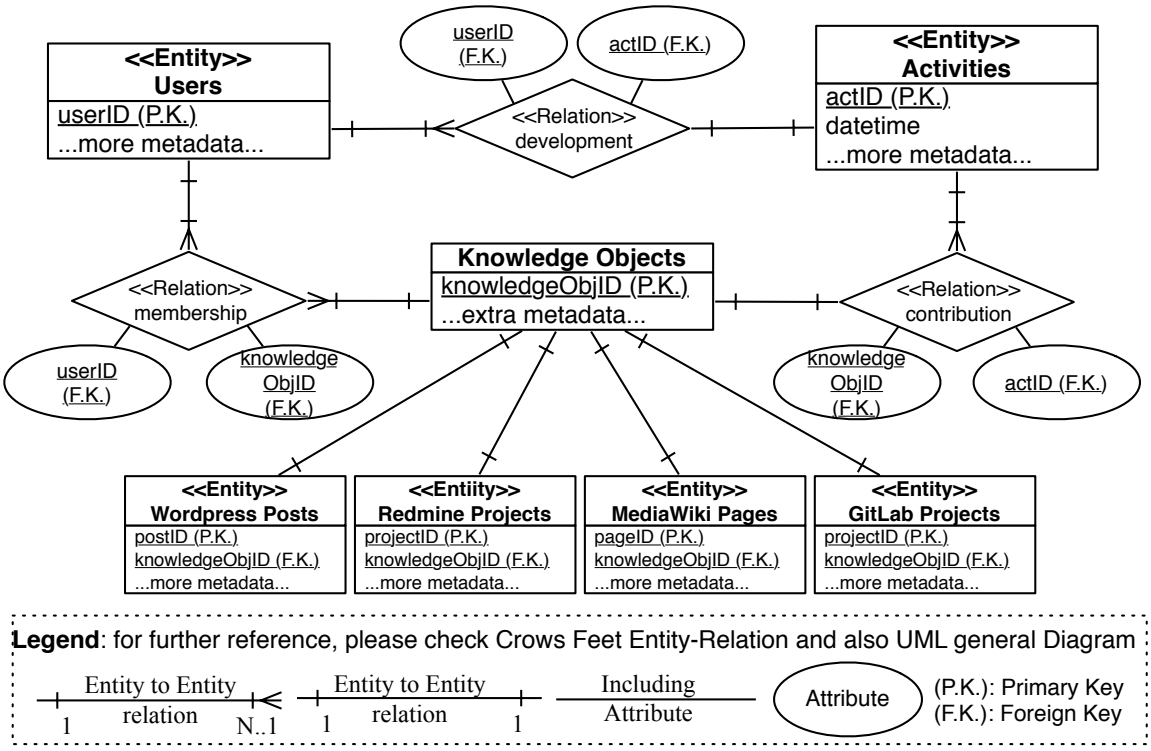


Figure 4.7. Data Schema for Data Fusion

Since the scale of activity data (see Table 4.2), is much larger than that of others, it is more meaningful to investigate the activity data. On the other hand, the graph of knowledge objects, as being investigated by the correlation measure in Section 4.2, construct the knowledge objects in the seamless repository, such as technical know-how and project achievements. Therefore, the data collection task herein is to collect to raw data of development activities from different sub

repositories so as to retrieve the conceptual information of knowledge objects and the referential relations among knowledge objects.

4.3.1.2 Implementation of Data Fusion Component

This process reconstructs a graph-based representation of the achievements and knowledge in the seamless repository, in which the vertices denote the knowledge objects, and weighted-directed edges denote the referential relationships among knowledge objects. According to Section 4.2, the correlation among the vertices is measured by the conceptual and relational correlation, which is based on the statistical measure respectively to terms and links information from the development activity data.

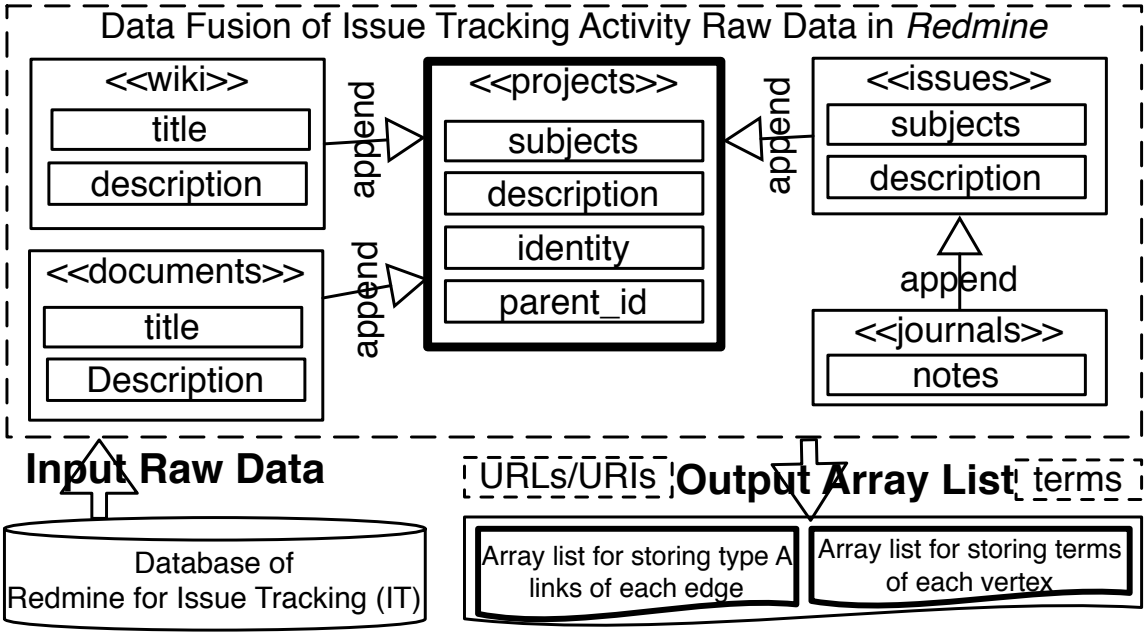


Figure 4.8. Sample of Data Fusion for Redmine

Figure 4.7 shows a high-level ER model to integrate the different data schema of each sub repositories. A major effort of the data fusion process is extracting the terms and links information. Here we take an example of data fusion to the development activity data in Redmine (as shown in Figure 4.8):

- *Initialization*: enumerate the *projects* in Redmine as vertices and the URIs (uniform resource identifiers) of the vertices, and initialize the computation memory by creating an array list for

storing the terms of each vertex and another array list for storing the links of each edge.

- *Targets selection*: investigate the data schema in Redmine, and select the attributes that record the raw data of development activities. In the previous example example, the body information of issues tracking activities is recorded in attribute *notes* in table *journals*, and the introduction of issues is recorded in attributes *subjects* and *description* in table *issues*.

- *Terms extraction*: parse the raw data in those selected attributes, then extract and append the all terms to vertices that them belong to according to Figure 4.8. For example, the terms from *issues*, *journals*, *wiki*, and *documents* are exported to the array of each vertex.

- *Links extraction*: parse the raw data to detect the URLs (uniform resource links) or URIs that are type A links (see Section 4.2.2) in the whole seamless repository not only Redmine. And store the links' start and end identifiers to the array of each edge.

- *Parsing the whole repository*: the aforementioned steps are also processed on all the other sub repositories besides *Redmine* to build the array list of terms and links.

A graphical representation of the knowledge and achievements in seamless repository is built, in which the vertices are represented as the array list of terms, and edges are represented as the array list of links.

4.3.1.3 Implementation of Data Comparison

This process computes the correlation among vertices in the aforementioned graph using the integrated conceptual and conceptual correlation measure. Given a vertex that stands for the knowledge objects that developer currently working on, the data comparison process returns the most correlated vertices for the selected vertex. The major challenge in this step is to improve the significance and computational efficiency of the correlation measure. The terms-frequency and links-ratio approaches are proposed in Section 4.3.2 to power the measure of conceptual and relational correlation respectively.

4.3.2 Correlation Measure for Knowledge Correlating

In order to measure the correlation among knowledge objects for development support, we conceive the approach of using Terms-frequency and (Neighbouring/Chained) Links-ratio to incorporate the conceptual and relational correlation. A group of correlation measures have been implemented as follows using python programming.

4.3.2.1 Conceptual Correlation based on Terms-frequency

Algorithm 2 Computation of *Conceptual Correlation* based on terms-frequency

```
1: from sklearn.feature_extraction.text import CountVectorizer          ▷ Vector operation Lib.
2: from sklearn.feature_extraction.text import TfidfTransformer         ▷ tfidf operation Lib.
3: from scipy import spatial          ▷ Spatial Lib. for computing cosine distance among KOs
4: let corpus be the list of terms sets for KOs
5: Let tfidf_array be the tfidf vectors for KOs
6: procedure COMPUTETFIDF(corpus)                          ▷ Compute tfidf vectors for KOs
7:     vectorizer = CountVectorizer(min_df=1);
8:     tf = vectorizer.fit_transform(corpus);                          ▷ Term-frequency for KOs
9:     transformer = TfidfTransformer();
10:    return transformer.fit_transform(tf).toarray();
11: end procedure
12: procedure COMPUTECONCEPTUALCORRELATION(tfidf_array)
13:    r_tf = [];                          ▷ Conceptual Correlation using terms-frequency
14:    for i in range(0, len(tfidf_array)) do
15:        r_tf.append([]);
16:        for j in range(0, len(tfidf_array)) do
17:            r_tf[i].append(1 - spatial.distance.cosine(tfidf_array[i], tfidf_array[j]))
18:        end for
19:    end for
20:    return r_tf
21: end procedure
22: r_conceptual(tf) = ComputeConceptualCorrelation(ComputeTFIDF(corpus))
```

As mentioned in Section 4.2.1, the terms are those linguistic symbols denoting the domain concepts in knowledge objects. The terms-frequency based *tfidf* measure statistically represents each *KO*'s domain concept coverage as a vector in an *L*-dimensional space framework, where *L* equals the count of identical terms. The computation for pairwise conceptual correlation is given Alg. 2, in which each *KO* is considered as document, all documents constitute the corpus. Several python libraries, such *sklearn*, are used for feature extraction on a large text corpus. The *tfidf* term weighting is used for purposes such as ignoring the terms which are frequent but meaningless.

The function *ComputeTFIDF* in Alg. 2 firstly tokenizes the terms and enumerates each term's occurrence in every single document, outputs the the frequency of all terms in all documents into *tf*, an L-by-L array list, and then returns the *tfidf* array list through *TfidfTransformer*. Finally, the *ComputeConceptualCorrelation* function returns the terms-frequency based all-pairs conceptual correlation $r_{conceptual(tf)}$ among *KO*s using Cosine similarity. Here, $r_{conceptual(tf)}$ is symmetric (or undirected) due to Cosine similarity.

4.3.2.2 Relational Correlation based on Neighbouring Links-ratio

The relational correlation is modelled based on links of type A as mentioned in Section 4.2.2. The neighbouring relational correlation $r_{relational(ln)}$ is an asymmetric correlation measure between pairwise neighbouring vertices, using the links-ratio of both the *in* and *out* directed links as shown in Eq. (4.6) and (4.7). The links extraction process in data fusion (see Section 4.3.1.2) outputs the array list of $links_{out}$. The program to get the array list of $links_{in}$ in outlined in Alg. 3.

Algorithm 3 Enumerating Links of *In* Direction

```
Let links_out be the enumeration of links type A out from KOs
procedure EnumerateLinksOfInDirection(links_out)
    links_in=[]                          ▷ Enumeration of links type A out from KOs
    for i in range(0, len(links_out)) do
        links_in.append([])
    end for
    for i in range(0, len(links_out)) do
        for j in range(0, len(links_out[i])) do
            links_in[links_out[i][j]].append(i)
        end for
    end for
    return links_in
end procedure
links_in = EnumerateLinksOfInDirection(links_in)
```

Algorithm 4 Computation of *Relational Correlation* based on Neighbouring Links-ratio

Let $links_{out}$ and $links_{in}$ be the enumeration of links type A *out* from and *in* to *KO*s.
procedure COMPUTENEIGHBOURINGRELATIONALCORRELATION($links_{out}$, $links_{in}$)
    $r_{ln} = []$ ▷ Pairwise relational correlation using neighbouring links-ratio
    $weight_{pairwise}$ =[] ▷ Array list of weight of pairwise directed edges among *KO*s
    for i in range(0, len($links_{out}$)) **do** ▷ Enumerate $weight_{pairwise}$
        $weight_{pairwise}$.append([])
        r_{ln}.append([])
        for j in range(0, len($links_{in}$)) **do**
            $weight_{pairwise}[i]$.append([0.0])
            for k in range(0, len($links_{out}[i]$)) **do**
                if $links_{out}[i][k] == j$ **then**
                    $weight_{pairwise}$[i][j] = $weight_{pairwise}[i][j]$ + 1.0
                end if
            end for
            $r_{ln}[i]$.append([0.0])
            if i == j **then**
                $r_{ln}[i][j] = 1.0$
            end if
        end for
    end for
    for i in range(0, len($links_{out}$)) **do**
        r_{ln}.append([])
        for j in range(i+1, len($links_{in}$)) **do**
            $p^i_{i,j}$=$p^i_{j,i}$=$p^j_{i,j}$=$p^j_{j,i}$=0
            if $weight_{pairwise}[i][j] > 0$ **then**
                $p^i_{i,j} = weight_{pairwise}[i][j]$/len($links_{out}[i]$)
                $p^j_{i,j} = weight_{pairwise}[i][j]$/len($links_{in}[j]$)
            end if
            if $weight_{pairwise}[j][i] > 0$ **then**
                $p^i_{j,i} = weight_{pairwise}[j][i]$/len($links_{in}[i]$)
                $p^j_{j,i} = weight_{pairwise}[j][i]$/len($links_{out}[j]$)
            end if
            $r_{ln}[i][j] = p^i_{i,j} + p^i_{j,i} - p^i_{i,j} \times p^i_{j,i}$
            $r_{ln}[j][i] = p^j_{j,i} + p^j_{i,j} - p^j_{j,i} \times p^j_{i,j}$
        end for
    end for
    return r_{ln}
end procedure
$r_{relational(ln)}$ = ComputeNeighbouringRelationalCorrelation($links_{out}$, $links_{in}$)

Alg. 4 gives the function *ComputeNeighbouringRelationalCorrelation* to take the $links_{out}$ and $links_{in}$ as input, and the $r_{relational(ln)}$ as output. It first transforms the all-pair *in* and *out* directed links into all-pair directed and weighted edges, and stores that information in the array list $weight_{pairwise}$;

and the $weight_{pairwise}$ of non-neighbouring pairs equals to 0 by default. Then, it computes the links-ratio as formulated in Eq. (4.6), and calculates the $r_{relational(ln)}$ as formulated in Eq. (4.7). The $r_{relational(ln)}$ of non-neighbouring pairs equals to 0 by default.

4.3.2.3 Relational Correlation based on Chained Links-ratio

The chained relational correlation $r_{relational(lc)}$ is an extension of $r_{relational(ln)}$. According to Eq. (4.8), each $r_{relational(lc)}$ is formulated as the maximum joint probability of $r_{relational(ln)}$ along with the chained edges rather than a single edge of a neighbouring pair. It is similar with the all-pair shorted path problem, and function *ComputeChainedRelationalCorrelation* implements a variation of the *Floyed-Warshall* algorithm for calculating the maximum joint probability over the chains. A critical chain is the set of transitive edges which achieve the maximum joint probability of of $r_{relational}$ between any given pairwise vertices. The pseudocode for relation correlation measure based on chained links-ratio is given in Alg. 5.

Algorithm 5 Computation of *Relational Correlation* based on Chained Links-ratio

```
Let r_relational(ln) be all-pair relational correlation using neighbouring links-ratio
procedure COMPUTECHAINEDRELATIONALCORRELATION(r_relational(ln))
    r_lc = []                          ▷ All-pair relational correlation using chained links-ratio
    for i in range(0, len(r_relational(ln))) do
        r_lc.append([])
        for j in range(0, len(r_relational(ln))) do
            if r_relational(ln)[i][j] > 0 then
                r_lc[i].append(r_relational(ln)[i][j])
            else
                r_chained[i].append(0.0)
            end if
        end for
    end for
    for k in range(0, len(r_relational(ln))) do
        for i in range(0, len(r_relational(ln))) do
            for j in range(0, len(r_relational(ln))) do
                if r_chained[i][k] × r_chained[k][j] > r_chained[i][j] then
                    r_chained[i][j] = r_chained[i][k] × r_chained[k][j]
                end if
            end for
        end for
    end for
    return r_chained
end procedure
r_relational(lc) = ComputeChainedRelationalCorrelation(r_relational(ln))
```

4.3.2.4 Integrated Conceptual and Relational Correlation

According to Eq. (4.9), the integrated correlation incorporates the conceptual and relational correlation. There are two variations of integrated correlation. Function *ComputeIntegratedCorrelation* take $r_{conceptual(tf)}$, and $r_{relational(ln)}$ or $r_{relational(lc)}$ as input, and the $r_{integrated(tf,ln)}$ or $r_{integrated(tf,lc)}$ as output. The pseudocode for implementing the integrated conceptual and relational correlation measure is given in Alg. 6.

Algorithm 6 Computation of *Integrated Conceptual and Relational Correlation*

Let $r_{conceptual(tf)}$ be the all-pair conceptual correlation using terms-frequency
Let $r_{relational}$ be the all-pair relational correlation
Let $r_{relational(ln)}$ be all-pair relational correlation using neighbouring links-ratio
Let $r_{relational(lc)}$ be all-pair relational correlation using chained links-ratio
procedure COMPUTEINTEGRATEDCORRELATION($r_{conceptual}$, $r_{relational}$)
    $r_{integrated}$ = [] ▷ integrated correlation
    for i in range(0, len($r_{conceptual}$)) **do**
        $r_{integrated}$.append([])
        for j in range(0, len($r_{conceptual}$)) **do**
            $r_{integrated}[i]$.append($1 - (1 - r_{conceptual}[i][j]) * (1 - r_{relational}[i][j])$)
        end for
    end for
    return $r_{integrated}$ ▷ return the integrated conceptual and relational correlation
end procedure
$r_{integrated(tf,ln)}$ = ComputeIntegratedCorrelation($r_{conceptual(tf)}$, $r_{relational(ln)}$) ▷ Integrated correlation based on terms-frequency and neighbouring links-ratio
$r_{integrated(tf,lc)}$ = ComputeIntegratedCorrelation($r_{conceptual(tf)}$, $r_{relational(lc)}$) ▷ Integrated correlation based on terms-frequency and chained links-ratio

Chapter 5

Case Study

5.1 Seamless Integration

We have built a seamless collaborative workflow repository for pervasive teamwork, which is currently dedicated for innovative projects in information technology (IT) domains. The current achievement through a one-year practice in an enclosed research group is listed in Table 5.1. For example, Redmine has achieved 118 research and development projects which are contributed by 58 participants and 6 groups. In said projects, participants generated 693 activities including adding issues, solving bugs, and adding system specifications. Similarly there are 82 projects contributed by 41 participants in the Gitlab, 983 activities including participants' Git commit to the update of source code. Furthermore there are 213 pages and 970 activities in Mediawiki, 50 instances of OS image which host the IDE of participants' system in VirtualBox repository. In the integrated collaborative repository, the BLE AltBeacon based proximity tracking app tracks the onsite collaboration in two locations: one is a laboratory office and the other is a seminar room. Furthermore, there are eight group mailing lists in Postfix repository, and 15 channels with 20 users in Slack.

Table 5.1. Achievement in One-year Practice of Using the Seamless Teamwork Repository in an Enclosed Research Group

Integrated Repositories	Repository Scale			
	Users	Groups	Archives	Activities (Logs)
Redmine (Version-2.4)	58	6	118 Projects	693
GitLab (Version-7.0)	41	14	82 Projects	983
MediaWiki (Version-1.23)	—	—	213 Pages	970
Wordpress (Version-4.0)	32	4	47 Posts	603
VSFTP (Version-3.0) / SAMBA (Version-4.2)	–	–	4 Storage	2 TB (Aprox.)
VirtualBox	–	–	50 Instances	1 TB (Aprox.)
BLE AltBeacon	5	–	2 Locations	16478
Postfix (Version-4.0) / SquirrelMail (Version-1.2)	—	8 Mailing list	—	—
Slack	20	—	15 Channels	937

5.1.1 Services in Layer-1: Support of Sharing

5.1.1.1 Demonstration of Portfolio Service

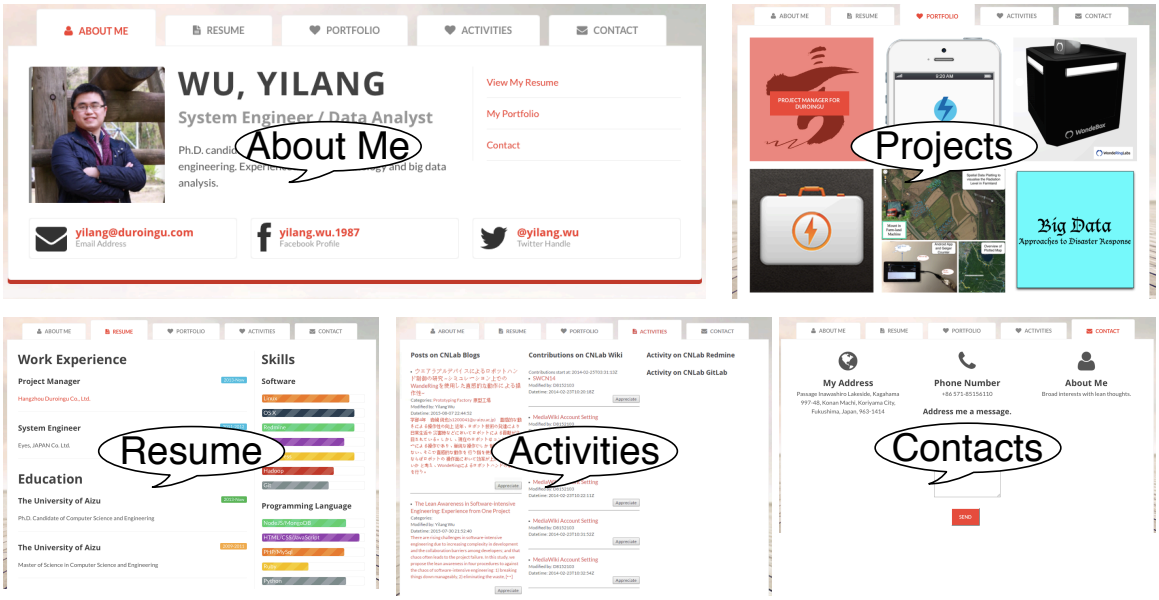


Figure 5.1. Demonstration of Portfolio Service

The portfolio service is a web page service of automatically collecting each user's portfolio information from the support platform to make it easier for users to know each others' background and profile information. The system design and implementation has been illustrated in Figure 3.8.

An open source one page responsive HTML resume template [54] is used for web page representation. Figure 5.1 shows a set of portfolio demonstration pages: the “About Me” page shows the basic profile information; the “Resume” page enumerates the items of biography information and levelled skills; the “Activities” page shows the head information of collaborative workflow activities in the platform.

5.1.1.2 Demonstration of Workflow Templates Service

Figure 5.2. Demonstration of Workflow Template Service: the Result of Graduation Thesis Backup

We take laboratory students’ graduation thesis backup as a case study to formalize the workflow process. Students are guided to use the services in the support platform to backup their thesis related data in specific sub repositories. Their thesis related files are achieved in SAMBA server, development issues and weekly reports are submitted to Redmine, development source codes are pushed to GitLab, and their demonstration video and images are published to Wordpress. And finally, a web-based digital book of introducing graduation thesis projects is generated (see Figure 5.2), viewers can flip the pages of the digital book, and also click the link to view the demonstration videos, slides, thesis, development issues, and also source codes.

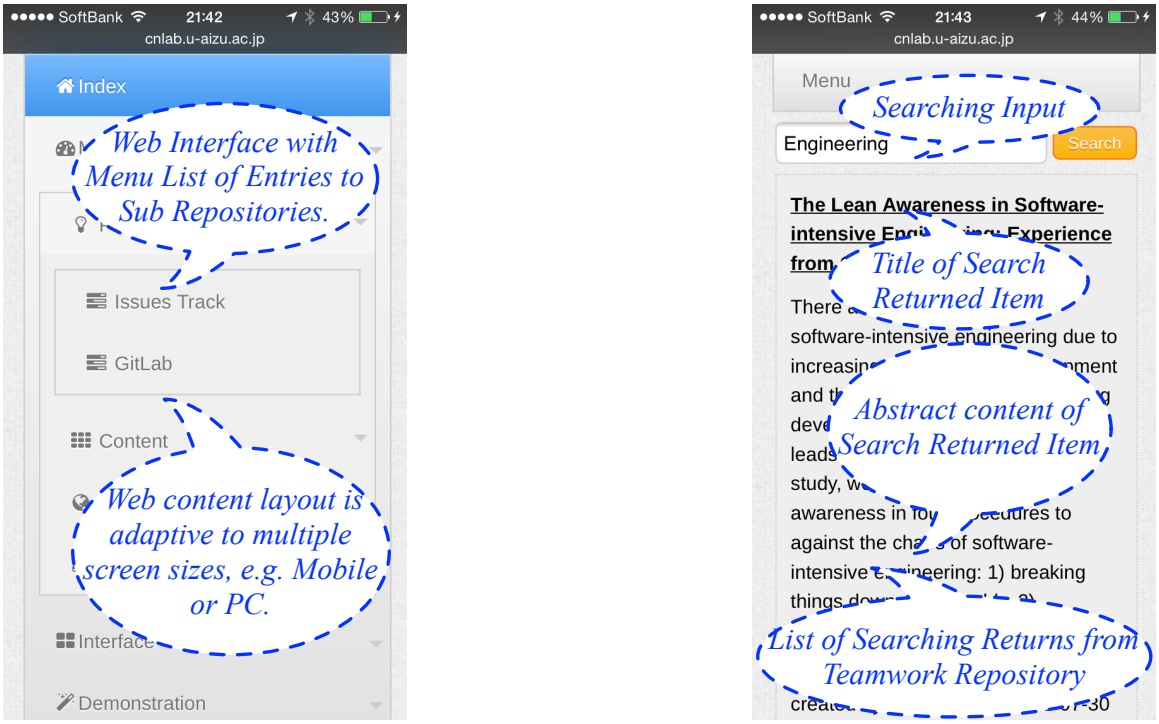
By using the workflow templates for graduation thesis backup, it is easier for the graduating students to finish the backup process. Meanwhile, rich and valuable information can be better

shared to the new students in laboratory.

5.1.2 Services in Layer-2: Support of Interconnection

The layer-2 integrates the *CW* sub repositories by connecting the data and functionality. We develop an integrated web portal in Section 5.1.2.1 as a single entry to access and search the data and functionality in sub repositories. And we also implement the notification component in Section 5.1.2.2 in purpose of delivering the news faster and connecting participants closer.

5.1.2.1 Demonstration of Web Portal Service



(a) Web Portal of Unified Entry

(b) Web Portal of Search Entry

Figure 5.3. The Web Portal of Seamless Repository for Pervasive Teamwork

Figure 5.3 illustrates the user interface of the web portal that integrates the sub repositories. The Web portal follows the Responsive Web Design [55], making it suitable to work on every device and every screen size. And it also improves the content visibility for participants in pervasive teamwork. For example, the categorized and nested navigation (see Figure 5.3(a)) makes the sub repositories easily accessible. And the integrated search function (see Figure 5.3(b)), utilizes the HTTP REST (Representational State Transfer) API and ODBC (Open Database Connectivity) protocol to connect

the local repositories more comprehensively. Furthermore, web portal also enables the search to multiple sub repositories simultaneously with a single input, making the whole repository more transparent for participants.

The Web portal can quickly search the unknown keywords to get the background knowledge for the communicators, so that the users do not have to stop their communication by unknown keywords. And they can also learn the knowledge behind the keyword again by searching through the Web portal after the communication. Such a web portal service is important since the commonly used search engine cannot search local repositories.

5.1.2.2 Demonstration of Notification Service

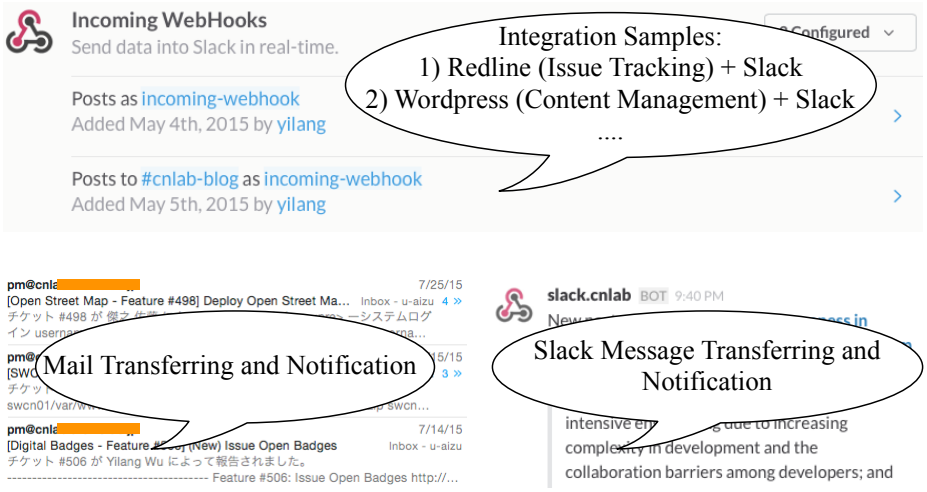


Figure 5.4. Mail Transferring and Messaging based on Postfix and Slack

Figure 5.4 shows the deployed communication service of mail transferring and notification messaging based on Postfix and Slack. Once there is any new *CW* activities comes in the *CW* sub repositories, the participants will be notified either through the mail transferring or Slack instant messaging. Participants can customize their notification settings so that to be notified only with the *CW* activities that they concern.

The notification service which is based on the mailing transferring and instant messaging eases the communication and enhances the connectivity among participants. And the smoother communication also eases the workplace conflicts.

5.1.3 Services in Layer-3: Support of Visualization

The layer-3 brings participants deeper and more attractive insights to the collaborative workflow, so that they will achieve better common sense in teamwork. And we implemented the real-time visualization (in Seciton 5.1.3.1) of the teamwork involvement to bring participants attractive insights to the real-time status of the teamwork. And we also implemented the heat-map (Section 5.1.3.2) to visualize the temporal hotspot of the collaborative workflow in teamwork.

5.1.3.1 Demonstration of Teamwork Involvement Animation Service

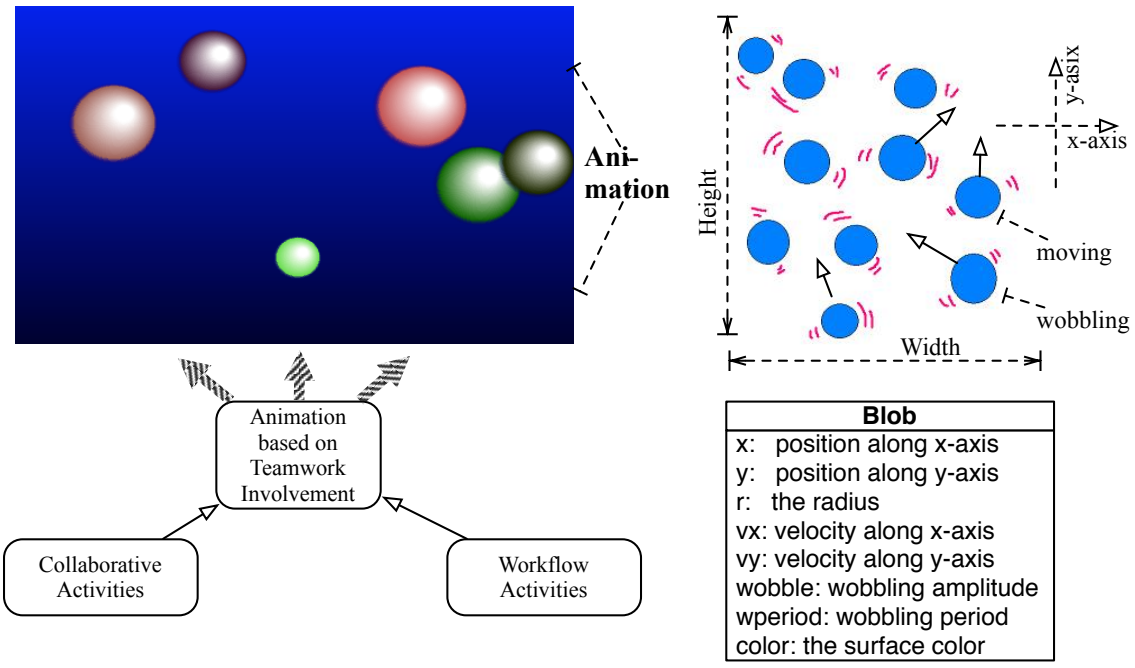


Figure 5.5. VR Simulation for Immersive Collaboration Experience

To bring users immersive experience in collaboration, we try to use the VR technology to visualize the real-time activities in teamwork. The movement intensiveness of blobs represents participants' involvement. The pseudo-code is given in Algorithm 1. A Web based animation visualizes the teamwork intensiveness through animating the moving and vibration and the scalable size of the blobs (see, Figure 5.5). The color of the blobs is used to identify different participants.

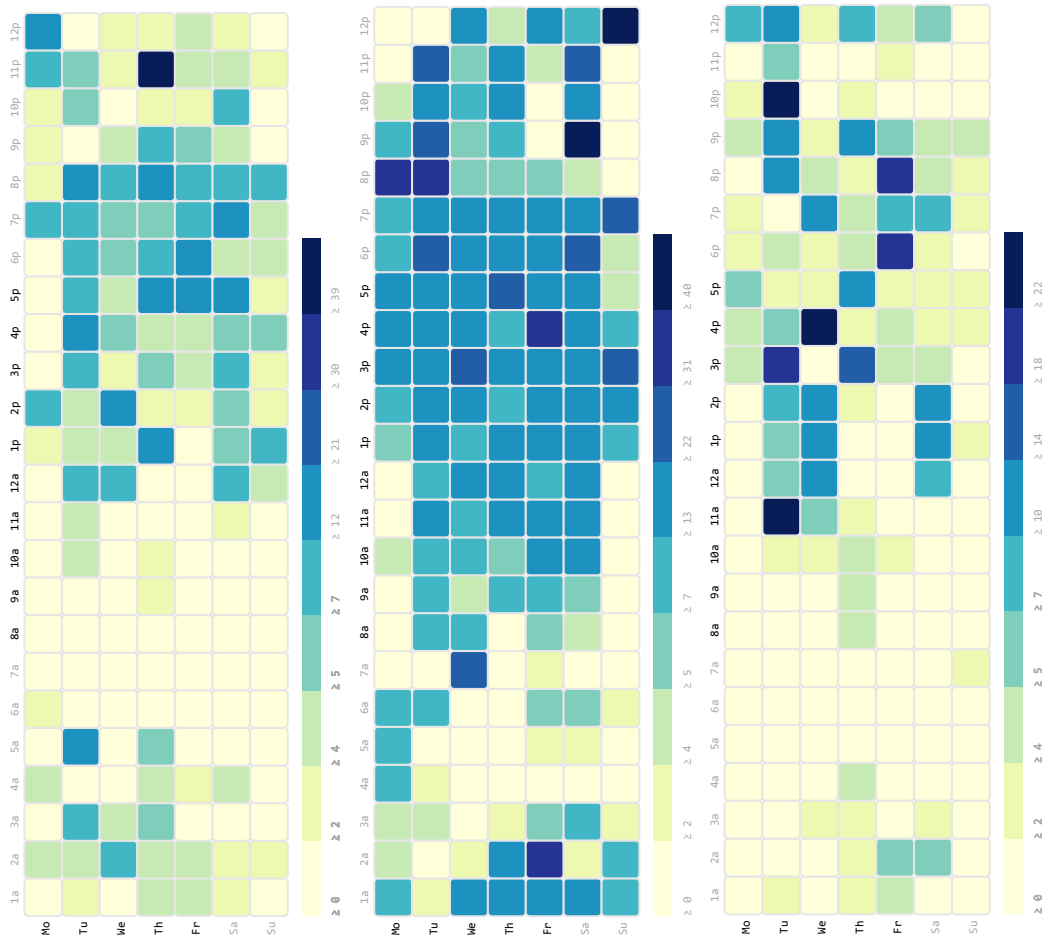
The teamwork involvement animation brings participants an attractive insights to the real-time activeness of all participants. And it encourages the participants to take continue *CW* activities to sustain the animation.

Algorithm 7 TI Animation of Teamwork Involvement

procedure TIANIMATION(P, I) ▷ P: list of participants, I: list of participants' teamwork involvement
  $blobs \leftarrow$ new $Blob[count(P)]$ ▷ Initialize the Blobs
  while true **do**
    for h from 0 to $count(P) - 1$, k from 0 to $count(I) - 1$ **do**
      $blobs[h].color \leftarrow$ assign color according to $P[h]$;
      $blobs[h].vx \leftarrow$ assign x-speed according to $I[h, k]$;
      $blobs[h].vy \leftarrow$ assign y-speed according to $I[h, k]$;
      $blobs[h].wobble \leftarrow$ assign wobbling degree according to $I[h, k]$;
    end for
    for h from 0 to $count(P) - 1$ **do**
      Render wobbling for $blobs[h]$
      Render moving for $blobs[h]$
    end for
  end while
end procedure

5.1.3.2 Demonstration of Teamwork Involvement Heat-map

Besides the real-time teamwork involvement animation, we provide a daily and hourly heatmap (in Figure 5.6) of participants' *CW* activities though our seamless repository. Variables for this heat-map include online workflow activities tracked by Redmine (in Figure 5.6(b)), on-site collaborative activities tracked by the BLE proximity app (in Figure 5.6(b)), and online collaborative activities through Slack (in Figure 5.6(c)). The horizontal axis is an enumeration of days (Monday to Sunday) and the vertical axis is an enumeration of hours (1:00 AM to 12:00 PM). The color axis enumerates the color representation for the activeness of participation involvement. Darker shades imply more active participation for that time period.



(a) Workflow Activities (b) (Onsite) Collaboration Activities (c) (Online) Collaboration Activities

Figure 5.6. Daily and Hourly Heat-map of Collaborative Workflow Activities during 1-year Pervasive Teamwork based on Seamless Repository

We implemented animation and heat-mapping to help participants get a better understanding of the real-time and historical status of the teamwork. This can make teamwork more interesting and enhance it cohesion.

5.1.4 Discussion on Seamless Integration

5.1.4.1 Relieved Barriers in Collaborative Workflow

We deployed and developed *CW* repositories for critical *CW* activities, and made an attempt to overcome the heterogeneity and workflow complexity barriers by improving the support of sharing through services of portfolio and workflow templates. We developed and deployed a Web portal, communication channels, and notification service, and made an attempt to overcome the skills gap,

workplace conflicts, and poor communication barriers by improving the support of interconnection. We also implemented a teamwork involvement animation and heat-map service to ease the teamwork disruption and less immersive barriers by improving the support of representation.

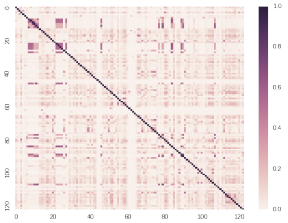
It is a long-term effort to validate the comprehensive impact of seamless integration. However, active teamwork involvement has been observed; the sustainable growing number of active projects, activities, users, and groups implies decreasing barriers after introducing the seamless teamwork repository.

5.1.4.2 Discovered Patterns from Collaborative Workflow

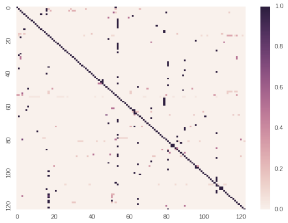
And by comparing the 3 sub heat maps in Figure 5.6, several teamwork behavioral patterns are found. For example, the collaborative workflow are mainly done at day time. And by comparing the grids of (MO, 7p∼9p), the heat of onsite workflow activities is much higher that of online workflow activities and collaborative activities, it is because that there is a presentation seminar at evening time on Monday and all participants have to join the onsite collaboration. And also according to the comparison of grids (MO, 11p∼12p), the participants left the laboratory but still continue their workflow pervasively in remote. It is a significant evidence of pervasive teamwork without the spatial and temporal limitations. Such discovered *CW* patterns brings participants a better understanding to the teamwork. And the patterns are helpful to optimize the plan for collaborative workflow; for example, a better scheduling of the tasks.

The seamless repository strengthens the collaborative workflow in pervasive teamwork, featured in availability, connectivity, and transparency. It has eased the *CW* barriers such as heterogeneity, workflow complexity, poor communication, skills gap, workplace conflicts, teamwork disruption and less immersive experience. However, it still far from being a smart system to understand participants' needs and provide proactive support to participants' collaborative workflow; participants have to deliver their requests to get the support. Thus our future work, firstly is to improve the less matured components as show in Table 5.1. Secondly but most importantly, it is to design and implement the proactive support system, which can utilize the exisiting knowledge and experience to support the real-time collaborative workflow.

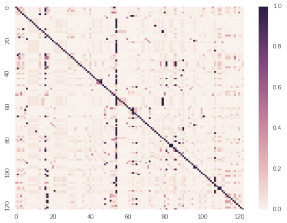
5.2 Knowledge Correlating



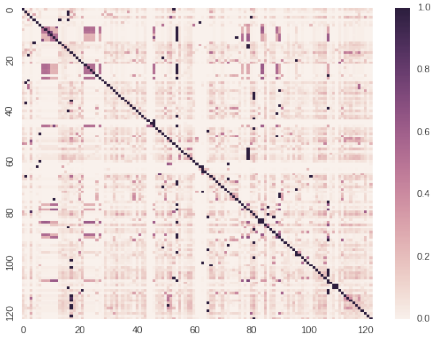
(a) $r_{conceptual(tf)}$: Conceptual Correlation using Terms-frequency



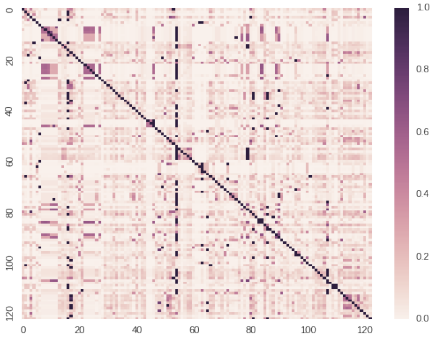
(b) $r_{relational(ln)}$: Relational Correlation using Neighbouring Links-ratio



(c) $r_{relational(lc)}$: Relational Correlation using Chained Links-ratio



(d) $r_{integrated(tf,ln)}$: Integrated Correlation using Terms-frequency and Neighbouring Links-ratio



(e) $r_{integrated(tf,lc)}$: Integrated Correlation using Terms-frequency and Chained Links-ratio

Figure 5.7. Comparison of Information Coverage of Correlation Measure to the Development Activity Data from the Seamless Repository

Five variations of correlation measures have been implemented to quantify the correlation among knowledge objects for development support. Experiment to test the performance of correlation measures is then run on the sample data set of development activities collected by the integrated seamless repository (see Table 4.2). The main purpose of activity awareness for development support, as shown in Figure 2.4, is to take advantage of the other developer's diverse knowledge and achievements to ease the problems related to current development, so that development can take more effort focusing on innovations to new problems. Therefore, the correlation measurement results should reflect the diversity of the pairwise collections among knowledge objects that were gradually contributed by many other developers. The comparison of information coverage of the five correlation measures is given through graphical evaluation in Section 5.2.1. A further analysis about the boundary conditions of the correlation measures is given in Section 5.2.2.

5.2.1 Comparison of Information Coverage using Graphical Evaluation

The correlation plots, as shown in Figure 5.7, illustrate the terms-frequency and links-ratio based all-pair correlation measures to the knowledge objects extracted from the sample data of development activities in the seamless repository. We measured approximately 120 knowledge objects in this experiment. The correlation value ranges from 0.0% to 100.0%, and each knowledge object is fully correlated with itself. The correlation value is represent by the darkness of the plotted colour, while the darker colour signifies higher correlation between the pairwise knowledge objects. Figure 5.7(a) shows that the $r_{conceptual(tf)}$ is symmetric; it considers only the factor of terms-frequency, and outputs undirected pairwise correlation measures. The other four correlation measures, as shown in the remaining plots, are asymmetric; they consider the factor of links-ratio, and output the directed pairwise correlation measure. The directed pairwise correlation of the links-ratio has the potential to be more diverse than the undirected pairwise correlation from the terms-frequency. The $r_{relational(lc)}$ is able to detect more correlated pairs than that of $r_{relational(ln)}$, since pairwise chains are more diverse than the pairwise neighbouring links in the graph of knowledge objects. Meanwhile, the integrated correlation $r_{integrated(tf,ln)}$ and $r_{integrated(tf,lc)}$ are more comprehensive than the other three measures, and $r_{integrated(tf,lc)}$ is more diverse than $r_{integrated(tf,ln)}$.

Table 5.2. Comparison of Correlation Measures

Correlation Measures	**Involved Attributes**			**Complexity**	**Coverage**	**Computation / Visualization**	
	TF	LN	LC				
$r_{contextual(tf)}$	◯	×	×	$\mathcal{O}(H \times N^2)$	Low	Alg. 2	Fig. 5.7(a)
$r_{relational(ln)}$	×	◯	×	$\mathcal{O}(N^3)$	Middle	Alg. 3 and 4	Fig. 5.7(b)
$r_{relational(lc)}$	×	×	◯	$\mathcal{O}(N^3)$	Middle	Alg. 5	Fig. 5.7(c)
$r_{integrated(tf,ln)}$	◯	◯	×	$\mathcal{O}((L+N) \times N^2)$	High	Alg. 6	Fig. 5.7(d)
$r_{integrated(tf,lc)}$	◯	×	◯	$\mathcal{O}((L+N) \times N^2)$	High	Alg. 6	Fig. 5.7(e)

Additional Description	
Coverage Comparison	$C(r_{integrated(tf,lc)}) > C(r_{integrated(tf,ln)}) > C(r_{contextual(tf)})$; $C(r_{relational(lc)}) > C(r_{relational(ln)})$.
Annotation	TF: Terms-frequency; LN: Neighbouring Links-ratio; LC: Chained Links-ratio. L: set size of covered terms in dictionary; N: set size of knowledge objects ◯: involved attributes; ×: not involved attributes.

Table 5.2 illustrates the difference in attributes, implementation, and computational complexity among those correlation measurements. The available attributes are terms-frequency, neighbouring links-ratio, chained links-ratio, and the attributes selection for correlation measure results in different information coverage. The computational complexity of those five correlation measures are

close to each other's when the scale of identical terms (L) is close to the scale of knowledge objects (N). However, normally L is much larger than N, so the correlation measures only based on links-ratio will be more computationally efficient than those only based on terms-frequency. Therefore, the $r_{integrated(tf,lc)}$, which integrates terms-frequency based conceptual correlation and chained links-ratio correlation, is the most comprehensive measure among those five variations. It achieves the best performance in information coverage, though being less computationally efficient. The graphical evaluation in Figure 2.4 shows the feasibility of implementing the use case of development support using activity awareness.

5.2.2 Boundary Conditions

It requires further investigation to evaluate the significance of using the development activity data to aware the correlations among knowledge objects for the purpose of development support. There are several boundary conditions limiting the performance of the knowledge correlation to provide development support.

- *Data quality*: The quality of knowledge information contributed by developers will affect the result of the integrated correlation measure. For example, the terms (linguistic symbols) that are used by developers in the development activities should understandable, meaningful, and also commonly used by other developers with similar domain knowledge. And the links among knowledge objects should also be judged by developers' professional experience before being appended. There will be quality constraint by fostering developers' conscientiousness in providing the quality knowledge and also by warming the teamwork.

- *Data scale*: Either the scale of knowledge objects or the scale of development activities in singular knowledge object will affect the result of the integrated correlation measure. The larger amount of objects and the more details in each object, the better result from the correlation measure. Therefore, there will be a scale constraint to the amount of knowledge objects in repository and the volume of the singular knowledge objects. We recommend repository scale to be over $O(10^2)$, and activity amount in singular objects to be over $O(10)$ based on our current experience.

- *Investigation scope to data attributes*: The investigation scope to the data attributes will affect

the result of the integrated correlation measure. There are other types of correlations besides the contextual or relational ones, which use the data attributes of terms frequency or (neighbouring/chained) links-ratio. For example, the data attribute of development time expense and developers' profile information are also playing important roles in correlating the knowledge objects. Therefore, there will be a scope constraint to investigate the major attributes based on the observation to the majorities.

- *Balancing impact of different correlations*: The impact balance among different correlation measure will also affect the result of the integrated correlation measure. The impact of the different correlations can not simply equal to each other due to unbalance amount of data attributes: for example, the amount of terms might be much less than that of links in some repository. That means that the impact of the correlation using terms-frequency has to be decreased due to insufficient data of terms, and vice versa. Therefore, there will be an impact factor constraint according to the ratio between the amount of terms and links.

- *Developers' response to the development support*: Developers' response to the development support is also important feedback condition to improve the correlation measure. For example, developers' may not be interested in jumping over too many times over the chained knowledge objects through the URLs. In such a case, therefore there will be a maximum length constraint to the chains in the TFCLR based measure.

5.2.3 Future Improvement of Knowledge Correlating

5.2.3.1 Future Improvement of Correlation Measure

In the graphical evaluation, we have explored the performance of terms-frequency and neighbouring/chained links-ratio based correlation measures in quantifying the correlations among knowledge objects. However, the significance of the measurement needs further validation in practice. There are many factors that may affect the significance of correlation measurement to development activity data, and finally lead to the fitness of development support. It needs a continuous effort to improve the measurement to diverse correlations. First of all, there could be further variations to improve the correlation measure; for example, it could be beneficial to reduce the size of identical terms L by ignoring the noise terms. Secondly, the integrated correlation currently only

considers terms-frequency and (neighbouring/chained) links-ratio, there are also other factors in the development repository that have not been included yet. For example, the working time cost, the developers' profile information of knowledge objects, and the collaborations of developers both on-site and remotely could also be considerably important correlation factors. Last but not least, the results of correlation measurement are also affected by the weight definition to the correlation factors according to specific engineering requirements. For example, the conceptual correlation can be adjusted to be less weighted in calculating the integrated correlation if there is a significant amount of outlier information in textual contents in the development activity data.

5.2.3.2 Future Improvement of Seamless Integration

As the scale of seamless repositories grows and collaborative workflow is standardized, diverse correlations among not only knowledge objects but also developers become more measurable in the future. With long-range perspective, developers using such a seamless repository can be more efficient in sharing experience and productive in focusing on innovations. Thus, we took external efforts to investigate more approaches for development support. For instance, we investigated the matching between developers' engineering skills and knowledge objects in one study [56], and proposed principles of lean software development to optimize the human resources of developers for developing minimal viable products. Furthermore, we extended the seamless repository to track the physical activities in development environment, and developed a BLE (Bluetooth Low Energy) proximity technology based CICO (Check-in-Check-out) system in another study [38] to records developers' check-in data in physical workplaces. Such a proximity dataset is rich in spatial and temporal information of development activities, making it possible to understand developer's physical collaboration in teamwork, and improve the customization of development support for different developers. In a summation, our future effort will be in three directions. One is to improve and add new functions to the seamless repository to track the development activities more comprehensively. And another effort is to improve the analysis on the development activity data to measure diverse correlations. A third effort is to evaluate and improve the development support to help the practical development in practice.

Chapter 6

Summary and Future Work

6.1 Summary

The teamwork nowadays has been empowered by diverse support systems, however, the collaborative workflow gaps still commonly exisit among co-workers who are different in preferences of support systems and needs of domain knowledge. The different but persistent personal preferences of using the the support systems require new support of seamless integration. And co-workers' different background knowledge and their different purposes of utlizing knowledge require the support of knowledge correlating.

This dissertation aims at bridging the collaborative workflow gaps by providing a new support platform. For the purpose of seamless integration, three scenarios are given to show the necessity of improving the supports of sharing, interconnection, and visualization to bridge the gaps of information, communication, and representation respectively. And a three-layered framework is specified to show the approaches for seamless integration of multiple support systems in to a seamless repository. And the information security and system scalability design is also considered to guarantee the quality of service after seamless integration. And for the purpose of knowledge correlating, a graph model is presented to organize the knowledge objects from different support systems. And a correlation measure based on terms-frequency and chained links-ratio (TFCLR) is proposed to quantify the conceptual and relational correlation among the knowledge objects. Then the system specification and technical implementation are also given to measure the correlations among knowledge objects by analyzing the raw data of collaborative workflow activities.

The support of sharing is proposed to bridge the gap of information, including the services of portfolio and workflow templates. The service of portfolio helps teamwork participants know others more quickly by learning from their shared portfolio in repository, and vice versa. The service of workflow templates helps teamwork participants process tasks more feasibly by breaking them down onto feasible templates, and match with execution plan in repository.

The support of interconnection is proposed to bridge the gap of communication, including the services of web portal, communication channels, and notification. The service of web portal helps teamwork participants harness the skills by cooperating with others in executing the predefined tasks in repository, and exploring over Web portal. The service of communication channels help teamwork participants deliver needs more meaningfully be contacting in preferable channels, and with reference to predefined pieces in repository. The service of notification helps teamwork participants warm up in workplace by setting up repository preference, and notifying activities in repository to achieve interests.

The support of visualization is proposed to bridge the gap of representation, including the service of heat-map and animation. The service of heat-map helps teamwork participants get a broad sense of team by reviewing the historical teamwork performance visualized by repository. And the service of animation help teamwork participants get a narrow sense of individual participant by reviewing the real-time teamwork status visualized by repository.

The support of correlation measure is proposed to bridge the gap of knowledge by utilizing the correlated knowledge objects through the constructed knowledge graph in repository.

Comparing with other support systems, the seamless integration in this platform has better functionality in sharing, interconnection, and visualization. And comparing with other collaboration measure, the TFCLR measure achieves better performance in information coverage and usability, and also has tolerable performance in speed and feasibility.

6.2 Plan of Future Work

The gaps of collaborative workflow for project-based development commonly exist in various organizations, such as startup teams, commercial corporates, educational institutes, manufacturing factories. Each type of organization relies on different collaborative workflow tool sets, and has

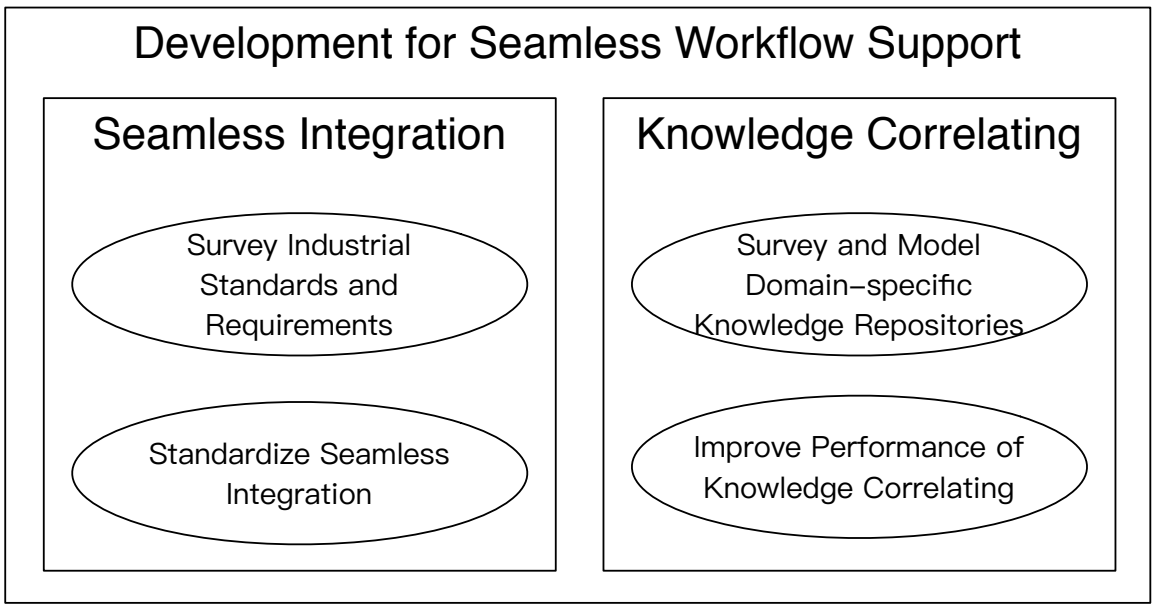


Figure 6.1. Development for Seamless Workflow Support

different problems in correlating their domain specific knowledge. The continuous future work based on current study will focus on development for seamless workflow support (see Figure 6.1), including the sub issues of generalizing the seamless integration of different support systems, and specialize the knowledge correlating for the domain specific knowledge repositories.

In order to well standardize the seamless integration to satisfy the industrial specifications and requirements of support systems, a new survey is under going about industrial standards and requirements of system integration of support systems. And the technical improvement to the previously development support platform will be planed based on the standardized specification from the survey. Then the planed improvement will be implemented through the application of seamless support system for collaborations not only in experimental projects within laboratory but also in other projects that are more practical and challenging in daily life.

And in order to signify the knowledge correlating to be adaptable to different domain-specific knowledge, a new survey and modeling to different domain-specific knowledge repositories is under going. Based on the survey results, new improvement of correlation algorithms will be the next step to achieve improved performance in the knowledge correlation measure in different domain-specific knowledge repositories, for example through the approach of involving more correlation factors.

Bibliography

[1] B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, J. Tao, and Y. Zhao, "Scientific workflow management and the kepler system," *Concurrency and Computation: Practice and Experience*, vol. 18, no. 10, pp. 1039–1065, 2006.

[2] W. Van der Aalst, "Petri net based scheduling," *Operations-Research-Spektrum*, vol. 18, no. 4, pp. 219–229, 1996.

[3] P. R. Cohen and H. J. Levesque, "Teamwork," *Journal of Noûs*, pp. 487–512, 1991. [Online]. Available: http://web.media.mit.edu/~cynthiab/Readings/cohen-teamwork.pdf

[4] J. Wu, L. Ping, X. Ge, Y. Wang, and J. Fu, "Cloud storage as the infrastructure of cloud computing," in *Proceedings of International Conference on Intelligent Computing and Cognitive Informatics (ICICCI)*, June 2010, pp. 380–383.

[5] G. Graefe, F. Halim, S. Idreos, H. Kuno, S. Manegold, and B. Seeger, "Transactional support for adaptive indexing," *The International Journal on Very Large Data Bases (VLDB)*, vol. 23, no. 2, pp. 303–328, 2014.

[6] J. Golbeck and C. Halaschek-Wiener, "Trust-based revision for expressive web syndication," *J. Log. Comput.*, vol. 19, no. 5, pp. 771–790, 2009.

[7] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, 2013.

[8] S. Dustdar and H. Gall, "Pervasive Software Services For Dynamic Virtual Organizations," in *Processes and Foundations for Virtual Organizations*. Springer US, 2004, pp. 201–208.

[9] J. Noll, S. Beecham, and I. Richardson, "Global software development and collaboration: barriers and solutions," *Magazine of ACM Inroads*, vol. 1, no. 3, pp. 66–78, 2010.

[10] G. Sadowski-Rasters, G. Duysters, and B. Sadowski, *Communication and Cooperation in the Virtual Workplace: Teamwork in Computer-mediated-communication*. Edward Elgar Publishing, 2006.

[11] A. H. Littlejohn and L. A. Stefani, "Effective use of communication and information technology: bridging the skills gap," *The Journal of the Association for Learning Technology (ALT): Research in Learning Technology*, vol. 7, no. 2, 1999.

[12] C. K. De Dreu and L. R. Weingart, "Task versus relationship conflict, team performance, and team member satisfaction: a meta-analysis," *Journal of Applied Psychology*, vol. 88, no. 4, p. 741, 2003.

[13] J. Cardoso, "About the complexity of teamwork and collaboration processes," in *Saint Workshops on Applications and the Internet Workshops (SAINT 2005 Workshops)*. IEEE Computer Society, 2005, pp. 218–221.

[14] L. R. Brawley, A. V. Carron, and W. N. Widmeyer, "Exploring the relationship between cohesion and group resistance to disruption," *Journal of Sport and Exercise Psychology*, vol. 10, no. 2, pp. 199–213, 1988.

[15] R. L. Jackson and E. Fagan, "Collaboration and learning within immersive virtual reality," in *Proceedings of the Third International Conference on Collaborative Virtual Environments*, ser. CVE '00. New York, NY, USA: ACM, 2000, pp. 83–92. [Online]. Available: http://doi.acm.org/10.1145/351006.351018

[16] Y. Zhang and Y. Zhou, "Transparent computing: a new paradigm for pervasive computing," in *International Conference on Ubiquitous Intelligence and Computing*. Springer, 2006, pp. 1–11.

[17] W. M. van Der Aalst, A. H. Ter Hofstede, B. Kiepuszewski, and A. P. Barros, "Workflow patterns," *Distributed and parallel databases*, vol. 14, no. 1, pp. 5–51, 2003.

[18] J. Cardoso, “Business process quality metrics: Log-based complexity of workflow patterns,” in *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*. Springer, 2007, pp. 427–434.

[19] L. Bobbitt, G. Pelton, W. Morrison, J. Miner, J. Rossie, and D. Walker, “Asynchronous workflow participation within an immersive collaboration environment,” Dec. 10 2009, uS Patent App. 12/133,226. [Online]. Available: https://www.google.com/patents/US20090307189

[20] J. McDonough, R. Torchon, M. Walton, and W. Crooks, “Distributed, collaborative workflow management software,” Mar. 11 2004, uS Patent App. 10/311,449. [Online]. Available: https://www.google.com/patents/US20040049345

[21] Jean-Philippe Lang, “Redmine: A Flexible Project Management Web Application,” http://www.redmine.org, accessed: January 2015.

[22] D. Bertram, A. Voida, S. Greenberg, and R. Walker, “Communication, collaboration, and bugs: The social nature of issue tracking in small, collocated teams,” in *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, ser. CSCW ’10. New York, NY, USA: ACM, 2010, pp. 291–300. [Online]. Available: http://doi.acm.org/10.1145/1718918.1718972

[23] Atlassian Software Company, “Atlassian JIRA: Issue & Project Tracking Software,” https://www.atlassian.com/software/jira, accessed: January 2015.

[24] G. S., “Revision control system (rcs) in computational sciences and engineering curriculum,” in *Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment*, ser. XSEDE ’14. New York, NY, USA: ACM, 2014, pp. 76:1–76:3. [Online]. Available: http://doi.acm.org/10.1145/2616498.2616576

[25] GitLab Inc., “GitLab: Create, review and deploy code together ,” https://gitlab.com, accessed: January 2015.

[26] GitHub, Inc., “GitHub: Build software better together,” https://github.com, accessed: January 2015.

[27] Atlassian Software Company, “Atlassian Bitbucket: Git and Mercurial code management for teams,” https://bitbucket.org, accessed: January 2015.

[28] F. Sousa, M. Aparicio, and C. J. Costa, “Organizational wiki as a knowledge management tool,” pp. 33–39, 2010.

[29] Wikipedia Online Community, “Comparison of wiki software,” http://en.wikipedia.org/wiki/Comparison_of_wiki_software, accessed: January 2015.

[30] Team of MediaWiki Project, “MediaWiki: a free softwared open source wiki package written in PHP,” https://www.mediawiki.org/wiki/MediaWiki, accessed: January 2015.

[31] PukiWiki, “PukiWiki FrontPage,” http://pukiwiki.sourceforge.jp, accessed: January 2015.

[32] S. K. Patel, V. R. Rathod, and S. Parikh, “Joomla, drupal and wordpress-a statistical comparison of open source cms,” in *Trendz in Information Sciences and Computing (TISC), 2011 3rd International Conference*. IEEE, 2011, pp. 182–187.

[33] T.-H. Wu, “Method of transferring resources between different operation systems,” Nov. 22 2005, US Patent No. 6968370.

[34] Bluetooth Special Interest Group, SIG, “Specification of the Bluetooth System Covered Core Package Version 4.0,” 2010. [Online]. Available: https://www.bluetooth.org/en-us/specification/adopted-specifications

[35] Slack Technologies, Inc., “Slack: Be less busy,” http://slack.com, accessed: January 2015.

[36] J. Oikarinen and D. Reed, “Internet Relay Chat Protocol,” *Internet Engineering Task Force (IETF)*, pp. 1–65, 1993.

[37] R. B. Jennings, E. M. Nahum, D. P. Olshefski, D. Saha, Z.-Y. Shae, and C. Waters, “A study of internet instant messaging and chat protocols,” *IEEE Network Magazine*, vol. 20, no. 4, pp. 16–21, July 2006.

[38] Y. Wu, J. Wang, L. Jing, Y. Zhou, and Z. Cheng, “A cico system based on ble proximity,” in *2015 IEEE 7th International Conference on Awareness Science and Technology (iCAST)*, Sept 2015, pp. 180–183.

[39] M. Cusumano, D. B. Yoffie *et al.*, "Software development on internet time," *Computer*, vol. 32, no. 10, pp. 60–69, 1999.

[40] A. Fuggetta, "Software process: a roadmap," in *Proceedings of the Conference on the Future of Software Engineering*. ACM, 2000, pp. 25–34.

[41] K. M. Benner, M. S. Feather, W. L. Johnson, and L. A. Zorman, "Utilizing scenarios in the software development process," *Information system development process*, vol. 30, pp. 117–134, 2014.

[42] Centers for Medicare & Medicaid Services, "Selecting a development approach," *Centers for Medicare & Medicaid Services*, pp. 1–10, 2008. [Online]. Available: http://www.cms.gov/Research-Statistics-Data-and-Systems/CMS-Information-Technology/XLC/Downloads/SelectingDevelopmentApproach.pdf

[43] D. Bertram, A. Voida, S. Greenberg, and R. Walker, "Communication, collaboration, and bugs: The social nature of issue tracking in small, collocated teams," in *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, ser. CSCW '10. New York, NY, USA: ACM, 2010, pp. 291–300. [Online]. Available: http://doi.acm.org/10.1145/1718918.1718972

[44] M. A. Junichi Kanai, "Redmine as a Web-Based Collaboration Tool in Engineering Design Courses," *American Society for Engineering Education*, vol. II, pp. 8989–9001, 2013.

[45] B. Leuf and W. Cunningham, "The wiki way: quick collaboration on the web," 2001.

[46] J. Loeliger and M. McCullough, *Version Control with Git: Powerful tools and techniques for collaborative software development*. " O'Reilly Media, Inc.", 2012.

[47] R. Burke, "Hybrid recommender systems: Survey and experiments," *User modeling and user-adapted interaction*, vol. 12, no. 4, pp. 331–370, 2002.

[48] M. Balabanović and Y. Shoham, "Fab: content-based, collaborative recommendation," *Communications of the ACM*, vol. 40, no. 3, pp. 66–72, 1997.

[49] G. Linden, B. Smith, and J. York, "Amazon. com recommendations: Item-to-item collaborative filtering," *IEEE Internet computing*, vol. 7, no. 1, pp. 76–80, 2003.

[50] A. Rajaraman and J. D. Ullman, “Data mining,” in *Mining of Massive Datasets*. Cambridge University Press, 2011, pp. 1–17, cambridge Books Online. [Online]. Available: http://dx.doi.org/10.1017/CBO9781139058452.002

[51] “AltBeacon, The Open and Interoperable Proximity Beacon Specification,” http://altbeacon.org, accessed: January 2015.

[52] A. Huang, “Similarity measures for text document clustering,” in *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, 2008, pp. 49–56.

[53] A. Strehl, J. Ghosh, and R. Mooney, “Impact of similarity measures on web-page clustering,” in *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, 2000, pp. 58–64.

[54] mRova Solutions, “A Free One Page Responsive HTML Resume Template,” http://www.mrova.com/free-one-page-responsive-html-resume-template, accessed: December 2014.

[55] J. Bryant and M. Jones, “Responsive web design,” in *Pro HTML5 Performance*. Springer, 2012, pp. 37–49.

[56] Y. Wu, K. Sato, L. Jing, J. Wang, and Z. Cheng, “The lean awareness in software-intensive engineering: Experience from one project,” in *2015 IEEE 7th International Conference on Awareness Science and Technology (iCAST)*, Sept 2015, pp. 168–173.