"Modeling Communities in Information Systems: Informal Learning Communities in Social Media"

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades einer Doktorin der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Zinayida Kensche, geb. Petrushyna, MSc.

aus Odessa, Ukraine

Berichter: Universitätsprofessor Professor Dr. Matthias Jarke

Universitätsprofessor Professor Dr. Marcus Specht

Privatdozent Dr. Ralf Klamma

Tag der mündlichen Prüfung: 17.11.2015

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek online verfügbar.

© Copyright by 2016
All Rights Reserved

Abstract

Information modeling is required for creating a successful information system while modeling of communities is pivotal for maintaining community information systems (CIS). Online social media, a special case of CIS, have been intensively used but not usually adopted for learning community needs. Thus community stakeholders meet problems by supporting learning communities in social media. Under the prism of Community of Practice theory, such communities have three dimensions that are responsible for community sustainability: mutual engagement, joint enterprises and shared repertoire.

Existing modeling solutions use either perspectives of learning theories, or analysis of learner or community data captured in social media but rarely combine both approaches. Therefore, current solutions produce community models that supply only a part of community stakeholders with information that can hardly describe community success and failure. We also claim that community models must be created based on community data analysis integrated with our learning community dimensions. Moreover, the models need to be adapted according to environmental changes.

This work provides a solution to continuous modeling of informal learning communities in social media. In particular, it makes the following contributions: 1. A metamodel of learning communities and its specific cases in social media. 2. A process of continuous community model creation that consists of four phases that model, refine, monitor and analyze learning communities. The phases and their realizations can be used to model any learning community with the purpose to support community evolution and to improve social media facilities to satisfy community needs. 3. Methods for community data analysis and storage have been exploited for retrieving learning community states to manage competences in a collaborative space and specifying culturally sensitive requirements of communities towards social media. 4. Our formal representation of a learning community has been used to model early requirements of learning communities and their evolution and to validate the effectiveness of possible community changes through multi-agent simulation.

Zusammenfassung

Für die Entwicklung erfolgreicher Informationssysteme ist Informationsmodellierung notwendig während die Modellierung von Gemeinschaften für die Pflege und Weiterentwicklung von Community-Informationssystemen (CIS) eine entscheidende Stütze bilden kann. Soziale Online-Medien, ein Spezialfall von CIS, werden häufig zur Unterstützung von Lerngemeinschaften verwendet obwohl sie ohne Anpassung nicht zu diesem Zweck geeignet sind. Durch das Prisma der Theorie über praxisbezogene Gemeinschaften betrachtet haben solche Gemeinschaften drei Dimensionen die ihre Zukunftsfähigkeit bestimmen: gegenseitige Verbindlichkeit, ein gemeinsames Unterfangen, sowie den Zugriff auf das gleiche Repertoire an Ressourcen.

Existierende Modellierungslösungen greifen entweder auf die Perspektive der Lerntheorien zurück oder basieren auf einer Analyse von Lern- oder Gemeinschaftsdaten in Sozialen Medien aber kombinieren nur selten beide Ansätze. Daher führen bestehende Lösungen zu Gemeinschaftsmodellen die nur einen Ausschnitt der Interessengruppen mit notwendigen Informationen versorgen und kaum hinreichend sind um Erfolg und Scheitern von Gemeinschaften zu erklären. Stattdessen müssen Gemeinschaftsmodelle auf Basis einer Integration von Datenanalysen sowie der Dimensionen von Lerngemeinschaften erzeugt sowie an sich ändernde Umgebungen angepasst werden.

Diese Arbeit stellt eine Lösung für die kontinuierliche Modellierung von informellen Lerngemeinschaften in sozialen Medien dar. Insbesondere leistet sie die folgenden Beiträge: 1. Ein Metamodell für Lerngemeinschaften und Spezialfälle dieses Modells in sozialen Medien. 2. Einen Prozess zur Erzeugung und kontinuierlichen Weiterentwicklung von Gemeinschaftsmodellen der aus vier Phasen besteht: Modellierung, Verfeinerung, Beobachtung und Analyse. Die Phasen und ihre Umsetzung können zur Modellierung beliebiger Lerngemeinschaft zur Unterstützung ihrer Evolution und zur Weiterentwicklung der Funktionalität sozialer Medien genutzt werden um so Anforderungen der Gemeinschaften zu erfüllen. 3. Es wurde ein Verfahren zur Analyse und Speicherung von Gemeinschaftsdaten genutzt das es ermöglicht Eigenschaften von Lerngemeinschaften zu erkennen und so die Entwicklung von Kompetenzen in kollaborativen Umgebungen sowie die Spezifikation kulturell sensitiver Anforderungen an soziale Medien ermöglicht. 4. Die formale Darstellung von Lerngemeinschaften wurde zur Modellierung früher Anforderungen von Gemeinschaften und ihre weitere Entwicklung genutzt. Die Effektivität möglicher Änderungen an Gemeinschaften wurde mit Hilfe von Multiagentensimulationen verifiziert.

Acknowledgments

This work finalizes an important part of my life and I could not close it without naming the people that I want to thank for their advice and support.

First of all, I want to thank my principal advisor Prof. Matthias Jarke for precious discussions on the contents of this work and for giving me the opportunity to freely research on a challenging topic which I enjoyed a lot. I would like to specially thank PD Dr. Ralf Klamma who served as coadvisor and gave major advice and recommendations that helped me to finish the dissertation work. Moreover, I thank Prof. Marcus Specht for his willingness to become my second advisor for this thesis.

During my work I cooperated with my colleagues and friends at the chair, particularly, Anna Hannemann, Yiwei Cao, Manh Cuong Pham, Dominik Renzel, Istvan Koren, Milos Kravcik, Dejan Kovachev, Mohsen Shahriari, Dominik Schmitz, Petru Nicolaescu, Michael Derntl, Katya Neulinger, Sandra Geisler, Stefan Schiffer, Daniele Glockner, Gabriele Hoeppermanns, Tatiana Liberzon, Reinhard Linde, Chao Li. I want to thank them for inspiring, interesting, and often funny academic and non-academic discussions as well as for our fine team work in teaching activities. Moreover, I thank my former students including Julian Krenge, Florian Oberloer, Alexander Ruppert, Ergang Song, Alex Tritthart and many others for their support in realizing parts of this work.

This research would not have been possible without the funding by the TEMPUS project on Cairo University E-learning Center, the EU FP7 project on Responsive Open Learning Environments, the Lifelong Learning project on Teacher's Lifeling Learning Network and further support from BIT Research School and RWTH Aachen University. I thank my colleagues in these projects for years of collaborative research, particularly, Prof. Katherine Maillet, Prof. Magda Fayek, Prof. Martin Wolpers, Alexander Nussbaumer, Felix Mödritscher, Riina Vuokari, Prof. Peter Sloep, Mark Kröll and Prof. Markus Strohmaier. Furthermore, I give my thanks to organizers and students I collaborated with at Joint Technology Enhanced Learning summer schools including Prof. Ambjörn Naeve, Nicolas Weber, Sandy El Helou, Joris Klerkx, Anna Lea Dyckhoff, Hannes Ebner, Fridolin Wild and others.

Last but not least, I thank my daughter, Helena, for giving me time for writing, my husband, David, for taking the load off and making relevant comments for the work, my parents and family for support from the beginning till the end of this thesis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Inexpected actions of users become a major problem in development of information systems (IS). In the Web 2.0 users have a huge impact on IS sustainability and therefore modeling collective influence of users and retrieving their requirements to IS is important and relevant for IS maintenance. An online social medium, a community information system in the Web 2.0, is continuously affected by its heterogeneous users that make it challenging for the medium to correspond to customer needs.

Social media bring opportunities and affordance (Gibson, 1977) for communities but at the same time complexity of a community structure increases due to the amount of ways how peers can collaborate increases due to types of activities the social media allow. Furthermore, due to social media functionalities the amount of produced and shared information and the amount of heterogeneous representations of these information complicate investigation of communities.

Understanding of communities is further impeded by community context — learning is one of these. Learners have different learning goals, different levels of expertise, different preferences in communication and different visions how to realize goals. Even though, they cooperate to find solutions to the same issues and thus organize themselves in informal learning communities to fulfill the same goals. Belonging to the same community indicates similar purposes and direction of knowledge as well as trust between collaborators (Wenger, 1998) but heterogeneity of users causes learning communities to be complex organisms with many entities that need to be considered by modeling or investigating communities and by estimating community success.

Although success of learning depends on communities of learners (Wenger, 1998) and their activities (Vygotsky, 1978; Engeström, 1987; Iandoli and Zollo, 2008), social media play a role in learning communities' success and communities have to consider digital media as new actors in their environment since they allow to measure learning success. Active in social media communities learners get better scores than others (Anderson et al., 2014) while professionals get benefits from participating in social media discussions. So three quarters of Korean software developers do a better job because they collaborate in online communities while 68% of them refine their professional skills due to their participation (Ala-Mutka et al., 2009). Being a part of a

community its members not only gain knowledge but as well learn to handle responsibilities like in Wikimedia[1], where active members are granted privileges and responsibilities. Furthermore, learners gain lifelong learning competences (European Parlament and the Council, 2006) such as collaborating, critical and reflective thinking and metacognition in online communities (Antoniou and Siskos, 2007; Xie et al., 2008).

Maintaining community needs of online learning communities and modeling learning communities considering complexity of cooperation between learners in social media is hardly done manually. Communities consist of thousands or millions of users that perform millions of activities and produce millions of artifacts, such as texts. Informal learning communities, groups of online collaborating peers that are not limited by institutional frames, include heterogeneous learners with different learning styles, expertise and pace of learning. Collections of learners are continuously changing and a flow of new learners into communities may turn it on its head as the learners have different knowledge, learning goals and strategies than community members. Therefore, to support both community needs and community stakeholders we require social software that allows to estimate learning community changes, retrieve community needs and model learning communities (Klamma, 2013).

Some approaches define community needs (Hilts and Yu, 2011; Ferreira and Silva, 2012) and community models (Suh and Lee, 2006) though just a few of them (Kleanthous and Dimitrova, 2007, 2010) consider the gap existing between technology and people-based theories (Iandoli and Zollo, 2008) such as learning theories. Even though, they did not bring results of technological and learning theories together. Replicating an experience of social science[2], this work deals with modeling of learning communities in social media regarding learning theories and technology.

The work follows two purposes. Firstly, community models can support community stakeholders in estimating communities and their success and secondly they help to estimate community issues and needs and what are solutions to the issues based on experience of other communities.

1.1 Research Questions and Methods

In the end it is the users who are making social media to live that is why success of social software — tools of social media — is usually estimated by the number of people that participate in them. Such a measure indicates the popularity of the software though it can not estimate the success of communities that use the software (Klamma, 2010). Therefore, social software recently includes many tools that exploit and analyze data to provide personalized or group-oriented information about social media users.

Many research approaches investigated collaborations of learners (Dascalu et al.,

[1]The non-profit Wikimedia Foundation `http://www.wikimedia.org/`, Last access on 29.07.2014

[2]the gap between information modeling and social science provoked Gilbert and Troitzsch (2005) to create and simulate models of societies

2010; Scheffel et al., 2011; Ferguson et al., 2013) and provided reports for them or their peers (Upton and Kay, 2009; Florian et al., 2011), though the created tools worked only with a particular social medium and these tools estimated collaborations of learners but in most cases did not consider communities as items of investigation. Estimating community success operating with tools that measure learners' activities or collaborations only is complicated. Even considering only corresponding to communities activities as in (Kleanthous and Dimitrova, 2007, 2010), their achievements need to be associated as well with learning theories to get relevant results to retrieve community models and needs.

My work provides a framework for maintaining community models and needs by answering these questions:

- *How to model online learning communities connecting technology and learning theories?* Partnership between technology outcomes and learning theories interpretations is a prerequisite for a truly objective modeling of learning communities. This is because learning plays a pivotal role on formation and development of communities while technological analysis provides unbiased information about activities of communities in social media. Furthermore, no concrete approach for modeling online learning communities was proposed so far.

- *How can community models be effectively refined?* Information modeling represents a number of statements about some artifacts. Community models can represent unprejudiced information about community states and actors and therefore can be used for comparison, extension, analysis, and simulation of learning communities that helps to find most suitable or efficient models.

- *How to monitor and analyze learning communities to support them and respond to community needs?* Social media facilitate collaboration between learners and creation of communities while each community is different in the number of users, content of topics and goals and many other characteristics. According to Anderson (2006) each community owns its niche in the long tail of communities, where a niche can define popularity of a community or a community topic. Communities may be similar in 0structure and can include similar artifacts such as texts, though their needs and states usually differ. But technologies required for the investigation are similar and therefore can be reused if social media data is appropriately represented. Furthermore, to provide objective data for community modeling one choice of technologies should be done in such a way that they satisfy learning theories' requirements.

1.2 Approaches and Contributions

This work deals with *Modeling* in general and with *Community Modeling* in particular. I focus my research on informal learning communities that emerge in online social media. Since learning has an impact on community structure and states and communities

affect learning process, I consider learning communities as communities of practice (CoP) (Wenger, 1998) where learners influence learning processes in their communities (Vygotsky, 1978; Engeström, 1987). CoP is a well-established concept that is usually used in Technology Enhanced Learning (TEL) for investigating learning communities. In CoP learners share knowledge collaborating in trust, following the same goals and rules, and possessing common understanding and tools.

Another important aspect I consider is a technology — a social medium in our case — that plays a pivotal role for informal online learning communities. Learning theories miss the explanation of media operations that trigger learning process and occur between media and communities (Klamma, 2010). Transcriptivity theory (Jäger, 2002; Spaniol et al., 2007; Jäger et al., 2008) proposes a solution that helps to describe the flow of knowledge between communities, individuals and media. The theory emphasizes media operations that explain information exchange and cognitive processes of learners.

One of the challenges of this work is to enable the synergy of technology and learning theories. Learners in social media communities have been leave traces describing their learning activities thus the traces provide a foundation for mining learning communities data. Before mining data we need to store it in an application- and media-independent way. We consider the Mediabase model (Klamma and Petrushyna, 2008) as the data model for community environment that acknowledges Actor Network Theory (Latour, 2005) to represent different entities of community environment whether they are human or non-human. Data modeled using metaentities from the Mediabase model, are valid as an input for numerous social software without any transformation. Since we have to operate with large amounts of data and retrieve it using complex queries we utilize the data warehouse technology (Jarke et al., 1999) to achieve efficiency in data queries. After the data is stored, we appeal to data mining and information retrieval and techniques of data science to extract interesting and relevant data from communities.

Community data analysis provides information to detect community states but the data do not inform about community needs, issues and solutions. Replicating the ATLAS framework (Klamma et al., 2006a) that aims to access community needs, I extract community needs based on modeling, refinement, monitoring and analysis phases. The monitoring is performed using data warehouse techniques and the Mediabase model as described in the previous paragraph. The analysis include a number of data science techniques such as community detection and evolution, intent analysis, emotional analysis, and named entity recognition. While for the modeling and the refinement phases we appeal to approaches originating from Artificial Intelligence — agent-based models and multi-agent simulation.

Agent-based models emphasize each agent, their goals, activities, strategies and payoffs (Macal and North, 2009). The modeling approach and its simulation using multi-agent systems have been broadly adopted in social science to represent communities (Gilbert and Troitzsch, 2005). Such a representation gives a clue about community processes and the outcomes of these. We utilize an agent-based and goal-oriented

modeling approach (Yu, 1997) that emphasizes agents' networks and dependencies, such as goals and is used to detect early requirements of users.

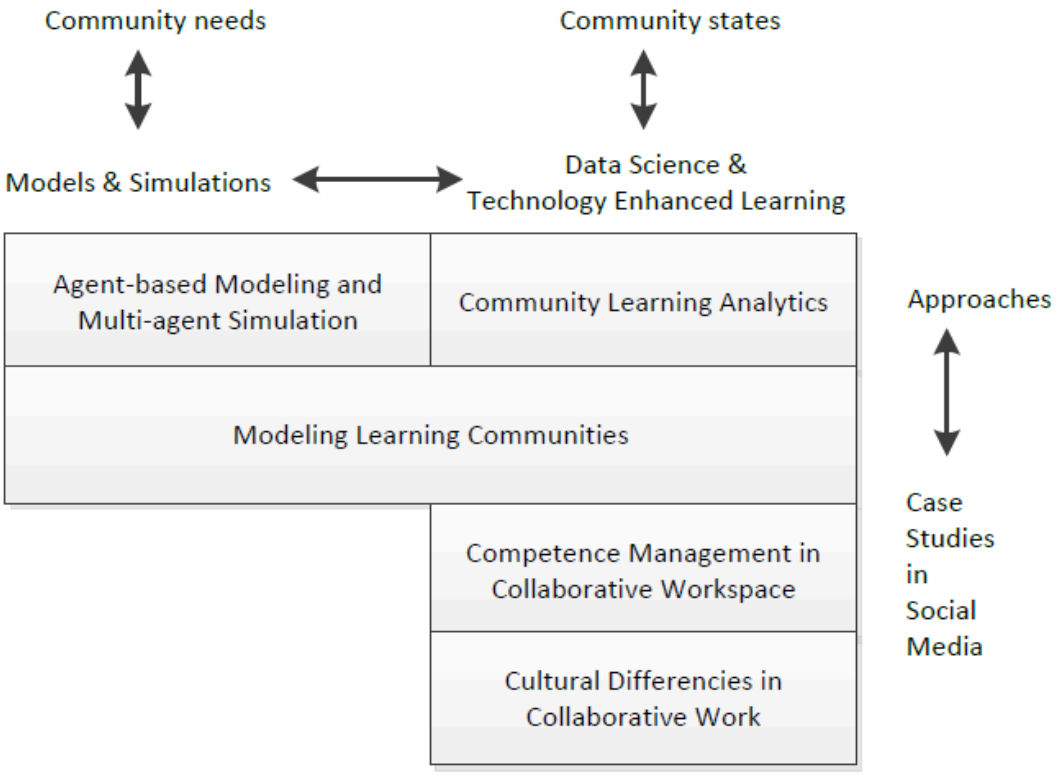


Figure 1.1: Background methodology, approaches and applications for this work

Figure 1.1 depicts our main goals (community needs and states), the research areas and their specializations — the methodology of the work on the top and with applications on the bottom that we designed and executed within this work. Since the whole process of accessing community needs consists of investigating communities and their modeling, I firstly apply the methodology to an online learning forum where the whole process has been performed. Community data have been collected, communities analyzed, classified, modeled and simulated. Such a process of accessing community needs has been performed continuously to capture changes. This experiment was established in the context of the ROLE project[3] that aims to support self-regulated learners with tools that allow the learners to independently direct their learning. So community models detected in the forum can indicate communities suitable for self-regulated learners (Zimmerman, 1990) according to their requirements and goals.

Learning communities in forums operate with discussion artifacts usual for other media. While other media provide other functionalities that enhance collaborative

[3]Responsive Open Learning Environment EU FP7 IP project `http://www.role-project.eu/`, Last access on 20.05.15

work such as collaborative editing or a collaborative task. Within the TeLLNet project[4] we investigate collaborative projects created on the eTwinning portal[5], the collaborative space for European teachers. eTwinning, similar to other collaborative workspaces, allows to organize projects and collaborate within them. To prove the broad applicability of our methodology in monitoring and analysis we investigate learning communities of eTwinning that are groups of teachers that collaborate in projects. Furthermore, we estimate competences of teachers and their peers in communities according to preferences of policy makers (European Parlament and the Council, 2006) and requirements for workers to enhance self-monitoring and self-reflection metacompetencies (Cheetham and Chivers, 2005).

Test beds of the ROLE project include learners belonging to different countries and cultures. To design an appropriate learning environment differences in process of learning caused by culture differences have to be considered (Uzuner, 2009; McLoughlin and Oliver, 2000; Gunawardena et al., 2003). Since Wikipedia includes two hundred and eighty eight language projects[6], these provide a foundation for the research of learning communities within different cultures. Therefore, we take 13 Wikipedia language projects as an example to investigate informal learning communities from different cultures, where peers are collaborating with each other through editing Wikipedia articles. We find differences in learning interactions and activities of learners that can be used to enhance design of digital learning environments used in particular countries.

1.3 Outline of the Dissertation

This work has an interdisciplinary character since it applies computer science approaches to learning science entities, i.e., learning communities. Outcomes of the approaches detect community states and community needs with regard to learning theories' interpretation of learning communities. The following chapter starts with the related work in the field of learning science as it is one of the pillars of this work. The chapter illuminates development of learning theories and considers theories that emphasize learner activities, their environment and collaboration. After that I observe applications that monitor, analyze and model learning communities and other communities in social media. In the end, I review information modeling approaches suitable for this work.

Chapter 3 introduces the methodology how we materialize community needs by the community model creation process. Each phase of the process is described with corresponding technologies. The data management solution is introduced in Chapter 4

[4]Teachers' Lifelong Learning Network `http://www.tellnet.eun.org/web/tellnet`, Last access on 15.10.2014

[5]The eTwinning portal `http://www.etwinning.net/en/pub/index.htm`, Last access on 20.04.15

[6]The Wikipedia article about Wikipedia `http://en.wikipedia.org/wiki/Wikipedia`, Last access on 9.05.2015

that describes the Mediabase cube. It explains how the Mediabase cube is designed and shows examples of its usage.

Discussions between learners are a common reason to organize informal learning communities. I approbate my methodology by modeling informal learning communities in forums in Chapter 5 where we realize all phases of community model creation. Furthermore, I concentrate on special cases of informal learning communities. In Chapter 6 I describe how we monitor and analyze learning communities at workplace and community stakeholders receive information on their own and their peers' competences. In the last case study, in Chapter 7 differences of learning communities that depend on learner culture were investigated applying monitoring and analysis of Wikipedia article editing.

Finally, Chapter 8 discusses lessons learned in this work and provides an outlook on the field of modeling of informal learning communities in online social media.

Chapter 2

Background and State of the Art

Social media caused a tremendous improvement in learning environments since they provide means that learners can use to share information, consume learning resources and communicate with other learners. Social media for learning allow to learn anywhere, anytime and anything. The ease of creation of learning resources and communication with other peers facilitates emergence of numerous informal learning communities that share common goals and solve issues together.

To avoid ambiguity, in this chapter I will first observe development of the role of learning communities in learning theories. I will define the meaning of community modeling for the thesis and review the research in the area of monitoring, analysis and modeling of learning collaborations and communities. Having introduced the shortcomings of the research, I discuss information modeling and modeling approaches that indicate a solution for modeling of learning communities and help to discover community needs.

2.1 Context of Learning Theories

Classical learning theories are divided into behavioral, cognitive and constructive theories. The *behavioral* view on learning originates from Skinner (1954). He conceptualized a 'teaching machine' that guides learners to come to the right decisions through changes in a learning environment caused by the teaching machine. The learning environment prescribes knowledge and interaction consequences of users to come to correct solutions, in other words interactions of learners are shaped. There is no room for learner creativity, experimentation, conceptual thinking and reasoning. Nevertheless, behavioral approaches are relevant in learning environments for small children that have no or few competencies in guiding their learning processes and constructing knowledge on their own (Zimmerman, 1990).

Pask and Scott (1972) focused more on learners than on learning environments, paying attention to their cognitive processes. They considered learners as individuals respecting their learning styles. Even though, cognitivism neglects learner abilities in

knowledge creation (Ravenscroft, 2003) and lacks motivational, emotional and social aspects of learning (Rey, 2009).

Piaget (1973) had a cognitive and constructive view on learning. His supporters, the *cognitive constructivists*, support the idea of learning by discovery. New knowledge is assimilated to existing situations or existing knowledge is accommodated to new situations. For instance, learners can create their own world using the programming language LOGO (Papert, 1980), but teachers found that learning by discovery as highlighted in cognitive constructivism is not enough. Learners need support in interactions (Ravenscroft, 2003).

Social constructivism emphasizes the role of interactions for learning and states that companionship has a great impact on a learning process. Vygotsky (1934/1986) claimed that a learning child gets a better outcome in collaborating than alone. Moreover, he named collaborations as a trigger for intellectual development of collaborative peers. Collaboration makes learning more efficient.

Another learning theory that pays attention to the role of society in learning is the *social learning theory* (Bandura, 1971) called later *social cognitive theory* (Bandura, 1986). The name was changed since Bandura (1986) stressed the role of cognition during learning processes. Bandura (1971), independently from Vygotsky (1934/1986), emphasized the importance of society in learning, where a learner becomes experienced from behaviors of her peers.

2.1.1 Networked Learning

Learning has social sources (Palinscar, 1998). This idea was supported by many works that emerged in the late nineties. For example by Wenger (1998), who defined the Community of Practice theory based on three pillars. CoP members understand and accept community culture and are able to contribute to their CoP — their *joint enterprise*. This dimension depends on goals, norms and policies of communities. The collaboration of the members and trustful attitude towards each other is pivotal for the *mutual engagement* dimension. The members establish different kinds of artifacts or use resources, such as tools, stories or styles that define the *shared repertoire* dimension. These three dimensions define the boundaries of communities of practice. Learning appears when CoP members expand CoP boundaries and extend shared practices of the CoP.

Wenger (1998) focuses only on the description of cooperations between community members and cooperation conditions such as rules and policies. While collaborations according to Dillenbourg (1999) initiate learners' *cognitive processes*. He represented some phenomena like reasoning according to information acquired from social interactions, appropriation or interpretation of facts, and mutual awareness about peers' knowledge. Results of these *cognitive processes* are *cognitive effects* that were of interest to many learning theories. In *reflective learning* Boud et al. (1985) described cognitive effects as *outcomes* while according to self-regulated learning (Zimmerman, 1990) peers produce cognitive effects as they reason about their learning processes.

Vygotsky (1978) marked these cognitive effects as triggers for learner development.

Although all these works are talking about relevance and importance of cognitive processes, their analysis is complex. Even though, some works investigated cognitive processes without involvement of medical procedures. Stahl (2006) and Jäger (2002) explained a circulation of knowledge between a community and an individual that involves cognitive processes. Jäger et al. (2008) defined as well a role and operations of media for the circulation. His transcriptivity theory specifies three processes: *transcription*, *addressing* and *localization*. Based on pre-scripts — original information or personal experience (Wenger, 2000) — learners create *transcripts*, i.e., adapt information using media for a particular circle of people — a community. The *addressing* process arouses particular community group's interest in shared knowledge and the *localization* process localizes the knowledge according to community practice and policies.

2.1.2 Self-Regulated Learning

Investigation of learning processes of an individual learner is required for correct estimation of learning processes in communities and networks. Zimmerman (1990) called those learners that guide their learning processes individually self-regulated. He worked out the *cycle of self-regulated learning* (SRL) consisting of three phases: *forethought*, *performance* and *self-reflection*. In the *forethought* phase a learner works with beliefs and experiences to understand her goals and orient herself. She considers her a-priori skills and plans a learning process. In digital environments such a learner sets or updates her profile, e.g., what she has learned, what was her progress before, what kind of learning materials she had used, what her weaknesses and strengths are (Nussbaumer et al., 2011). She plans future steps, decides which topic to learn and how to learn (*self-instruction, self-efficacy, self-motivation*).

In the next phase the learner looks for learning resources and materials to follow her learning goals. Moreover, in this phase she may start to reflect on a learning topic: finds pros and cons, answers related questions by other learners (*self-evaluation*, *self-improvement*).

In the *reflection* phase the learner reacts on her progress and learning success. She supports other learners in a topic by answering questions of varying complexity. Based on her learning interests, she decides about her future plans and learning goals (*self-instruction*, *self-evaluation*, *self-motivation*).

2.1.3 A New Era for Learning Communities

Since the rise of the World Wide Web numerous social media services appear that allow learners to gather in communities and discuss their issues. Social media is broadly used as an environment for informal learning since gathering and sharing information in social media is accessible anywhere and anytime. Cross (2007) explained the term *informal learning* as improvised learning with no schedule or place with physical

frames where learning should happen. While a path in formal learning is predefined like a ride with a train, a path in informal learning is chosen by a learner like a ride in a car, although the goal of both rides can be the same. Social media make it possible to share and collaborate anywhere, across the globe, at any time and independently from an institutional or organizational context. Many online learning environments, such as Coursera[1], EdX[2], the Khan Academy[3], and LiveMocha[4], propose learners to guide their education themselves. Furthermore, social media facilitate collaborations between learners. Theories of networked learning such as (Dillenbourg, 1999) and (Wenger, 1998) serve as foundations of numerous works about learning in social media.

Facilitating learning from one side, social media produce tremendous amounts of data that overwhelm attempts of community stakeholders to understand communities. Social software exploited by many social media to support community users can help to solve the issue. So market-leading social networks such as Facebook , Twitter, LinkedIn recognize the importance of communities for users fidelity and develop recommender algorithms to find friends and communities that make users adhere to the social media. Before developing recommender algorithms or any other algorithms, scholars used to model users (Brusilovsky, 2001).

Bandura (1971) already emphasized the role of modeling for learning. User modeling allows learners to learn from the mistaked made by others, adopt successful behaviors, and mimic activities to fulfill their goals. Recent specifications such as IMS Learning Design[5] and Grapple Core[6] allow to model users in learning environments in an interoperable way. But these models are either limited to formal learning or pay no attention to the presence of communities (Derntl et al., 2014). While echoing Bandura's attitude to modeling, community modeling can serve for similar purposes as user modeling, for example, for observing positive and negative experience. Community models can include information about the behavior of all community members, interactions between these members, patterns of media usage, member and community types. Similar to learners that observe models of others and imitate successful models, communities can learn from other communities by observing their models and reproducing them. If a community suffers from some issues it can mimic another community that managed to overcome similar issues by analyzing a model of another

[1]An education platform Coursera `https://www.coursera.org/`, Last access on 23.07.2014

[2]An education platform EdX `https://www.edx.org/`, Last access on 23.07.2014

[3]Non-profit educational organization the Khan Academy `https://www.khanacademy.org`, Last access on 23.07.2014

[4]Free-online language learning platform `http://livemocha.com/`; Last access on 24.02.2015

[5]The homepage from the IMS Learning Design `http://www.imsglobal.org/learningdesign/`, Last access on 10.12.2014

[6]Grapple Core user modeling format `http://wis.ewi.tudelft.nl/rdf/grapple-core.owl`, Last access on 10.12.2014

community while successful communities can share their experience using their models.

2.2 Modeling Learning Communities

In the following, I review existing solutions that model learning communities, collaborations of users or do any investigation that facilitates modeling, e.g., collecting or storing data. Furthermore, I highlight the solutions' benefits and flaws. Additionally, I review modeling approaches that support awareness of community stakeholders about processes inside a community and provide the stakeholders with additional and sufficient information that can help them to lead communities to the states where learners acquire knowledge efficiently.

2.2.1 Terminology

One of the ways of expressing information about principles, premises and objectives is modeling. Models show the understanding of subjects and create abstractions of the complex world, its description and analysis. A model can emphasize some parts of the complex world with more details while omitting other parts (Jeusfeld et al., 2009).

Models are used in many areas: for building a car or a machine, for specifying the structure of a language or for explaining a painting style. These models can be used as examples or starting points for creating new machines, languages and painting styles.

Community modeling is influenced by user modeling that is broadly used and applied in learning environments. Since the 80's, user modeling has evolved from overlaying models that store the assessment of user knowledge in each relevant item (VanLehn, 1988) over models based on collections of learner knowledge and interests categorized with the help of taxonomies and ontologies (Middleton et al., 2004) to time-dependent models that store relationships between previous actions and outcomes of learners (Mayo and Mitrovic, 2001). User modeling approaches create user models that are later utilized for personalization, adaptation, and recommendation in learning environments (Brusilovsky, 2001).

To create community models it is not enough to operate with user models because a community is not just the set of its users. Using only this set, one can hardly observe progress and success of the community since the user models represent individual characteristics. Success of learning processes in communities can only be objectively estimated by investigating the whole structure of a community where each item such as a learner, a technology or a community play a role.

Soller et al. (2005) specified two types of tools that utilize user models. *Mirroring* tools for collaborative learning help to observe learner progress in a learning task. User models provide an abstract view on individual interactions but not interactions of a whole group, thus it is possible to estimate discourse and knowledge of each individual learner but not of the whole group. Collecting user models together we can

estimate community knowledge and discourse but we can not investigate the process of knowledge evolution and the flow of information in a community. Community models estimate not only the knowledge of the whole group, but the learning progress and the activity of each member compared to others. For community modeling the mirroring tools have to observe the progress of a whole learning community in practicing or achieving a community learning goal.

The other type of tools, *guiding* tools (Soller et al., 2005), inform how to moderate learning to achieve a desired state for learners while guiding tools for communities can provide information required for achieving a particular community state such as an innovative community state (Petrushyna et al., 2010). Mirroring tools usually feed guidance tools with required information.

In the following, I will review existing solutions that mirror and guide students during collaborative learning in learning environments or social media. Mirroring and guidance tools firstly collected data about learners and communities, then stored it, analyzed and used outputs for modeling.

2.2.2 Monitoring and Storage of Learning Community Data

Here I discuss solutions that collect learning community and social media data. After that I present existing approaches to data storage of learning community data.

2.2.2.1 Monitoring Learning Community Data

Monitoring of learners and their communities is achieved if scholars use Learning Management Systems or other learning environments whose data is accessible to their stakeholders (Florian et al., 2011; Verbert et al., 2012; Arnold and Pistilli, 2012). Otherwise activities of learners have to be tracked using other tools such as Greasemonkey[7] (Dawson, 2010) and (Macfadyen and Dawson, 2010), or preliminary installed tools or virtual machines on personal computers (Wolpers et al., 2007; Scheffel et al., 2011). Commercial alternatives track user activities, such as Wakoopa[8] that monitors social media and RescueTime[9] that analyzes user activities and differentiates between working tasks and fun. All these approaches collect data about user activities but fail to collect data about community activities.

Another way to reach data is to monitor activities of learners that use learning services on a server that include additional services for monitoring. The ROLE[10] middleware includes an observation mechanisms, where the MobSOS service (Renzel et al., 2008) captures calls of services from clients and the CAM service monitors user activities in widgets (Govaerts et al., 2011). The data is stored in the MobSOS and the

[7] A Firefox extension that customizes the webpages `http://www.greasespot.net/`, Last access on 14.08.14

[8] Behavioral data collecting service `http://www.wakoopa.com`, Last access on 14.08.14

[9] Time management service `https://www.rescuetime.com/`, Last access on 10.12.2014

[10] EU IP FP7 project `http://www.role-project.eu/`, Last access on 24.02.2015

CAM repositories correspondingly and it is available to ROLE stakeholders only.

Some social media provide free access to their data through APIs, like Facebook and Twitter APIs, though their access is limited[11]. The problem of data retrieval from social media can be solved by following *Information Retrieval* approaches. These approaches retrieve statements from social media texts by analyzing their web pages and particularly Document Object Model (DOM) trees (Insa et al., 2013; Song et al., 2013; Pappas et al., 2012). Other works are based on regular expression rules (Adelberg, 1998; Lin and Ho, 2002; Vieira et al., 2006; Pappas et al., 2012). Furthermore, Pappas et al. (2012); Song et al. (2013) and (Uzun et al., 2013) used multiple measures to find the main content in HTML documents. All these approaches concentrated on detecting web pages' content only with relevant information avoiding, for instance, advertisements. Furthermore, the approaches are content-oriented and payed no special attention to users that create content of web pages and their collaborations with other users.

Tracking of social media for the learning purpose is possible using Web crawlers. Klamma et al. (2007) presented a watcher, a Perl script that emulated a browser and collected data devoted to learning. Watchers developed according to the same concept collected data from forums (Krenge et al., 2011; Petrushyna et al., 2011; Hannemann and Klamma, 2013), mailing lists (Klamma and Petrushyna, 2008), and wikis (Klamma and Haasler, 2008a,b; Petrushyna et al., 2014a). A commercial crawler by Salesforce[12] tracks 650 M different sources but it focuses on data extraction of commercial information about products, their consumers and communities.

2.2.2.2 Data Management Solutions for Learning Community Data

Storage of collected data has to be efficient to provide a quick responses for queries from mirroring and guiding tools.

Data Sharing Initiatives: Several initiatives collected different data sources and mapped them to one schema. Reffay and Betbeder (2009) included data repositories from K-12 mathematical learning environments and shared the collected data. They developed an XML-based formalism to represent any collected data using the same schemata. They considered both actors and the environment with tools used during a pedagogical scenario. Furthermore, communication tools were described with great precision as well as activities of learners. Another initiative for sharing educational data, the PSLC (Pittsburgh Science of Learning Center) Datashop, was described by Koedinger et al. (2008). It included any electronic artifact connected to a course or study and stored interactions with online courses and intelligent tutoring systems. Student actions were labeled as correct or incorrect and categorized according to hypothetical competencies. The Datatel initiative used a Learner Action Model to describe

[11] for example traffic limitations for the Twitter API

[12] Social Studio `http://www.salesforce.com/marketing-cloud/features/social-media-marketing/`, Last access on 10.12.2014

data sources taken from different Technology Enhanced Learning projects and initiatives (Verbert et al., 2012). The sources included a diverse set of actions and resources with some context information, though information about learners were limited.

All of these initiatives have been collecting data sources and mapping their data to the given schemata. The schemata allow to reuse and redesign learning objects across various learning environments but omit consideration of communities. This makes modeling of communities complicated and time-consuming.

Data Models for Interaction Traces Storage: Renzel et al. (2008) proposed the MobSOS Communication Monitoring Data Model where learning activities of the ROLE Personal Learning Environments were stored (Renzel and Klamma, 2013). The data model had the purpose to support information systems management and therefore it included entries like sessions and requests of users to services. Entries like learners and communities need to be deduced from the collected data and therefore require some additional effort.

The Contextualized Attention Metadata (CAM) (Wolpers et al., 2007) investigated user activities in different applications. Using events such as activities, one can detect whether documents are connected to the same affair. The schema of CAM, that included neither users nor communities, was designed to store events and actions but not to store data devoted to communities.

Settouti et al. (2011) collected interaction traces from different resources like log files from learning environments. Their Trace-Based Learner Modeling Framework organized learners' profiles in an RDF-based representation of knowledge models. But the framework captured only interactions of each learner and not interactions within a community.

In the center of the Karam et al. (2012) approach was a user in social media such as Facebook or Twitter. The authors considered not only traditional information for user modeling like age and gender but also a representation of a user in one or more communities. Their data model included users, their social activities and profiles in social network sites. The approach is promising, as it considers the usage of social media but the model lacked a connection to community-based learning where learners are not only sharing resources in a community and interact in social media but as well refine and extend their knowledge by consuming information shared by others.

Suh and Lee (2006) collected data using a monitoring agent and stored the data in a so-called workspace. The authors defined three types of interactions they differentiated in the workspace. These were *participant to participant*, *participant to resource* and *participant to learning* interactions. Such a structure helped the authors to view relationships between learning peers, learning materials, and learning processes and to follow the outcomes of learning. Even though, they missed a *participant to community* interaction that has a great influence on learning in communities (Vygotsky, 1934/1986).

The objectives of community modeling are to estimate community success and efficiency, to suggest relevant resources considering collective actions and to provide

awareness about community states and needs to community stakeholders. A data management solution for community modeling must emphasize both users and communities since users conclude communities and define community goals, topics and perspectives for the future existence. Therefore, in a possible data management solution a community has to be an entity.

2.2.3 Mirroring Tools for Learning

Here I describe tools that let learners view the information about their learning processes and additional information that may include simple analysis and visualization.

Meerkat-ED is a toolbox for analyzing student collaborations (Rabbany k. et al., 2012). It creates networks based on students' collaborations and visualizes them. Furthermore, Meerkat-ED proposes to visualize a set of terms appearing in a selected thread as a network of topics. By selecting terms in the visualization one can get information about a user that used the terms. The toolbox can serve to support both teachers and learners in the self-monitoring task (Zimmerman, 1990). Moreover, it only collects and analyzes students' collaborations but not student communities and their progress.

Florian et al. (2011) let users view information about activities and competencies in Moodle[13] in self, peers or class panes. Depending on their role, users had access to different perspectives. The authors provided learners with panes that included information about learners' competencies and competencies of their peers. In the approach a community is defined according to institutional frames while it can include a few informal communities.

Upton and Kay (2009) exercised similar perspectives for student groups that were active in different media and visualized data in the *Narcissus* application in different views such as a group view. The views allow to compare activities of students in different teams and activities that appeared in various media. The investigations of communities were limited to the student groups only that were organized in formal settings. Upton and Kay (2009) exercised similar perspectives for student groups that were active in different media and visualized data in the *Narcissus* application in different views such as a group view. The views allow to compare activities of students in different teams and activities that appeared in various media. The investigations of communities were limited to the student groups only that were organized in formal settings.

2.2.4 Guiding Tools for Learning

One of the categories of support for group activities during collaborations is *peer interaction* (PI) support (Magnisalis et al., 2011). PI support aims to improve in-group

[13] A learning platform `https://moodle.org/`, Last access on 30.12.2014

communication and develop a common understanding between group members. Another category, called *domain specific support*, adapts learning environments according to user knowledge and helps users to acquire information more efficiently. In the following, I focus only on PI approaches while some of them are with elements of domain specific support.

2.2.4.1 Discourse Analysis

Learners in communities are usually interacting with each other using different artifacts of media. Many artifacts include texts like posts in blogs, comments, replies in threads and so on. Discourse analysis estimates knowledge of individuals and communities and can characterize the shared repertoire of a community (d'Aquin and Jay, 2013). It indicates community goals (Krenge et al., 2011) or specifies directions of community knowledge changing (Dascalu et al., 2010).

Chat conversations of students were analyzed by Dascalu et al. (2010) who investigated knowledge exchange between individuals and communities explaining it using the cognition theory of Stahl (2006). Their polyphonic conversation analysis of CSCL[14] scripts revealed how knowledge is transferred from one peer to the other and what forms it obtained. It emphasized divergent and convergent interactions between learning peers. An application based on polyphonic conversation analysis was positively evaluated by tutors of lectures that included chat rooms with investigated conversations. Even though, the authors addressed formal communities only and analysis of community chat conversations can not be used for analysis of other chats because no patterns were defined.

Generally, analysis of texts may help tutors to mediate discussions. In approaches such as (Scheuer and McLaren, 2008) the authors created machine-learning classifiers that defined categories like "reasoned — claimed" or "contribution — contraargument" and additionally defined positive or negative discussion situations. Ferguson et al. (2013) created a different classifier that detected various exploratory dialogues like critical views, evaluations, explanations and others. These approaches concentrated only on analysis of generated contents from individuals but did not consider community discussions where it will be useful to know how critical communities are or if there is a match between a peer and a community in knowledge or ways of discussions.

2.2.4.2 Activities

Many approaches estimated only user activities, for instance, approaches using the Context Attention Metadata (CAM). In (Scheffel et al., 2011) CAM was used for collecting information about learner activities on personal computers and finding error patterns based on the sequence of activities.

[14]Computer-supported cooperative/collaborative learning

Florian et al. (2011) created activity-based learner models of students in Moodle. The models provided learners and teachers with information about their learning processes, acquired competences, learning processes and competencies of peers and a whole classroom.

Settouti et al. (2011) collected user interactions and described learner profiles continuously through provided ontologies using RDF representations. User profiles and their activities were stored in RDF-based repositories. But the approach pays no attention to the role of a learner community and learning resources that would be useful for community modeling.

2.2.4.3 Learning Groups

The multi-agent distributed environment for collaborative learning I-Help captured interactions between learners and between learners and instruments as interactions of agents (Vassileva et al., 2003). Each user got an agent that helped to realize user activities. User models were created based on these activities. Any tool of the environment was as well represented by an agent. Additional agents like diagnosis and matchmaking agents had the task to follow all user activities and to find knowledgeable or suitable learning partners. This multi-agent solution facilitated the simulation of learning situations in I-Help. Nevertheless, the approach did not consider a community as an agent while including it in the environment would make it easier to investigate learning processes in communities.

The approach of Suh and Lee (2006) included a facilitator agent that generated advice for learners according to their communication activities, types of discussion messages, group interaction patterns and group cohesiveness. The agent informed learners not only about individual progress but as well about the progress of learning peers. Furthermore, the approach detected patterns of group interaction that described different group interaction processes though important actors for learning communities are not emphasized as well as no attention is given to the medium used.

Perera et al. (2009) operated with data from a learning environment that included different media. The authors detected clusters of users according to information such as the average number of collaborations, the amount of content produced on average, the average number of tasks users accepted to do, the average number of tasks users created and some other characteristics. The results showed activity patterns of stronger and weaker students as well as patterns of stronger and weaker groups. This approach performed a simple analysis of data produced by learners in formal settings.

Martinez et al. (2011) identified and modeled collaborative situations where students communicated orally and interacted using a multi-display environment. The authors tracked groups of student activities on multi-displays and recorded their discussions. They constructed classification models of collaborative activities that appeared in different media. The models can help teachers to estimate the degree of collaboration among students. But no formal representation of patterns was presented in the work as well as no analysis of content of user interactions was considered.

Solutions in (Upton and Kay, 2009) and (Florian et al., 2011) focused on learning groups as well though these groups were formally organized and authors concentrated on investigations of activities of learners and sharing of these activities with other learners.

2.2.4.4 Collaborations

While investigating collaborations in communities one can reveal interesting and important facts. Considering students as nodes in a graph and interactions, such as doing homework together, as connections between the nodes one can investigate the graph using Social Network Analysis (SNA) (Wasserman and Faust, 1994) and detect tightly connected groups — communities — using community detection algorithms such as the Girvan-Newman community detection algorithm (Newman and Girvan, 2004).

Changes of students' roles and teachers' roles indicate changes in the structure of learning groups. Marcos-García et al. (2009) used SNA and detected some patterns of students such as a *student coordinator* or an *isolated student* and changes of patterns in groups. The solution by Suh and Lee (2006) included computational models that estimated social interactions of each user (ego-centric networks) and group collaborations. Their facilitator agent informed how cohesive the community was based on the computational models, how interactive a user was or which dialogue patterns (*question and answer*, *application*, *agreement*, etc) community content included. The approach is one of the first that combined content and interactive analysis for collaborative learning.

2.2.4.5 Social Media and Massive Open Online Courses

With the rise of the World Wide Web social media became popular to use as an additional instrument for collaboration. Wikis, forums, Facebook and Twitter are full of posts, videos and comments that are created for the purpose of learning.

Both Karam et al. (2012) and Abel et al. (2011) modeled users based on their activities and properties that appeared in social media such as Twitter or Facebook. Abel et al. (2011) utilized generic user modeling formats like Grapple Core (Abel et al., 2009). The Semantic Web service of their framework, the U-Sem, enriched user profiles with named entities (Grishman and Sundheim, 1996) extracted from user texts. But the U-Sem framework concentrated on user profiles and left communities of users out of consideration. Instead, Karam et al. (2012) payed special attention to social life of users. In their data model they included communities and their networks. The authors considered as well relationships of users to the communities and networks. Such a perspective allowed to capture cases when a user was part of many different communities and networks. However, this approach emphasized activities of users but not artifacts users operated with, such as e-mails, posts, or comments.

MOOCs usually include a set of video lectures together with short quizzes and assignments. Such media like forums or chats are as well included in MOOC platforms

for supporting collaborative learning and providing a feedback channel. Some scholars conducted their research on learners in MOOCs. Anderson et al. (2014) followed behaviors of learners like watching videos and submitting assignments and defined patterns of these learners. Furthermore, the authors investigated activities of learners in forums where they found that activities of video watching depend on defined patterns: a learner who has been doing all assignments and watching all lectures is much more probable to be active in a forum than others. In another MOOC Kizilcec et al. (2013) defined clusters of students according to their participation in the course. The authors investigated as well the forum activity of learners devoted to different clusters. Both works found that patterns of video viewing and submission of assignments could predict forum activity. But no abstraction of clusters or patterns was created so that information could not be used for other MOOCs or learning environments.

2.2.5 Modeling of Collaborative Learning

In Technology Enhanced Learning there are many modeling approaches that modeled learners and collaborative learning. The type of modeling differs: there are models based on specifications (Nodenot and Laforcade, 2006; Settouti et al., 2011; Abel et al., 2011) and conceptual models (Soller, 2001; Suh and Lee, 2006; Kleanthous and Dimitrova, 2007, 2010).

Soller (2001) modeled effective collaborative learning teams. He emphasized the need of participation in a group by all group members to reach the best result on a topic. The author considered a number of important dimensions in his Collaborative Learning Model that can be used for a system to adapt and support learning teams. Some of them are common understanding, shared goals, helping each other and benefiting from each other. The model consisted of indicators for effective collaborative learning and was used to design and develop tools for collaborative learning though the model did not consider communities and learner activities.

Suh and Lee (2006) detected and represented several community patterns based on activities of all students of communities. But their models included only users and did not emphasize roles of other actors such as media or community content that can help community stakeholders to estimate situations in communities better.

One of the few solutions that model communities is represented by Kleanthous and Dimitrova (2007) that was refined later in Kleanthous and Dimitrova (2010). For modeling, they considered information about individual users, relations between these users, and user knowledge in learning topics. The authors chose organizational psychology as a background for community modeling. They emphasized transactive models that describe relationships between individuals and their knowledge. Moreover, they considered shared mental models that described aspects similar to *shared repertoires* of Wenger (1998). Another aspect for community modeling was cognitive consensus that correlates with *mutual engagement* of Community of Practice and *symmetry* of Dillenbourg (1999). Kleanthous and Dimitrova (2007) modeled the flow of knowledge between individuals through transactive models and shared mental models

but missed the explanation of the flow of knowledge between communities and individuals. The continuation of their work resulted in community knowledge sharing patterns (Kleanthous and Dimitrova, 2010) that consider only communications between individuals in communities. Due to their formal representation the patterns can be used for other communities in other environments. Both works concentrated on user activities but learning resources were out of the research scope.

Using the IMS Learning Design (LD)[15] it is possible to model users, their roles and activities. Paramythis (2008) and later Derntl et al. (2014) investigated usage of IMS LD for adaptive collaboration support and support of learner interactions. In both works the IMS LD was criticized because of its inflexibility. Paramythis (2008) mentioned the possibility to attach roles to users only at design time while changes of these roles during runtime are impossible. Most runtime players offer only two collaborative services: the send-mail service and the conference service while other external or internal services need either the extension of the IMS LD or the service description schema (Derntl et al., 2014). Furthermore, learners can not interact with each other unless these two services were initiated by a learning designer. The IMS LD proposes the useful feature of interoperability that allows to use learning design for different environments and contexts but it still requires extensions that allow user interactions easily and increase the amount of information about user interactions.

A community modeling approach needs to capture users, their communities, social media and dependencies between these. All these entities and their dependencies allow to get a full picture about communities.

2.3 Information Systems Background

In this work, I appeal to information systems, particularly social software, for the support of learning communities. In this section I introduce information modeling essentials and social media as an information system with a number of different influential actors that constitute a user world (Jarke et al., 1992) that has to be emphasized to define community states and needs. Afterward, I describe some modeling approaches and discuss in detail an agent-based and goal-oriented approach used for social media modeling.

2.3.1 Information Modeling Essentials

Information modeling aims to represent applications and their environment. Modeling is pivotal for designing complex information systems as it includes a collection of instances that are used to describe an application, a collection of operations on the instances and a collection of constraints that define changes of instances' states.

[15]IMS LD `http://www.imsglobal.org/learningdesign/`, Last access on 23.05.2014

Information models can be divided into physical models, logical models and conceptual models (Borgida and Mylopoulos, 2009). The physical models focus mainly on implementation details, the logical models symbolize abstract models and hide implementations while the conceptual models include semantic terms that allow to model applications naturally and directly (Hammer and Mc Leod, 1981). Here and later I operate with conceptual models that I use to represent parts of the real world that are connected with applications. In the following, I further dig into the description of information models based on (Borgida and Mylopoulos, 2009) and describe existing modeling approaches.

Any information model of an application includes static aspects of the world like *individuals* (e.g., users, technologies, etc.), *classes* (e.g., newbies and experts), *subclasses* (e.g., experts in language learning), and *relations* (e.g., an expert *helps* a newbie). *Dynamic aspects* of applications are defined by a sequence of tasks that need to be performed to achieve a state. Another important dimension of information modeling, *intentional settings*, emphasizes things that individuals believe in and pursuit. The last *social settings* dimension considers organizational structures, roles, dependencies and group collaborations. Using static and dynamic aspects and intentional and social settings Mylopoulos (1998) defined a theoretical framework for modeling any information system, for instance, a social medium.

Social media, their users, artifacts they created, and communities they organized are *individuals* for information modeling. The individuals can belong to different *classes*. Social media users can belong to classes pre-defined by the social media, e.g., *admininistrators*, or discovered by applying data mining techniques (Fisher et al., 2006; Klamma et al., 2006c). Dependencies between individuals are denoted as *relations* or *relationships*, such as a dependency between a user and a medium where the dependency sets constraints on user activities and artifacts users can produce. Social media have a dynamic nature where a sequence of activities changes social media individuals (users or communities) — *dynamic aspects*. For instance, writing inappropriate content makes social media users unpopular in communities. Such users are candidates for excluding them from a medium. *Intentional settings* can include goals, subgoals and softgoals (Mylopoulos, 1998) of any individual of the medium. *Social settings* of informational modeling represent organizational structures, group collaborations and dependencies between individuals. Representation of user communities as well as their dependencies from each other or from the media is pivotal for social media modeling (Ahlqvist et al., 2008).

In the following, I shortly review existing modeling approaches and describe one suitable approach with more details.

2.3.2 Modeling Approaches

Jarke et al. (1992) defined four *worlds* that had to be considered for information modeling: 1) the objects an information system is about, the *subject world*; 2) the system itself with implementation details, the *system world*; 3) the environment where the

system functions, the *usage world* and 4) the process of software development of the system, the *development world*.

Modeling approaches such as UML or BPEL lack of *social* and *intentional* settings (Mylopoulos, 1998). Goal modeling approaches (van Lamsweerde, 2001) and (Rolland, 2007) introduced intentional ontologies but they missed *social settings* especially in connection of individuals to intentionality. *Social settings* of information modeling define the well-being of an individual through actor roles, dependencies from other actors, achievements of goals, and performance of tasks (Yu, 2009). Yu (1995) proposed the *i** modeling approach that highlights dependencies between actors and describes precisely the rationale of actors by performing some activities. The approach implements the *subject* and *usage worlds* and is useful for early requirements engineering that reveals goals and dependencies of individuals.

*i** stands for distributed intentionality that spreads on social networks of autonomous actors (Yu, 2009). It allows to emphasize different individuals, dependencies between the individuals with various types of these dependencies.

Short Introduction to *i

The *social* aspect of *i** emphasizes dependencies of actors from each other. Actors depend on other actors through resources, tasks and goals. It is beneficial for an actor to be dependent as it opens new opportunities for the actor (Yu, 2009). For example, if a user actor is connected to a medium actor and depends on it, the user can perform actions using the medium, e.g., spread news over many subscribers in a mailing list.

*i** aims to support not only social agents but as well to highlight their intentions. It highlights not an actual behavior but expresses *why* users perform particular actions. The intentions can explain why users choose one alternative over another one.

Strategic dependency models of *i** focus on connections between actors (Figure 2.1). A *depender* actor affects a *dependee* actor. Dependencies between actors are of different types: *resource, goal* or *softgoal* and *task* dependencies. An *i** model prescribes *goal* and *task* but cedes decisions according to goal or task achievements (*how to achieve*) to model customers. A *softgoal* dependency characterizes a quality dependency while for a *resource* dependency an artifact is important, e.g., the *Students* (actor) are dependent on the *Teacher* (actor) through the *Lecture* (resource dependency). In the case of a *resource dependency* a *dependee* demands a *depender* to maintain the resource. Some possible actor associations are when an agent *plays* a *Role* or when one actor is associated with another through the is_a or $part\_of$ generalization association.

Here I describe strategic dependency models but *i** provides a methodology to extend these by explaining rationale according to dependencies in strategic rationale models. In this work I deploy only strategic dependency models though the usage of strategic rationale models can be an extension of this work.

2.3.2.1 *i** Modeling of Social Media

Ferreira and Silva (2012) modeled a community using the *i** modeling approach. This work is an example of the usage of *i** for modeling dependencies between different

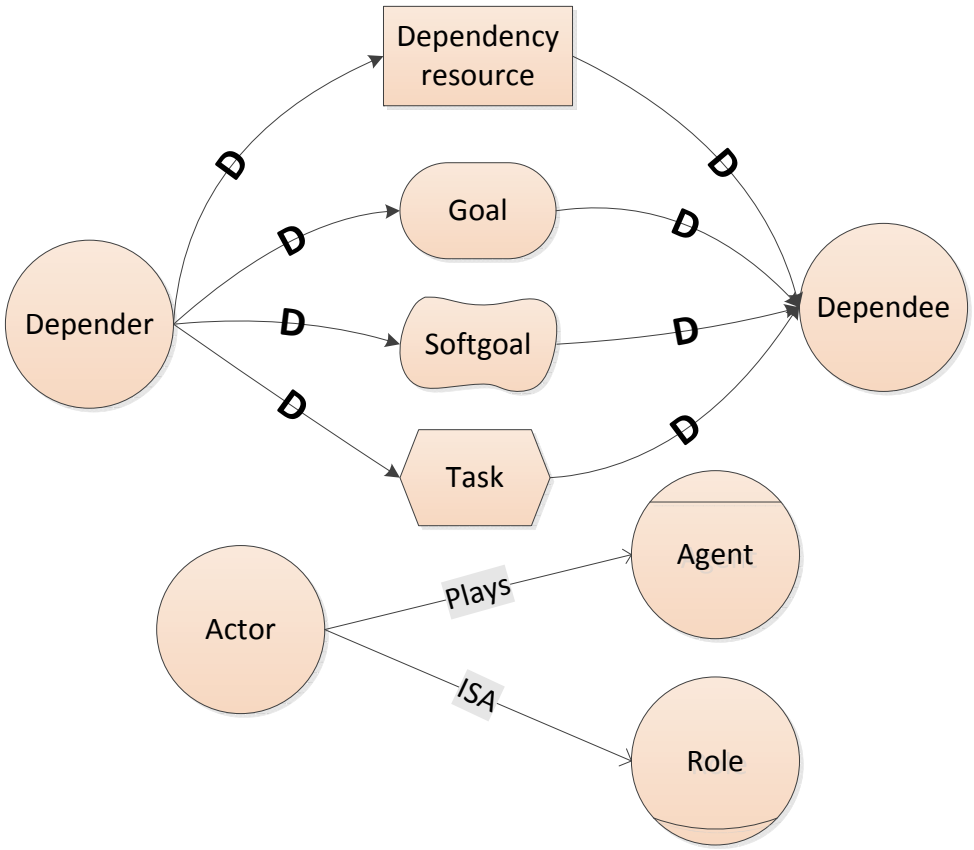


Figure 2.1: Different dependencies of actors and their associations

actors, human or non-human, in a social medium community. The authors defined activities that a community manager should perform in order to initiate interactions between community users. Later they described processes that appeared in a newsletter tracking system. The implicit connections between readers of a newsletter were defined by their interest in particular topics, their activities according to the newsletter like clicking a link or an image or forwarding a message. The authors described the environment of the newsletter tracking system together with important tasks and goals of newsletter readers and the community manager. However, the modeled community did not have a learning purpose.

Hilts and Yu (2011) created another *i** model of a social medium. It included users with special roles and their dependencies from each other and dependencies of users from a newly designed collaborative filter of the social medium. The filter aimed to support user goals and the *i** model aimed to prove the design of the filter.

In both works *i** models were created under assistance (not automatically), although creating *i** models automatically is possible if the data is freely available, which

is the case for social media. Current *i** tools focus mainly on user-driven model creation (Almeida et al., 2013), e.g., desCartes[16], iStarTool[17], TaoM4E[18] and jUCMNav[19]. OpenOME[20] and J-Pr*i*m[21] allows textual input to create models. It allows to prescribe a construction of *i** models that utilize Detailed Interaction Script (Grau et al., 2005) thus the tool expects users to assist by data gathering. One can create models using text commands using OpenOME but they have to be executed by humans. Modeling social media with *i** can be more efficient using a service that creates models automatically as soon as the required data is available. In such a case, models can correspond to current states of social media and their customers and define actual community needs that community stakeholders can react to immediately.

2.4 Summary

In this chapter I have reviewed the research on modeling learning communities. First of all, I explained the meaning of communities in learning theories. After that the chapter introduced *learning community modeling* approaches where either users and collaborations were modeled or models include only limited information about communities.

We have seen that there is a lack of modeling approaches that investigate learning communities considering both technological and learning-theoretical aspects. Furthermore, many approaches that collected community data specialized mainly on a small part of data about communities, such as community activities. Existing approaches that formalized community models focused on users and their interactions omitting either artifacts, or media as important actors in community models (Suh and Lee, 2006; Kleanthous and Dimitrova, 2007, 2010). Moreover, they missed to consider influence of communities on users.

Community stakeholders, such as community users, managers, and developers of community media, find it useful to know community states and needs that are emphasized in community models. Therefore, in this chapter, I reviewed information modeling and the *i** modeling approach that we use in the next chapters to create conceptual

[16]Design CASE Tool for Agent-Oriented Repositories, Techniques, Environments and Systems `www.isys.ucl.ac.be/descartes/`, Last access on 8.10.2014

[17]iStarTool home page `http://www.cin.ufpe.br/~ler/projects/istartool.php`, Last access on 8.10.2014

[18]Tool for Agent-oriented Modeling `http://selab.fbk.eu/taom/`, Last access on 8.10.2014

[19]Free, Eclipse-based graphical editor and an analysis and transformation tool for the User Requirements Notation, `http://jucmnav.softwareengineering.ca/ucm/bin/view/ProjetSEG/WebHome`, Last access on 8.10.2014

[20]Improved version of the Organization Modelling Environment `https://se.cs.toronto.edu/trac/ome/`, Last access on 8.10.2014

[21]J-Pr*i*m tool respects the Prim methodology that addresses *i** modelling from the process reengineering perspective `http://www.ideaciona.com/PhD/JPRIM/index.html`, Last access on 8.10.2014

models of learning communities in social media. The following chapter explains how information modeling can support learning communities. I describe the process of community model creation that results in automatically derived learning community models that specify both community states and community needs. Furthermore, we collect data of communities regarding different actors, human and non-human, that need to be mentioned in conceptual models of communities to clarify dependencies between community actors. Moreover, communities are analyzed considering CoP dimensions that trigger retrieval of information about communities sufficient for community modeling. The presented methodology has been used in several case studies where models of communities and more detailed analyses of communities have been performed.

Chapter 3

Supporting Learning Community Needs

In the previous chapter I have reviewed existing solutions for mirroring and guiding tools for collaborative learning and modeling approaches for communities and social media. Most of the described approaches failed to include communities as an entity type in their consideration. They collected and stored user activity data but not community data and analyzed data of learners that is not equal to data of communities. Moreover, no formal representation of learning community models was presented so far.

Searching for an appropriate representation of community needs and states we appeal to information modeling that can describe information systems such as social media together with their operations and constraints. Conceptual models that model social media naturally should include any actor that has an influence on the social media. Since we focus on learning communities that work with social media, the communities are emphasized in social media models since they belong to the usage world (Jarke et al., 1992). The conceptual models can specify community needs and states, for example using early-requirements modeling approaches such as *i** modeling.

Community stakeholders who are community members, community administrators, community analysts, developers of community information systems and many others, can benefit from community models. Observing the models, stakeholders have abstract views on communities. The models clarify important actors, their roles regarding learning theories and learner activities, and may indicate on issues or success in learning communities. Examining the models the stakeholders can detect solutions to the issues of their communities.

Communities are complex organisms that consist of many components such as users that initiate numerous actions. Therefore, we need a framework that defines a process of community modeling. In the following, I introduce a framework that continuously supports evolving communities by accessing community needs. The Architecture for Transcription, Localization and Addressing System (ATLAS) is the base of the framework (Klamma et al., 2006a) that emphasizes the need for tools (social

software) that measure, analyze and simulate communities. The tools provide information that let communities self-monitor and self-model themselves (Petrushyna and Klamma, 2008).

To recognize interactions between technological and human actors I adopt the ATLAS framework to the ATLAS i* model in Figure 3.1. In contrast to UML and other popular modeling approaches, *i** emphasizes agents and dependencies between actors such as goal, task, softgoal and resource dependencies (detailed descriptions can be found in Section 2.3.2).

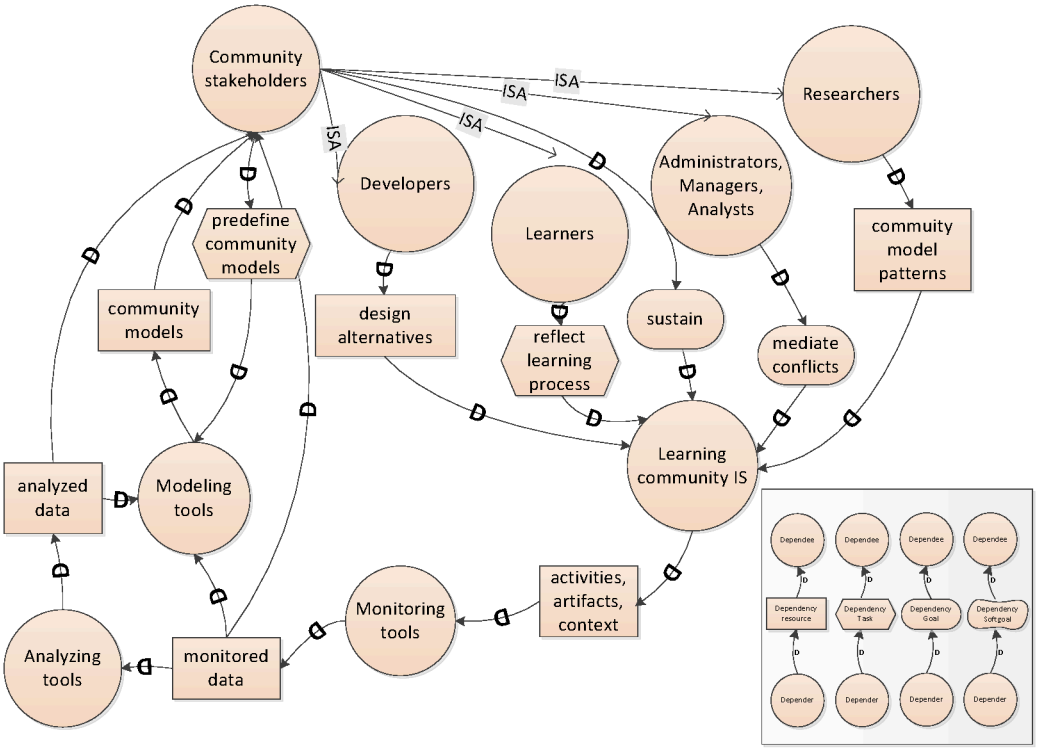


Figure 3.1: The *i** model of the framework for supporting community needs inspired by the ATLAS approach (Klamma et al., 2006a)

A *Learning Community IS (Information System)*, that is a social medium or a learning environment, gives access to *activities, artifacts* and *context* for *Monitoring tools*. I retrieved such tools in Section 2.2.2. The *Analyzing tools* actor depends on the *monitored data* and provides the *analyzed data*. I observed such tools in Sections 2.2.3, 2.2.4. The *Modeling tools* actor depends on both *monitored* and *analyzed data*. These tools are described in Section 2.2.5. They are responsible for *community models* that are the outcome of *Modeling tools*. *Community models* are abstract views on communities that show their structure, members with roles, human and non-human

actors of communities, and their dependencies and define types of communities. Resources of all these tools — the *monitored* and *analyzed data*, and the *community models* — are interesting for *Community stakeholders*, particularly for *Developers*, *Researchers*, *Learners* and *Administrators*. The *Developers* can provide *design alternatives* based on *community models* to support needs of communities. For instance, Hilts and Yu (2011) provided a design of a collaborative filter using *i** modeling to define user requirements in a social medium. The *Researchers* can find patterns of *community models* that indicate needs and possible solutions to issues of communities. In (Klamma and Petrushyna, 2010) we looked for competence gaps in communities involving patterns of community models. The *Administrators* can use *models*, *monitored* and *analyzed data* as estimations of community situations and *mediate conflicts*. Hannemann and Klamma (2013) opened analyzed data of communities for community stakeholders that estimated community health. Moreover, Kenett et al. (2014) assessed trust between community users using community models. *Learners* can reflect their learning processes based on the *monitored data* to self-monitor their learning processes. Operating with the data they become more active in collaborations with peers (Anderson et al., 2014) and become more active in reflection (Glahn et al., 2011). The

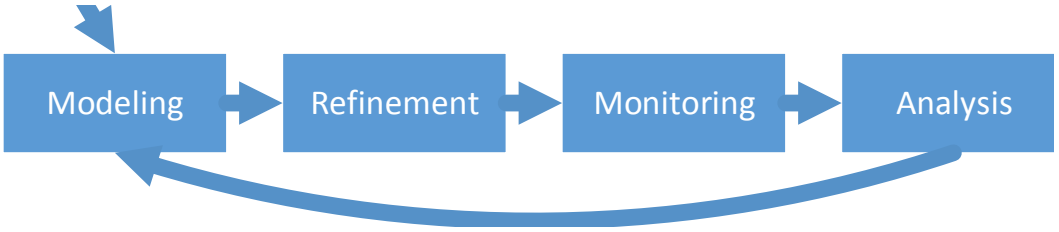


Figure 3.2: The process of community model creation

ATLAS *i** model reveals important actors that play a role in accessing and supporting community needs. It allows to construct a process of accessing community needs as depicted in Figure 3.2. Modeling is the first step of the process where any community can be represented by a general community model (Figure 3.3) or a model can be selected from the set of existing models (Figure 3.4, 3.5, 3.6). Doing so we avoid a cold start problem that appears because of sparsity or absence of data for modeling. If community stakeholders decide to consider their community as a special one such as *question-answer* or *innovative*, the stakeholders can refine their community according to the model they chose. In this refinement step the stakeholders can think about new roles and responsibilities for users, new topics and tasks for communities, and extensions or limitations for media. After that, communities can be monitored and analyzed. In the next iteration of the process after all steps are done community models are updated based on the output from the monitoring and analysis phases. The stakeholders can compare the current model with their vision and refine their community model if needed. To understand the relevance and efficiency of the refinement, the model can be simulated. After that, we can further monitor and analyze changes. The process of

accessing community needs has to be performed continuously since communities and their environment are changing and require continuous modeling (Jarke et al., 2008). In the following sections, I describe each step precisely, starting with the modeling step.

3.1 Modeling Learning Communities

In the previous chapter, I reviewed a number of works that modeled communities. Some of them selected a direction for the models, e.g., Soller (2001) viewed collaborations of learners under the prism of common understanding, Florian et al. (2011) payed attention to activities of learners, Kleanthous and Dimitrova (2010) focused on interactions describing them with the help of organizational theory while Suh and Lee (2006) considered social, emotive and cognitive factors to differ between learning groups. So what is a suitable learning-theoretical model for learning communities?

Earlier I have described theories that explained learning and the role of communities in learning. In this work I refer to learning communities as communities of practice (CoP) (Wenger, 1998) and I utilize *i** modeling approach for modeling communities in social media. These models can depict community states and needs since they include actors of communities and community environment, human and non-human actors, as well as dependencies between these actors. Models are either predefined by community stakeholders or defined after the monitoring and analysis phases. Before introducing three community models I explain a general community model that I take for granted to model learning communities in social media.

3.1.1 A General Community Model

To model a community in a social medium I consider the data model of social software (Klamma, 2010). In the model Klamma (2010) appealed to Actor Network Theory (Latour, 1999) that makes no difference between human and non-human actors by constructing a complex system. He emphasized five main actors: *medium, artifact, service, member* and *network* (Figure 3.8). I renamed some of the actors for the sake of clarity. The *service* is renamed to the *process* as I am interested in the processes activated by community users but not the services they use. Furthermore, the *member* is called the *agent* and the *network* is renamed to the *community* as these notations are more suitable for describing learning communities in media.

Figure 3.3 shows a general community model that depicts main actors of social software and dimensions of Communities of Practice (Wenger, 1998). The *mutual engagement* (ME) dimension of the CoP emphasizes collaborations for a common practice of a community though the CoP is not defined only by interactions but by the practical work they are doing together. For a learning community ME is defined not only by interactions of learners but as well during information sharing, e.g., as it happens in Wikipedia, that is devoted to the community practice — learning.

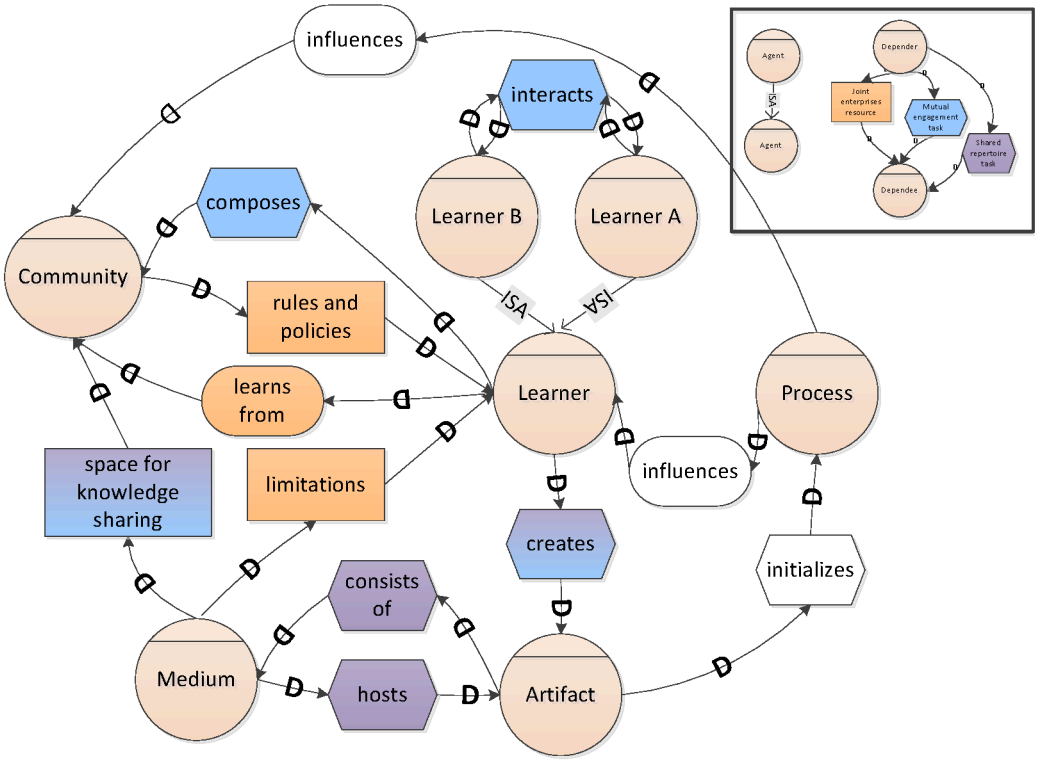


Figure 3.3: The general community model with actors of a social software (light orange).

The *Learner*, who is an agent in the data model of the social software, is pivotal for the *Community* as she *composes* communities since she *interacts* with other learners. Within the interactions the *Learner creates* the *Artifact*. For instance, forum communities can be constructed based on threads and posts which are *Artifacts* that allow forum users to *interact* with each other. The *Community* depends on the *Medium* as it provides *space for sharing knowledge*. Using the *Medium*, community members can share their practice. All these actors and dependencies conclude the ME dimension of the CoP.

The members of a CoP share practice, a *joint enterprise* dimension. It specifies being able to operate with software, having fun, wanting to share and contribute, learning, being kind and helping others, supporting and following community culture and many other facilities. In Figure 3.3 I include only some of the joint enterprise dependencies such as the *Learner* depends on the *Community* as the *Community* sets *rules and policies* and the *Learner learns from* the *Community*. Possibilities the *Learner* has are limited by the *Medium* as it provides only definite functions. For instance, in forum environments users can start new threads and post in existing threads but can not add

users into threads or start chat rooms.

The third dimension of CoP points out things a community produced or adopted during its existence — a *shared repertoire* (SR) dimension. The things are a *history of activities*, tools, concepts, stories and others. For social media, community artifacts constitute a part of SR while its medium plays a pivotal role in maintaining artifacts. Particularly, the *Medium consists of Artifacts* while the *Artifacts are hosted on* the *Medium*. The dependency between the *Medium* and the *Community* influences SR since the *Medium* provides a space where tools and concepts are exercised by the *Community*. Furthermore, without learners artifacts have not been created therefore the dependency between the *Learner* and the *Artifact* has an impact on the SR dimension.

One important fact of community existence is a knowledge flow between communities and individuals. Jäger et al. (2008) and Stahl (2006) payed special attention to the flow and Jäger et al. (2008) distinguished between processes of transcribing, addressing, and localizing that explain how the knowledge spread between community members and other communities using media. In the general community model the *Artifact initializes* the *Process* while the *Process influences* the *Learner* or the *Community*.

The presented model has a potential to express any kind of social media community considering learning theories prerequisites and media affordance. For example, Jäger et al. (2008) emphasized the transcription operation supported by a medium that lets a peer share knowledge with others. In the model the *Learner* acts according to limitations of the *Medium* and creates *Artifacts* that allow to transcribe knowledge that *influences* the *Community*. For a particular case, a learner in a collaborative workspace creates a shared space that initializes a *transcription* process that influences some peers or communities devoted to the space.

3.1.2 Specific Community Models

In this work we operate with the general model of a learning community and with specialized community models that inherit all actors and dependencies from the general model and extend the model with some additional dependencies and other items. These special models include only some of the actors and dependencies of the general model for the sake of clarity.

In the following, I describe three successful community models: *question-answer*, *innovative* and *dispute* (Petrushyna et al., 2010). These models present community patterns described with the help of actors and their dependencies. Community models can be extended by patterns that define trolling behavior of learners (Klamma et al., 2006c) or exploratory categories of learner discourse (Ferguson et al., 2013).

Question-answer community

In the *question-answer* community (Figure 3.4) I specify types of learners according to their knowledge: the *Novice* that is dependent on the *Expert*. Such a dependency is critical for the *question-answer* community. The *Expert* gains *prestige* by answering

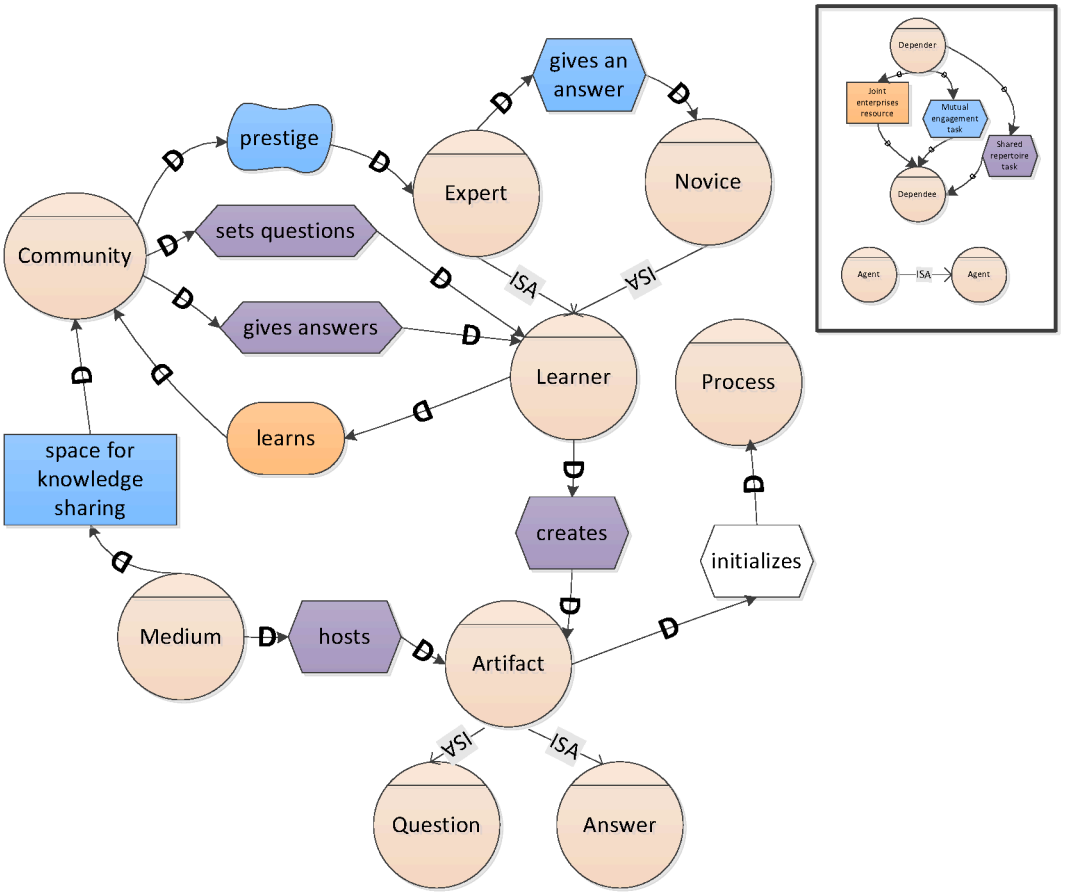


Figure 3.4: The *question-answer* community model

questions. Prestige can be one of the reasons why users help others in communities, together with global volunteering and social behavior (Fugelstad et al., 2012). Prestige is realized by gaining a popular position in a community network, by a number of useful answers or by respect of peers. One of the other characteristics of the *question-answer* community is classification of *Artifacts* into *Questions* and *Answers*. Furthermore, such a *Community* can exist because of *Learners* that *ask questions* and *give answers*.

Innovative community

An innovative community is a derivable from communities of practice (Coakes and Smith, 2007). Innovative ideas appear in such communities when innovative champions and their social capitals are community members. The *Broker* in Figure 3.5 is such a champion as she spans structural holes (Burt, 1992) and makes communication between otherwise isolated groups possible (*connect isolated groups*). She has a powerful position as she possesses information from both isolated groups. Granovetter (1973) and later Burt (2004) emphasized a critical role of brokers to enable information flow between structural holes that triggers innovative processes in communities.

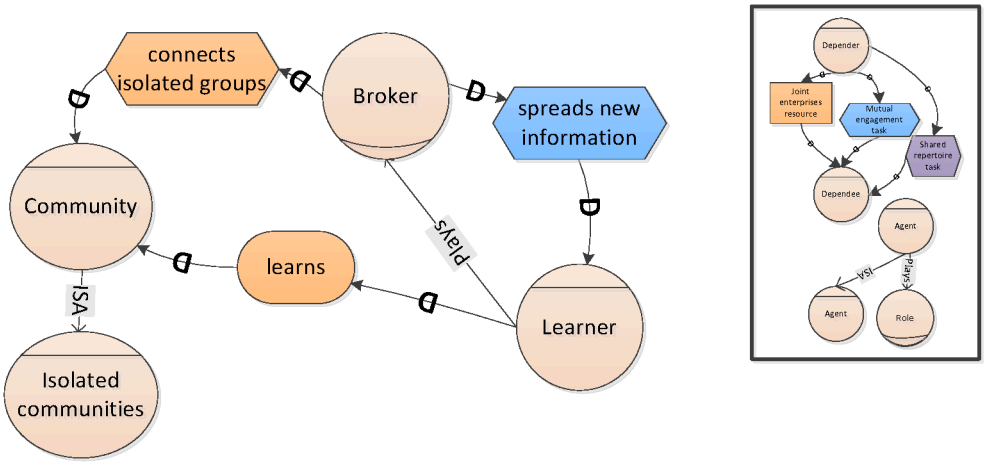


Figure 3.5: The *innovative* community

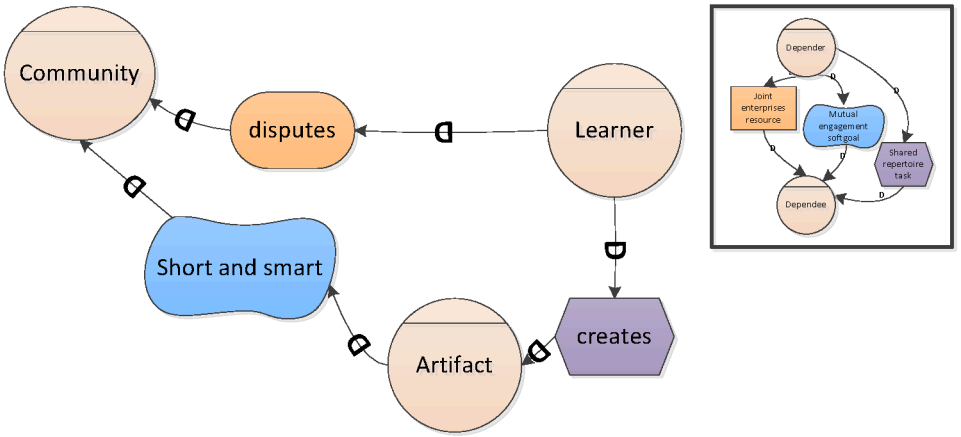


Figure 3.6: The *dispute* community

Dispute community

The *dispute* community model in Figure 3.6 refers to communities where a discussion is a way to find ground truth. The analysis of such communities (Wagner et al., 2012) showed that *Artifacts* should be *short and smart* to engage others to interact. Another work (Ferguson et al., 2013) specified different categories that can be used for refining the model and specifying a dispute character. The *Community* depends on the *Learner* as she initiates and maintains *disputes*.

The presented models can serve as starting points for community stakeholders. They can decide to consider their community with the general model from Figure 3.3 or choose one of the presented special cases of the model. When a model has been chosen or after the model is created based on monitoring and analysis of community data as described in next sections community stakeholders can be either satisfied with their community model or think about community refinement.

3.2 Refinement

A refinement phase is relevant for those communities that would like to improve their states. Community stakeholders can choose ways to change their communities so that they reach a state desired for the stakeholders, for example if they want that their community becomes an innovative community they need to engage a broker with her social capital to the community. Learners, researchers, developers, managers and other stakeholders can change learning communities in social media. For example, if learners cannot solve problems because of asynchronous collaborations, community managers have to recognize such a need and add a chat room to community media. In other cases a set of *artifacts* provided by a community is not enough for learner needs. Therefore, recommender systems (Manouselis et al., 2011) can help to find relevant *artifacts* that refine community models. Furthermore, a perfect group of learners (Alfonseca et al., 2006; Liu et al., 2009; Zakrzewska, 2010; Kyprianidou et al., 2012) can increase success of a community or indicating communities suitable for learners (Klamma, 2013) can increase efficiency of learning process.

Nevertheless, existing solutions are not generic and results of model refinements are estimated only after changes in community environment are applied. Alternatively, one uses agent-based models of learning communities that make it possible to simulate their evolution (Zhang and Tanniru, 2005; Li et al., 2008). Agent-based modeling (ABM) systems usually consist of agents, their environment, agent knowledge, and experience (Menges et al., 2008). The agents should be autonomous, situated, proactive, and social. It means the agents are fully responsible for collaborations in social systems due to their strategies and payoffs. For example, in forums users are collaborating due to threads and their messages. Agents who represent the users decide about their activities according to their payoff and strategies that can be influenced by the global rules such as preferential attachment or reciprocity strategies. Usage of global rules of social behavior approximates simulations of societies to their real states (Wunder et al., 2013). For example, basing on *preferential attachment* (Barabási and Albert, 1999) actors in societies get more connections with other actors if they have already a high number of connections. This fact is known as the *rich gets richer* or the Matthew principle. Alternatively, actors may form connections with those whom they already knew. Such a strategy is called *reciprocity* (Albert et al., 1999).

Since we use the *i** approach for community modeling, community models include autonomous agents with their goals and roles. *i** models have already been used in

simulation environments (Roesli et al., 2009). But *i** models do not emphasize directly strategies and payoffs of agents though these can be inferred according to user roles or given from outside. To create a multi-agent system of a community we formalize it as follows.

Definition 1 (A multi-agent social media network) *Formally, the entire system can be described as a tuple* $Soc = (A, Act)$ *where*

- $A = \{A_1, ...A_n\}$ *is a set of agents, where* $|A| = n \in \mathbb{N}$
- Act *is a set of possible (predefined) actions that are performed by agents* A *under the influence of a set of strategies* $S = \{S_1, ...S_l\}$*. Strategies define probability distributions of actions,*
- *agents in* A *have attributes* $X = \{X_1, ...X_m\}$ *and the attributes are assigned by a* ν *function:* $A \xrightarrow{\nu} \mathbb{R}^m$, $|X| = m \in \mathbb{N}$,
- *agents can have* $Roles$ *that are dependent on their attributes and assigned by a* η *function:* $\mathbb{R}^m \xrightarrow{\eta} Roles$*, where* $Roles = \{Role_1, ...Role_v\}$,
- *social relations* $R(t) \subseteq A \times A \times \mathbb{R}^+$ *appear as results of acts* Act *and can weaken or disappear after some time.* t *is a time point,*
- *a set of artifacts* $Artifacts_t = \{Artifact_1, ...Artifact_{k_t}\} \subseteq C$, $|A^t| = k_t \in \mathbb{N}$, *is created by agents* A *with the help of actions* Act *in a time point* t*.*
- *agents* A *can belong to communities of agents* C *that are defined by the function* $\theta(t)$*.* $A \xrightarrow{\theta(t)} C^t$*, where* $C^t = \left\{C_1^t, ...C_{k_t}^t\right\} \subseteq C$, $|C^t| = k_t \in \mathbb{N}$.

3.3 Monitoring Social Media Learning Communities

Monitoring tools collect and store data from social media communities. In the following, I first describe social media that I am investigating and related work on these media. After that I introduce the data model — the Mediabase model — that is the source of hierarchy dimensions used for the Mediabase Cube where the data is stored to perform efficient queries. In the end of the section I describe the design of watchers that collect the data.

3.3.1 Data

In this work we concentrate on diverse social media: forums, collaborative workspace and Wikipedia. First one includes discussions that are popular in other social media. The collaborative workspace eTwinning is similar to other collaborative spaces where users interact in spaces to enhance their working skills and knowledge. Wikipedia is one of the Wikimedia resources that allows to create and revise different digital artifacts such as articles.

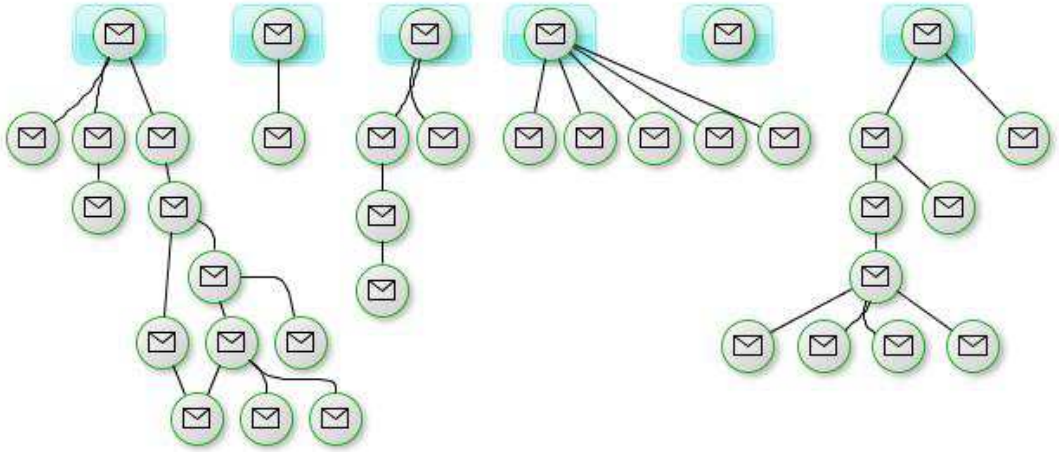


Figure 3.7: Possible structures of threads in forums

3.3.1.1 Forums

Forum users start threads, post questions and answer them. Forum posts are available on forum Web pages publicly or after authorization. A structure of a forum resembles a set of trees (Figure 3.7). A person starts a thread, a sequence of e-mails with the same topic. She sends a message, later gets a reply from one or several users that can be replied as well.

Forums are extendable since their interface may be changed and new topic areas are easily supported by extending forums with subforums. Some forums analyze users activities and based on these grant roles and access permissions. For example, in stackoverflow[1] a user can rate answers of others only after she posts some answers or questions.

A forum allows users to perform self-regulated activities (Zimmerman, 1990). These are asking questions (*self-starting*, *self-instruction*), answering questions (*self-efficacy*, *self-improvement*), reading discussions (*self-motivation*, *self-instruction*), and sharing relevant information (*self-efficacy*).

A forum allows collaborative processes through user interactions that trigger transfer of knowledge of a community to knowledge of individuals or vice versa. Data provided by interactions can be used for ranking users and constructing inner communities of the forum.

Studies of forums in the learning context showed benefits for learners in academic performance (Davies and Graff, 2005; Morris et al., 2005) and benefits additionally to formal learning (Deslauriers et al., 2011; Palmer et al., 2008). For example, Carceller et al. (2013) found that learners who participated more actively in a forum were more

[1]The developer forum stackoverflow `http://stackoverflow.com`, Last access 10.04.2014

successful with academic achievements. Recent forum studies by Johnson et al. (2012) emphasized the need of understanding student interactions as they can affect student knowledge evaluation and student motivation and help design learning environments efficiently (Anderson et al., 2014).

Forum communities investigated in Technology Enhanced Learning are in most cases not considered as Communities of Practice (CoP) (Wenger, 1998) and therefore analyses of the communities do not combine every CoP dimension: analysis of interactions, discourse and goals or intents. Moreover, in most cases modeling of forums is limited to mathematical or statistical models but it misses to emphasize the cooperation between learning theories and results of forum analysis.

3.3.1.2 Collaborative Space eTwinning

The collaborative teacher network eTwinning[2] is an initiative by the European SchoolNet[3] that provides a platform for collaborative projects of European schools, formal or informal professional development and social networking. Professional development of teachers is defined as a number of activities that enhance personal skills, experience and knowledge (OEC, 2009) while informal professional development considers participation in projects and networking.

Traditional forms of training seem less efficient for many teachers than networking with other teachers (US Department of Education, 1999). Teachers' cooperations improve educational processes and outcomes (OEC, 2009), where teachers share knowledge with each other and develop new knowledge jointly (Sloep and Berlanga, 2011). Networks such as Tapped-in[4], Teachernet[5], and eTwinning aim to support professional development of teachers.

Teachers can be viewed as self-regulated learners in the eTwinning where they plan their learning on their own (*self-instruction*, *self-starting*). In projects they interact with other teachers and thus initiate collaborative processes (*self-motivation*), exchange their knowledge with others, exercise communication skills (*self-efficacy*) and refine them (*self-improvement*).

Breuer et al. (2009) investigated collaborative projects of eTwinning and performed social network analysis on graphs organized by teachers cooperations. The authors provided eTwinners with information about their activities and activities of their peers and found that teachers require more evident measures and visualizations. Vuorikari and Scimeca (2013) detected that less than 20% of teachers stay active over 6 years in eTwinning while more than one third of eTwinners use social network tools but are not

[2]eTwinning European teacher network`http://www.etwinning.net/en/pub/index.htm`, Last access on 13.08.2014

[3]The European SchoolNet `http://www.eun.org/`, Last access on 13.08.2014

[4]Tapped In was the online workplace of an international community of education professionals till March, 2013`http://www.tappedin.org/`, Last access on 13.08.2014

[5]UK teacher network that is closed since middle of 2011 but its content is possible to view on the National Archives `http://webarchive.nationalarchives.gov.uk/*/http://www.teachernet.gov.uk/`

participants of any project. Therefore, teachers require support tools that engage them into participating further in projects and keep their communities alive.

3.3.1.3 Wikipedia

Rafaeli et al. (2009) defined Wikipedia users as a knowledge building community and Wikipedia as a place for *articulating individual knowledge*. The online encyclopedia Wikipedia[6] is administrated and maintained by just a few workers and millions of volunteers. Wikipedia is the largest example in the Internet of using crowd-sourcing for creating a base of knowledge. Contributors of Wikipedia can create and revise articles on different topics and maintain their own Wikipedia pages. Each article can include a discussion, where users talk about changes of the article. Proposed revisions are reviewed by Wikipedia administrators.

The Wikipedia is a social medium for self-regulated learning (Zimmerman, 1990) that allows users to read articles devoted to learning topics, revise articles (*self-efficacy, self-improvement* in case of revising own texts) and participate in discussions (*self-efficacy, self-improvement, self-instruction*). Wikipedia provides a playground for collaborative processes that are initiated by Wikipedia contributors through revising the same articles or discussing revisions. According to activities one can detect communities of Wikipedia contributors that are exercising sharing of a ground truth and therefore keep learning during the whole time of their practice.

Research on Wikipedia as a learning environment includes works where students contributed to Wikipedia instances by creating and revising articles devoted to a learning topic (Chao and Lo, 2011; Kessler and Bikowski, 2010; Kimmerle et al., 2009). At the same time Wikipedia provides a great base for cross-cultural analysis since it consists of two hundred eighty eight instances maintained in different languages. Investigation of different cultures is required to design appropriate learning environments for a particular culture (Nemoto and Gloor, 2011). For example, Hara et al. (2010) found that greater respect of hierarchical structure in society and preferences of working collectively are prevailing in eastern countries while authors from Wikipedia of western countries disagree more often.

Modeling Wikipedia communities, that emerge in different Wikipedia instances, can indicate variety in preferences of learners, their activities and types of their communities depending on their culture. I choose 13 Wikipedia instances to conduct analyses of their communities and users and detect patterns of communities that are influenced by cultures of community members.

3.3.2 Mediabase Model

The Mediabase was firstly presented in (Spaniol and Klamma, 2004) and later described and used in further works (Klamma et al., 2005, 2006b,c, 2007; Klamma and

[6]Wikipedia `http://www.wikipedia.org/`, Last access 31.07.2014

Petrushyna, 2008; Petrushyna and Klamma, 2008; Klamma and Petrushyna, 2010; Krenge et al., 2011; Song et al., 2011; Derntl and Klamma, 2012; Hannemann and Klamma, 2013).

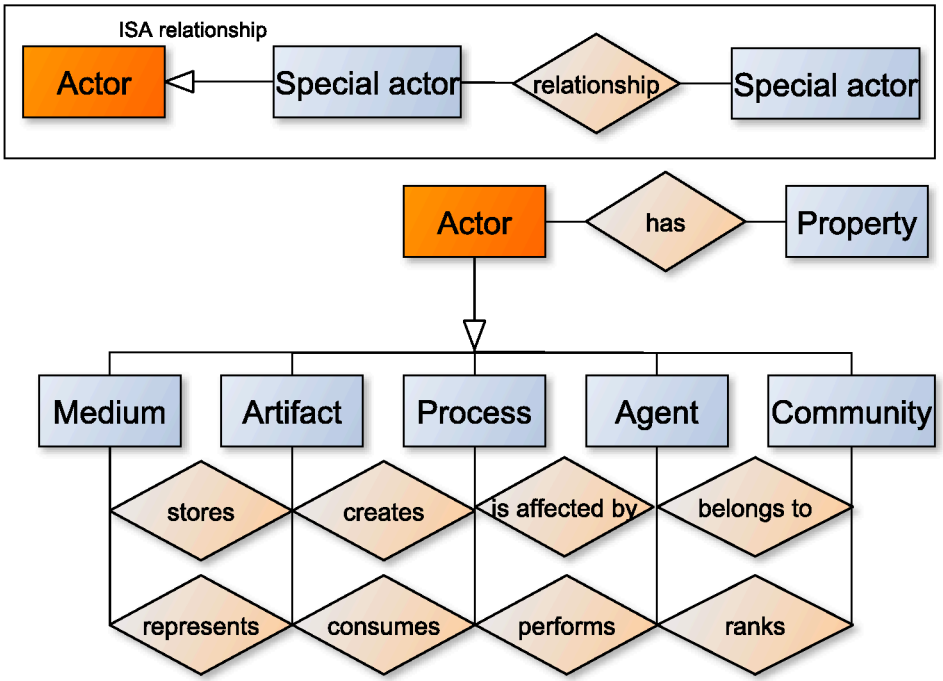


Figure 3.8: The Mediabase model (Klamma, 2010)

Klamma (2010) introduced the Mediabase model as the Actor Network Theory (ANT) Model of social software. The ANT gives us an opportunity to explain complex systems with human and non-human actors that transfer knowledge from one to the other actor (Latour, 1999). The Mediabase model includes the following actors: *Medium*, *Artifact*, *Process*, *Agent*, and *Community* (Figure 3.8). In any social software there is a *Medium* that creates an *Artifact*. If a forum is a *Medium* then a post is an *Artifact*. A *Process* creates or consumes an *Artifact*. An *Agent* initiates the *Process* or is influenced by it. For example, a forum user answers a question and initiates a *transcription* process of knowledge sharing while other users that read the answer are interpreting the information from the post according to their knowledge that initiates another process of *localization* (Jäger et al., 2008).

An *Agent* is a part of a *Community*. All forum users are members of a forum and therefore members of a forum community. A *Community ranks* an *Agent* according to initiated processes and social positions in the *Community*. For example, learners differ based on their collaborative activities (Lipponen et al., 2003).

3.3.3 Multidimensional Data Model

Traditional databases store data operated mainly by one application. *Data warehouses* (DW) provide a central point where the data is stored and refreshed. Jarke et al. (1999) defined DW as ”the right information in the right place at the right time with the right cost in order to support the right decisions”.

A warehouse collects data using wrappers that load data from data sources, such as social media. The operational data store (ODS) includes the data after transformations, done by wrappers. The ODS is used only to store current data. The global data warehouse collects all data gathered by wrappers from the beginning of the DW existence. Data marts, databases with parts of data from the DW, are created to solve a particular task and to provide data to a particular application.

DW data are used for Online Analytic Processing (OLAP) but it should be represented in a multidimensional data model (Chaudhuri et al., 2001). The model includes numeric measures that are used for OLAP. One or several measures constitute a *fact*. Descriptive properties of facts are dimensions that consist of hierarchies that allow to get detailed or aggregated information. The natural representation of the facts described by the dimensions is a multi-dimensional data cube.

DW multi-dimensionality provides easy access to different kind of data without a big effort spent on data selection and joins. Since social software data operates with a numerous set of fields and it is efficient to be stored in DW (Newman and Girvan, 2004; Tantipathananandh et al., 2007; Blondel et al., 2008; Aynaud and Guillaume, 2010; Chakraborty et al., 2013).

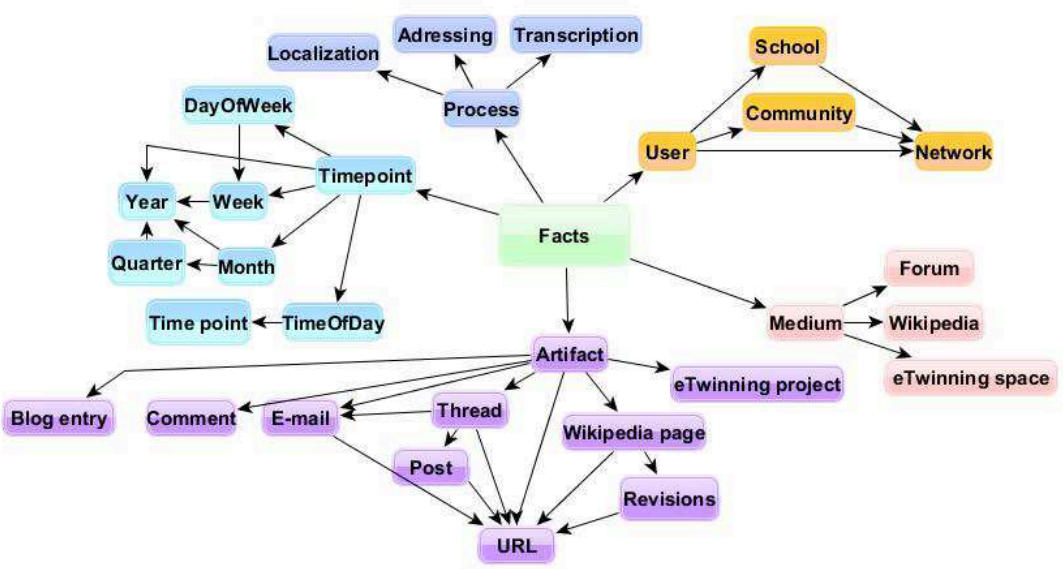


Figure 3.9: Dimension hierarchies of the Mediabase cube

3.3.3.1 Dimensions

The Mediabase cube dimensions are inspired by the described Mediabase model (check Figure 3.8). I adopt special actors of the model for a Mediabase Cube hierarchy in Figure 3.9.

Agent An *agent* dimension has 2 branches: user-community-network and user-school-network. Users are parts of some communities, networks and schools. The schools and communities are parts of networks. Therefore, the hierarchy is a directed acyclic graph.

Medium A *medium* dimension has a flat hierarchy and includes such elements like forums, Wikipedia, and eTwinnning. The dimension serves for selecting facts of the Mediabase cube for a particular medium.

Process A *process* dimension is triggered by activities such as posting, commenting, revising (a Wikipedia article), creating thread, creating a project, and others. These activities are dependent on affordance of media.

Time A *time* dimension allows to query for the cube facts in a given time interval. It consists of a time of a day, a day, a day of a week, a month, a quarter, and a year.

Artifact An *artifact* is a product of a medium. In forums one can create an artifact by starting a *thread* and by posting a *message*; in Wikipedia artifacts are *articles* and their *revisions*. In eTwinning users can send *e-mails*, create *projects*, *blogs* and *comment* on blogs, projects, prizes and other events.

The hierarchy of the artifacts is not a tree at all. A *URL* artifact has more than one parent as it may appear in posts, e-mails, comments, and other artifacts. *Posts* in forums or *e-mails* in eTwinning are collected in *threads* that are artifacts as well.

3.3.3.2 The Cube Model

The Mediabase cube is based on a multidimensional data model adopted from (Lin et al., 2008) with dimensions depicted in Figure 3.9.

Definition 2 *The collection of $Facts$ is stored in a 5-dimensional cube*

$$MBCube = (D_1, D_2, D_3, D_4, D_5, Facts).$$

- *Each row of the form*

  $$(d_1, d_2, d_3, d_4, d_5, f),$$

  where $d_i \in D_i$ is a dimension value for D_i and $f \in Facts$. $S(F)$ is a set of structural measures, $Se(F)$ is a set of semantic measures and $O(F)$ is a set of other facts. $\{S(F), Se(F), O(F)\} \subseteq Facts$ is a set of facts that match $d_1, d_2, ..., d_5$.
  $f = \{v_1, v_2, v_3, ...v_k\}$ is a set of fact values and $Facts$ is the multiset of facts.

- *A cell of the cube is of the form*

$$(d_1, d_2, ..., d_5 : S(F), Se(F), O(F)).\ (d_i \in D_i \bigcup \{*\}).$$

$d_i = *$ means that data is aggregated according to the dimension D_i. For example, one can query for all artifacts that were created in all media in a time period by the user A with no specific process.

- *A 1-D(imensional) cuboid is denoted as*

$$(d_1, d_2, ..., d_5 : \{S(F), Se(F), O(F)\})(d_i \in \{?, *\})$$

where ? means that D_i is the inquired dimension and, for example, the cuboid consists of the set of cells of all items devoted to the user A.

- *The subset of cells from a cuboid:*

$$(d_1, d_2, ..., d_5 : \{S(F), Se(F), O(F)\})(d_i \in D_i \bigcup \{*\} \bigcup \{?\})$$

For example, one can query for all posts (the artifact is the requested dimension) that appear in a particular time point in all media caused by the user A.

- *A cube without any specification of dimensions is*

$$(d_1, d_2, ..., d_5 : S(F), Se(F), O(F)).$$

3.3.4 Collecting Data

Since social media have different formats we need to specify wrappers that collect data into the data warehouse. Furthermore, the wrappers have to transform data according to the data model.

3.3.4.1 Forum Watcher

The *Forum Watcher* (FW) simulates a browser that is reading a forum and uses templates of forum websites created according to content models of forum pages. These templates define positions of forum sublinks, threads, forum user names, post contents and dates of posts. The *FW* scans forum information as it is defined in the templates and cleans extracted data.

The FW is a Mediabase crawler and thus all its entities are instances of the Mediabase actors. I show the entity relationship diagram of the FW in Figure 3.10. The *Forum* inherits from *Medium*. A *Forum* stores *Artifacts* such as *Posts* and *Threads*. A *Thread* contains *Posts*. An *Agent* of a *Forum* is a *Contributor* who performs a *Process* or is affected by a *Process*. The *Contributor* creates the *Thread* and/or writes the *Post*. We can extend the list with other contributor activities such as *replying*, *answering*, *questioning*, *arguing*, *discussing* and classify the activities according to transcriptivity

theory (Jäger et al., 2008). The *Contributor* belongs to a *Forum Community* that is an instance of the *Community*. For example, a forum can include many communities that are tightly connected groups and forum members can belong to several of them, i.e., communities overlap.

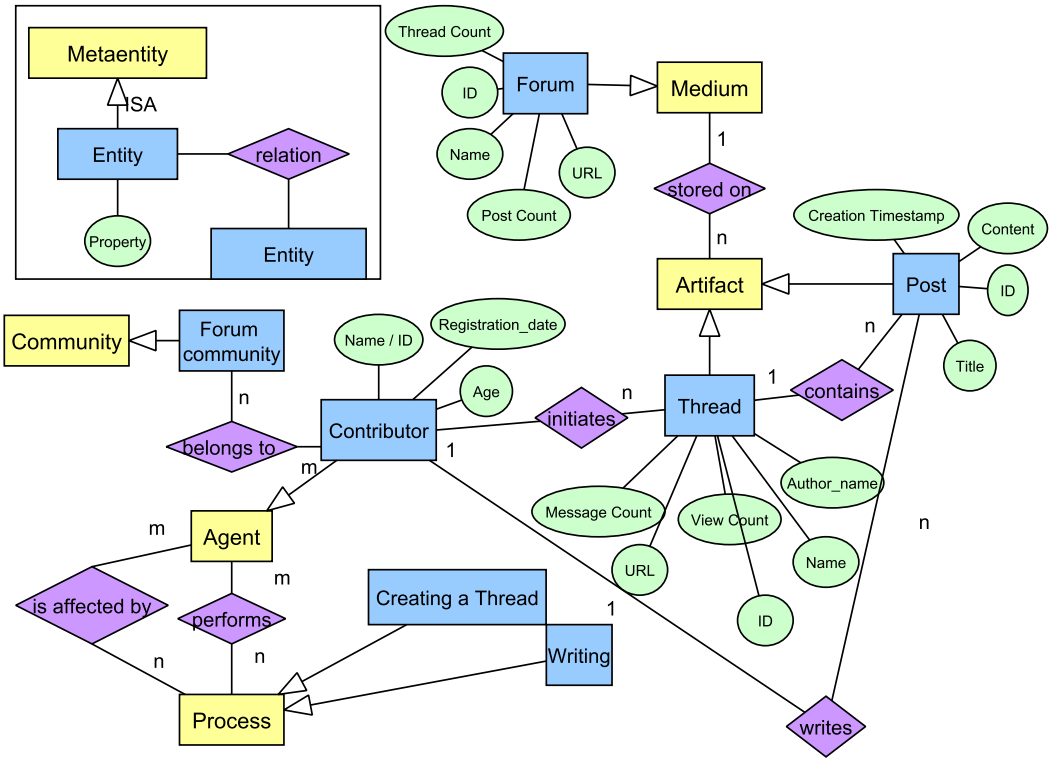


Figure 3.10: The Entity Relationship Diagram of the Forum Watcher

3.3.4.2 eTwinning Watcher

Within the TeLLNet project[7] we accessed several eTwinning data dumps. A data dump includes all teachers registered in eTwinning, their schools or institutions, their regions and countries with a given time dimension. Collaborative projects of teachers are characterized by the number of pupils, the subjects, the languages used and many other properties. From a data dump we can retrieve data about teacher activities such as participation in projects, posting messages on profiles of others, including other teachers in contact lists and writing and commenting on blogs and teacher or project profiles. Some projects and teachers are recognized for extraordinary quality by prizes and quality labels that we use later in our investigations.

[7]Teachers' Lifelong Learning Network project `http://www.tellnet.eun.org/web/tellnet`, Last access 27.02.2015

eTwinning is a *Medium* that lets teachers create collaborative *projects* that are one type of *artifacts* of eTwinning. *Teachers* are the *Agents* of an *eTwinning community*. They can initiate *Processes* by *creating a project, an e-mail*, or *a contact list*. The *eTwinning community* is a *Community* that can be defined based on collaborative processes between teachers.

3.3.4.3 Wikipedia Watcher

Klamma and Haasler (2008a) created the Wikiwatcher tool that can be used for retrieving Wikipedia data, visualizing their networks and performing simple social network analysis. The Wikiwatcher (WW) is a two-stage system as depicted in Figure 3.11: in the first stage the WW exports freely available XML dumps from Wikipedia and parses them. In the second stage the WW investigates Wikipedia.

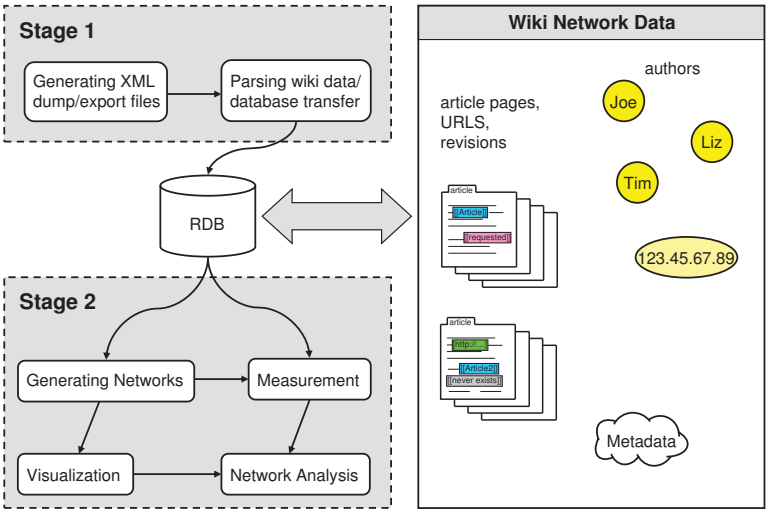


Figure 3.11: The 2-stage process of extracting and analyzing/visualizing data from Wikipedia (Klamma and Haasler, 2008a)

The WW considers *Wikipedia articles* as *Artifacts* of *Wikipedia* that is a *Medium*. Other *Artifacts* are *article revisions*, and *URL*s. Each article can have multiple revisions and refer to multiple URLs. An *Agent* in Wikipedia is a *Wikipedia contributor* that initiates the *Process* caused by *creating* or *revising articles* activities. Activities of different Wikipedians on the same articles help define *Wikipedia communities*.

3.4 Analyzing Communities

After retrieving the data about communities using the *monitoring tools* we have been analyzing the data. Results of the analysis reveal community states in a given time period.

3.4.1 Structural Measures

Forums, Wikipedia, and eTwinning can be considered as a graph G where agents are nodes N, e.g., forum users, Wikipedia contributors or teachers. The nodes are connected if they represent agents that collaborate. In case of a forum graph G its nodes N are connected through edges E if forum users represented by N participate in the same forum threads. In the Wikipedia case, nodes N of G, a Wikipedia graph, are connected through edges E if Wikipedia contributors represented by N revised the same articles. And finally, for eTwinning, nodes N are connected through edges E if teachers represented by N participate in the same projects. Alternatively, if one teacher T_1 wrote an e-mail to another teacher T_2 or commented blogs or profile pages of the teacher, then an edge $e_1 = \{n_1, n_2\}$ exists between two teachers' nodes. Following these mapping between nodes representing users and edges representing interactions we create graphs such as $G = (N, E)$.

Using Social Network Analysis (SNA) (Wasserman and Faust, 1994), we calculate different measures in all of the graphs. These measures can be used to estimate mutual engagement dimension (Wenger, 1998). In the following, I introduce a number of SNA measures relevant for this work. After that I mention some works in TEL where SNA is involved.

Definition 3 (Degree centrality) *For an undirected graph, the degree centrality of a node defines the number of edges it has. For a directed graph, we distinguish between in-degree and out-degree centrality. The in-degree centrality of v is the number of edges directed to a node v. The out-degree centrality of v is the number of edges that direct from the node v to other nodes.*

A node with a high degree centrality is a connector or a hub in a network. The user who is represented by the node interacts with others more actively.

Definition 4 (Closeness centrality) *is a measure that calculates how close a node is to other nodes.*

We use the inverse closeness defined by Sabidussi (1966). Freeman (1978/79) found that the inverse closeness estimates the node closeness more accurately but it depends on graph size.

$$C_c(u) = \frac{\sum_{v \in V} d(u, v)}{n - 1} \tag{3.1}$$

It calculates the number of paths from a node u to other nodes, i.e., $\sum_{v \in V} d(u, v)$ divided by the number of all paths in a complete graph of a network.

Therefore, nodes with the highest inverse closeness centrality occupy central positions in the network. Such nodes have a best view on the network and possess excellent positions to monitor information in the network. Nodes with closeness centrality close to 1 are far away from the network center. Newman (2004) found that authors with low values of closeness in a co-authorship network would be the first who received

new information. Moreover, they are better news reporters as they reach much more authors in a short time.

Definition 5 (Betweenness centrality) *Betweenness centrality of a node v indicates the influence of v on connections existing between other nodes. The betweenness centrality of v is defined as:*

$$C_b(v) = \sum_{s,t \neq v} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3.2}$$

where σ_{st} denotes the number of the shortest paths between any node s and t, and $\sigma_{st}(v)$ stands for the number of the shortest paths between any node s and t containing v.

Nodes with high betweenness centrality are called *brokers* (Freeman, 1977) that connect isolated groups of nodes, so-called *structural holes* (Burt, 1992). Newman (2004) showed in his own co-authorship network that most connections to other people go through several collaborators that possessed high betweenness scores. They control information flows in social networks as they tolerate the knowledge to flow between different network groups connected to each other due to brokers only. Therefore, brokers are expected to posses information from all their connections while other users with smaller betweenness operate with less amount of information.

Definition 6 (Local clustering coefficient) *The local clustering coefficient C of a node v in a graph network $G = (V, E)$ is defined by:*

$$C(v) = \frac{2|\{(i,j) \in E : i, j \in N_v\}|}{D(v) * (D(v) - 1)} \tag{3.3}$$

The neighborhood $N_v = \{w : (v, w) \in E\}$ contains adjacent nodes of v and $\frac{D(v)*(D(v)-1)}{2}$ defines the number of edges in a complete graph of the node v with all its neighbors.

The clustering coefficient is a measure indicating how likely it is for nodes to organize a cluster together. Usually real-world networks have a higher clustering coefficient due to sociability of humans and their inclination towards forming groups (Watts and Strogatz, 1998).

SNA measures were used in many works to investigate communities and learning communities in particularly. For example, Fisher et al. (2006) analyzed newsgroups and Klamma et al. (2006c) mailing lists. They built social networks and calculated user and network measures. Based on these, they defined special roles of users such as answering or questioner persons. An answering person usually responds to peers with lower in- and out-degree centrality (Anderson et al., 2014) that post and get replies to posts much less than other peers. Lipponen et al. (2003) differentiated in a student network between 2 clusters: one with high in- and out-degree and another with low

in- and out-degree centralities. Laat et al. (2007) defined roles of students using SNA measures: some of them provided new ideas, others organized the group activity. Another role called *troll* (Klamma et al., 2006c) described a behavior of a user that posts only in threads that she had started herself.

3.4.2 Semantic Measures

All media investigated in this work include language items such as texts that are results of mental processes (Ferguson et al., 2013). Investigation of users' texts reveals user knowledge, goals, and emotional attitude towards a discussed topic. Such an analysis can respond to the requirements from communities of practice in considering *joint enterprises* and *shared repertoire* dimensions while observing communities.

3.4.2.1 Emotional Analysis

Calvo and D'Mello (2010) reviewed numerous works that investigated affects or emotions of humans. Among a variety of approaches for detecting emotions in texts we choose a text-based approach. We utilize the Linguistic Inquiry and Word Count (LIWC) vocabulary (Pennebaker et al., 2007) and choose 9 categories of words from the vocabulary. The first six categories in Table 3.1 serve for recognizing *emotions* in texts. We use *posemo* and *negemo* categories for recognizing positive or negative emotions in texts while *anger*, *anxiety*, *swear* and *sadness* for defining negative attitude. The three last categories serve for discovering words that indicate cognitive work; *cogmech* includes words that denote cognitive process and *insight* and *achiev* include words related to hindsight. We differentiate between two possibilities: either texts express some cognitive work indicating relations to learning or texts include words characterized as flame or chatting.

Category	**Examples**
posemo	awesome, inspir*, super
negemo	depress*, impolite*, scary
anger	agress*, stupid*
anx	afraid, nervous*, shy*
sad	alone, fail*, miss
swear	ass, hell, sucks
cogmech	analy*, infer*, problem*
insight	explain, induc*, reason*
achiev	create*, excel*, skill

Table 3.1: Categories of LIWC with examples, * denotes the end of a word stem.

We create a language model classifier based on dynamic language models for each category based on training sets that include sentences from media, artifacts of which

need to be classified. The sets include sentences that include words from the training categories. The classifier applies the joint logarithmic probability that calculates a probability of a character sequence cs belonging to a category cat:

$$log_2 P(cs, cat) = log_2 P(cs|cat) + log_2 P(cat) \tag{3.4}$$

$P(cs|cat)$ is the probability of cs being in a language model for a category cat. $P(cat)$ is the probability of the category cat. Since the character sequence cs can belong to different categories we calculate probabilities for all categories. The highest score defines the category of the cs. The score is calculated as a division between the joint logarithmic probability of the cs from Equation 3.4 and the length of the cs.

$$score(cs, cat) = \frac{log_2 P(cs|cat) + log_2 P(cat)}{cs.length()} \tag{3.5}$$

This naive approach is sufficient for our case since we operate only with two categories in sentiment analysis (emotional or neutral) and with two categories for recognizing cognition in texts (cognitive or neutral). We denote a post as emotional if 50% of sentences include a sentiment, otherwise it is neutral. A post with words denoting cognitive work is classified in the same way.

3.4.2.2 Learning Concepts and Topics

Named entities (NE) are arguments of information units, sentences, such as companies, locations, products, or persons (Grishman and Sundheim, 1996). NE are recognized by, first of all, separating whitespace characters from non-whitespace characters to define meaningful units. This process is called *tokenization*. The input and output of the *tokenization* process may look like this:

Input: Firstly, I would like to introduce Paul Brown to all of you.

Output: Firstly TB , TB I TB would TB like TB to TB introduce TB Paul Brown TB to TB all TB of TB you. TB

In the next stage Part-of-Speech (POS) tagging attaches appropriate POS tags to each word. The conventional POS tags are *Noun*, *Verb*, *Adjective*, *Preposition*, *Adverb* and so on. Unfortunately some words are ambiguous, thus the POS tagging should consider the whole sentence as well. For example,

1. Plants/N need/V light/N and/P water/N.

2. Children/N plant/V trees/N in/P the garden/N.

POS tags are refined based on both word definitions and word context. Later, texts are divided into chunks, like nouns, verbs and preposition phrases. For instance,

[NP He] [VP said] [NP the project] [VP will be finished] [PP in] [NP the next week].

NP tags a noun phrase, VP — a verb phrase, and PP — a proposition phrase.

In the last stage, *domain analysis*, co-references are resolved or partial results are merged. This stage can be done by applying *supervised*, *semi-supervised* and *unsupervised* learning. Although supervised approaches , like Hidden Markov Models (Elliot et al., 1995), achieve excellent results they need a large annotated corpus while semi-supervised or unsupervised techniques use just a small set of labeled data or clustering for identifying patterns in texts and map them with known patterns.

Most recent and relevant services recognizing NE are evaluated by Rizzo and Troncy (2012). Some of them utilized resources from the Linked Open Data (LOD) Cloud[8]. LOD makes it possible to enrich data with other metadata stored in LOD datasets.

LOD facilitates connection of educational sources with each other and allows to enrich one educational item with the another one easily (d'Aquin, 2012). Lecturers of tens of universities in UK use the Talis Aspire[9] services to create and manage courses or lists of resources devoted to modules. Students can access these resources belonging to their and other universities if the resources are related. Another usage of Linked Data is the detection of missing relevant references in learning courses of Khan Academy [10] (Siehndel et al., 2013). d'Aquin (2012) used LOD for interpretation of results of sequence pattern mining of learning paths. The applications of LOD is only starting to emerge and to the best of our knowledge it has not been used to enrich a conceptual model of a learning community as well as it has not been considered as a technology that helps to investigate the *shared repertoire* of a community.

3.4.2.3 Intent Analysis

Strohmaier et al. (January 13) and Tatu (2008) showed that analysis of texts can reveal *intents*. Intents are parts of goals that impact learning communities, particularly their *joint enterprises* (Wenger, 1998). Furthermore, awareness of goals makes it easy to support learners in their learning processes (Ley et al., 2010) since the way how people acquire knowledge and perform tasks depends on their goals (Zukier, 1986).

Learning goals were recently investigated in Okoye et al. (2013) with machine learning and natural language processing approaches that they used to create trajectories that learners need to succeed to achieve goals. In the experiment learning goals were defined as sentences from books and papers that conclude the minimum knowledge of students and thus advice what they have to learn to get expertise on a learned topic. The proposed methodology answered what to do to learn a subject but because we operate with community activity data we are interested to know what are current learning goals.

Performing tokenization, POS tagging, and syntactic language patterns detection of texts we can find patterns such as $VB_1\_to\_VB_2$, such as *learn to calculate*, and the

[8]Linked Open Data cloud `http://lod-cloud.net`, Last access on 07.05.2014

[9]The Talis Aspire services `http://www.talis.com/`, Last access on 04.04.2015

[10]Khan Academy `https://www.khanacademy.org/about`, Last access on 08.05.2014

WRB_to_VB, e.g., *how to calculate*, that can indicate goals (Tatu, 2008). Here VB is a verb and WRB refers to a Wh-adverb (how, when, where, why). A combination of both patterns $WRB\_do\_I\_VB$, e.g., *how do I learn to write an essay*, is another possible expression of a goal.

3.4.3 Community Detection and Evolution

One of the pivotal steps for modeling learning communities is community detection and evolution. In a first set of experiments we operate with the Louvain algorithm (Blondel et al., 2008) for discovering communities of users and then map communities in different snapshots using Jaccard index. The shortcoming of such a solution is a high computation time when we are working with numerous large communities. Therefore, we refine our solution by adopting more efficient algorithms for detecting communities and defining events of community evolution where the algorithms are implemented in a distributed environment. In the following, we describe both approaches since both have been used later.

3.4.3.1 Time Intervals

Community detection algorithms require time intervals to define communities. Most works performing community detection do not give a rationale for the choice of a particular time interval, though the choice influences the community detection output (Morrison et al., 2012).

Time intervals can depend on events happening in a community. For learning communities exams or tests serve as events that initiate learners' activities. Intervals with fewer events require fewer passes of an algorithm and thus calculating intervals regarding events makes computations tractable:

A time interval $interval_j$ is calculated using the starting point t_j of an event e_j. I define a time point before the event e_j and a time point after it $before_j$ and $after_j$ correspondingly.

$$interval_j = (before_j, after_j)\ , after_j > before_j$$

Coefficients b and a are used for calculating $before_j$ and $after_j$.

$$before_j = t_j - l \times b, \text{ where } l \text{ is the length of the } interval_j$$
$$after_j = t_j - l \times a$$
$$a + b = 1$$

After intervals are defined I can depict time sliding windows in Figure 3.12. The windows have different length depending on a and b coefficients and the appearance of events. For each time window a corresponding event appears in the middle of the window.

Each interval is used to take snapshots of a network that include nodes and edges for the time interval. Communities are detected in these snapshots.

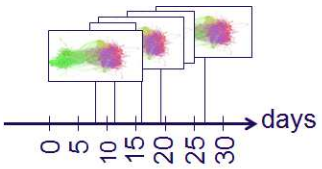


Figure 3.12: Time sliding windows that define snapshots for detecting communities

3.4.3.2 Naive Approach for Community Detection and Evolution

The Louvain algorithm has 2 phases: first, each node belongs to its own community; then, random nodes are merged into the same community and the community with the highest modularity (Newman and Girvan, 2004) value sustains. The communities obtained in the second phase are included in a new graph. After that, the algorithm repeats the phases with the exception that the communities found are stable and other nodes or communities can be added to the communities. Following such an iterative strategy the algorithm is more efficient in finding smaller communities than the algorithm by Newman and Girvan (2004) which in some cases fails to identify tightly connected small groups as new communities.

The communities c_{i_j} and c_{i_k} in the snapshot s_i include non-overlapping sets of users (forum users, eTwinnning users, Wikipedia users) in the time interval $interval_i$.

All users from communities c_{i_j} and c_{i_k} are in the set of users U_i

The snapshots s_i, s_r are defined by time intervals $interval_i \neq interval_r, i \neq r, i < r, r = i + 1$. Community users of these snapshots c_{i_j} and c_{r_k} may belong to sets of users in both snapshots

$$c_{i_j}(U) \subseteq U_i, c_{r_k}(U) \subseteq U_r$$
$$c_{i_j} \in s_i \ , c_{r_k} \in s_r$$

I map communities if the communities appear in consecutive snapshots and the modified Jaccard similarity (Gliwa et al., 2012) of the communities meets a given threshold.

$$Sim(c_{i_j}(U), c_{r_k}(U)) = max(\frac{c_{i_j}(U) \cap c_{r_k}(U)}{c_{i_j}(U)}, \frac{c_{r_k}(U) \cap c_{i_j}(U)}{c_{r_k}(U)}) >= threshold$$
$$threshold \in \{0, 1\}$$

If communities exist only in one snapshot, they remain unmapped. This approach finds evolving communities where the number of users increases or decreases. But other community evolution events are left without consideration such as splitting and merging of communities.

3.4.3.3 Community Detection and Evolution in a Distributed Environment

Many existing community detection algorithms are computationally intensive (Fortunato, 2010). Recently a number of algorithms appear that can be executed in distributed environments. These and other community-related algorithms use matrices

for their computations. Fatahalian et al. (2004) explained the efficiency of GPU algorithms that are working with matrices. For example, computation of modularity for community estimation with the Louvain algorithm (Blondel et al., 2008) is up to sixteen-fold faster than the same algorithm implemented on the CPU.

Propinquity algorithm We choose the algorithm from Zhang et al. (2009) for community detection with propinquity dynamics. It defines quality characteristics for each edge in a network. In each algorithm iteration it decides if an edge should disappear or remain as well as whether new edges should be added. After several iterations more edges appear within densely connected groups of nodes while edges within sparsely connected groups disappear. Therefore, after some iterations community structures become more obvious than before. In Figure 3.13 we present an example of changes after an iteration of the algorithm for all existing and possible edges. The edge $(1, 3)$ appears as its nodes have similar neighbors 0 and 2 and these neighbors are connected. The edge $(3, 4)$ disappears as edge nodes do not have similar neighbors.

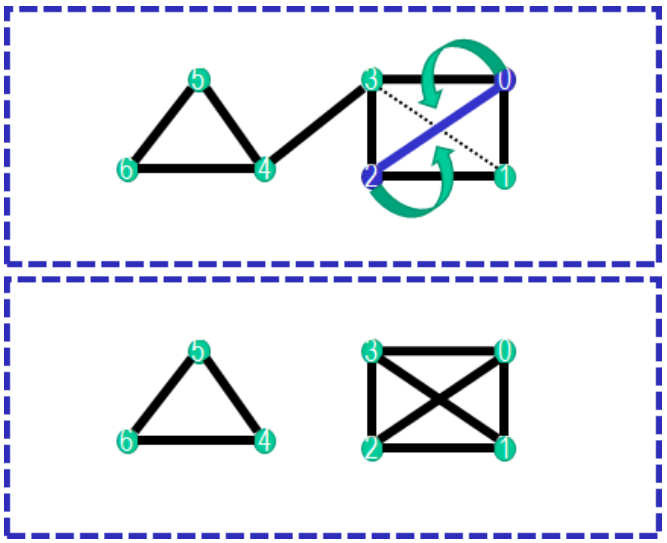


Figure 3.13: Example of propinquity value estimation

The pseudo-code of the community detection in GPU is presented in Algorithm 1 and 2. Propinquity values are defined in three steps. First of all, each of the existing edges gets a value of 1. In the second step we check the number of common neighbors of two nodes (*Couple Increment*) and in the third step we check the number of edges between these neighbors.

First of all, degrees of nodes are calculated in a parallel way and nodes are sorted according to degrees (Algorithm 1, ll. 3-4). Based on the degree, low or high, next steps are chosen. In most of the cases degrees of nodes are low. Then the function $generatePairsLowDegree$ finds neighbors of a node as it is depicted in Figure 3.14. The neighbors are sorted and new pairs are created by connecting one neighbor to the

Algorithm 1: Propinquity Algorithm in Pseudocode, Propinquity Handling

```
   input  : A set of edges and nodes in a network, Thresholds (k, α, β), Number of
            iterations t
   output: A set of sets of nodes (communities)
 1 for iteration 1..t do
 2     for each node do                          // in parallel threads
 3         node.calculateDegree();
 4     nodesWithDegree = sortNodesByDegreeAsc(nodes);
 5     for degree = 1..k do
 6         generatePairsLowDegree(nodesWithDegree(degree));
 7     for degree > k do
 8         generatePairsHighDegree(nodesWithDegree(degree));
 9     countNumberOfPairs();
10     for each pair do                          // in parallel threads
11         cn = findCommonNeighbors();
12         findEdges(cn);
13     for each pair do
14         if pair.propinquity > α then
15             edges.add(pair);
16         else if pair.propinquity < β then
17             edges.delete(pair);
```

next one. After such pairs are created, the new iteration creates a next set of pairs where neighbors connect with second next neighbors. The procedure continues until the first neighbor is connected to the last one.

In the case of nodes with high degrees the function $generatePairsHighDegree$ is used to create pairs as it is depicted in Figure 3.15. Such a function for high-degree nodes minimizes the usage of memory since it releases memory from nodes that are no more used for creating pairs. Pairs are generated for each node with all next neighbors thus generated pairs are sorted. Both approaches for pair generation produce the same pairs where some of pairs can be existing edges. The difference between both approaches is in usage of threads. In first case, one thread is attached to each node where a thread generates pairs consequently. In the second case, each pair of neighbor nodes of a high degree node is generated by a thread.

In the Algorithm 1 both approaches are described in ll. 5-8. In the next steps we find out neighbors of pair nodes and whether these neighbors are connected to each other (ll. 11-12). First of all, common neighbors of each edge node are detected (l. 11) in a parallel way and after that existing connections between the neighbors are defined (l. 12). Finally, we can compute the propinquity value for each pair (ll.13-17)

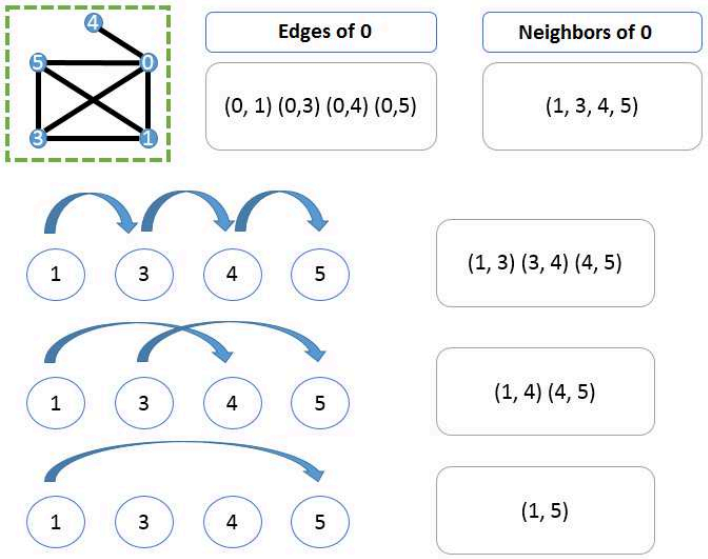


Figure 3.14: Example of propinquity value estimation for nodes with low degrees

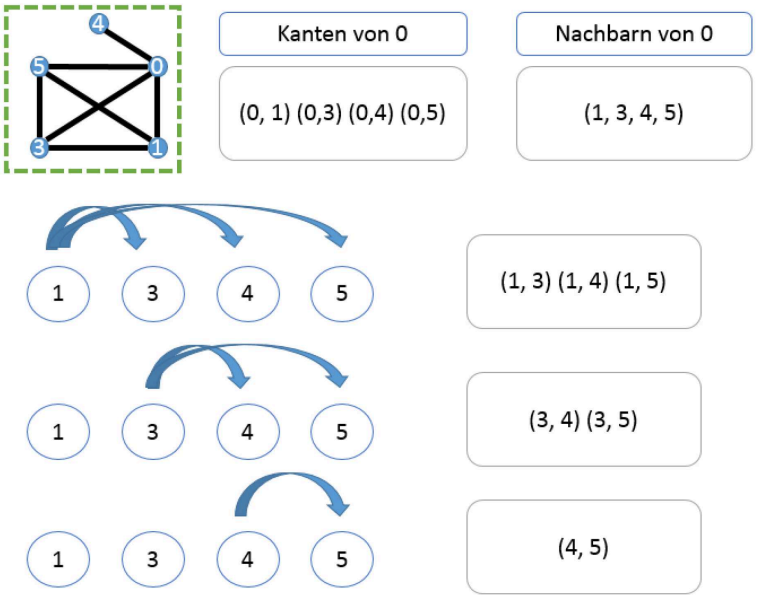


Figure 3.15: Example of propinquity value estimation for nodes with high degrees

and decide if we need to add the pair (l.15) or delete the pair and consequently the corresponding edge (l.17). After the set of edges is updated, they are used again for a new iteration.

After a given number of iterations is conducted, communities need to be specified

Algorithm 2: Breadth-first search of communities for the propinquity algorithm

```
   input  : A set of edges E and nodes V in a network
   output: A set of communities C
 1 create initially empty queue Q;
 2 initially empty sets of temporary visited (TMPV) and total visited (TV) nodes;
 3 initially empty set of communities C;
 4 move nodes without neighbors to TV;
 5 while TV ≠ V do
 6     select random node v from V;
 7     Q.enqueue(v);
 8     add v into TMPV;
 9     while Q is not empty do
10         v = Q.dequeue();
11         for each neighbor w of v do
12             if w is not in TV then
13                 Q.enqueue(w);
14                 add w into TMPV;
15     C.add(copy of TMPV);
16     TV = TV ⋃ TMPV;
17     TMPV = ∅;
```

in the resulting graph that includes a set of isolated communities or strongly connected groups of nodes. We use a breadth-first search as described in Algorithm 2 for this purpose. There we operate with two sets: a set of totally visited (TV) nodes, a set of temporary visited nodes ($TMPV$) and a queue Q (ll. 1-3). Nodes without neighbors directly appear in the TV set (l. 4). After that, an unvisited randomly chosen node is added to the queue and the $TMPV$ list (ll. 7-8). Presence of a node in the sets is checked in a parallel way. All its neighbors that are unvisited nodes are added to the queue (ll. 11-14) and to the $TMPV$ set while v is deleted from Q (ll. 10). Until any node exists in the Q, the process of neighbors adding continues (ll. 9-14). In other words, the while iteration stops when all connected nodes are reached and added into the $TMPV$ set. After that, we retrieve the nodes from the $TMPV$ set and store them as a community (l. 15). If any other unvisited nodes still exist, the breadth first search will run again (ll. 5-17). Currently the algorithm defines only non-overlapping communities. Edges that are deleted in Algorithm 1 can serve as sources for defining overlapping communities.

Event extraction

To follow changes of communities we adopt the event algorithm from Asur et al. (2009) where different events of community evolution are defined. Community events are *dissolve, form, merge, split*, and *continue* while events devoted to nodes are *appear,*

disappear, *join*, and *leave*. Examples of these events are shown in Figure 3.16. The first row depicts how a community continues to exist. The second row explains the process of merging while the third denotes the process of splitting. The event of forming defines the creation of a community that did not exist before. The event of dissolving represents dying of a community that existed before. Furthermore, we follow events of nodes such as the appearance of nodes, the disappearance of nodes, joining of nodes to a community (4 in Figure 3.16), and nodes leaving a community (4 in Figure 3.16).

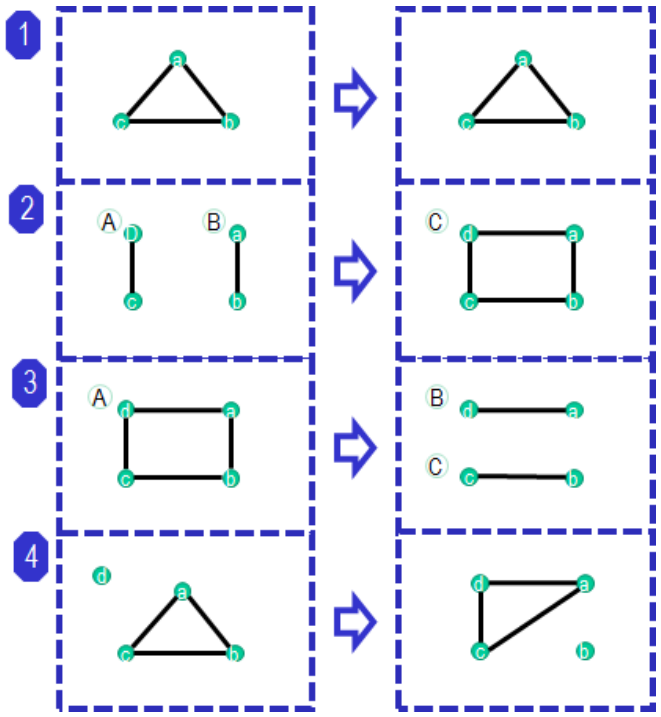


Figure 3.16: Examples of events for communities and nodes

Algorithm 3 describes how communities are compared with other communities from the next snapshot. First of all, we create vectors of communities in each snapshot based on nodes ids. After that, we create a matrix where rows are communities from both snapshots while columns represent ids. Therefore, communities are represented by bit vectors, e.g., 0 1 1 1 denotes the community with nodes 1,2,3.

According to the adopted algorithm of Asur et al. (2009), the calculation is realized by simple bit operations AND and OR. Costs of the algorithm execution should not be underestimated. Since all communities k_1 of a snapshot are compared to all communities k_2 of the other snapshot we peform $k_1 \times k_2$ calculations for only one event type, e.g *continue*.

For the *form* and *dissolve* events communities from a previous snapshot are compared to communities from the current snapshot using the AND operation. If no community exists that can be compared to another community, it indicates the creation of

Algorithm 3: Event Detection Algorithm

```
input : A list of snapshots with communities, thresholds for merge, split,
        continue events
output: Events occurring between consecutive snapshots
prev = pick first snapshot;
for each community do                          // create one-line vector
    create community vector;
for snapshot 2...k do                  // k is the number of snapshots
    cur = snapshot;
    for each community from cur ⋃ prev do
        create community vector;
    create matrix;
    // coefficients to define form, dissolve,
       continue, merge, split, appear, disappear, join
       and leave
    for each community vector C_1 in previous snapshots do
        for each community vector C_2 in current snapshots do
            calculate C_1|C_2;
            calculate C_1&C_2;
            for each community do
                define A, A*, B, D, D*, E, F, G, H;
                define events();
    set cur as prev;
```

a new community, if communities from previous snapshot were compared to communities of current snapshot, or the death of a community, if communities of the current snapshot were compared with the communities of the previous snapshot. The *continue* event is realized using the AND and OR operations to find nodes that appeared in both snapshots. The detection of *merge* and *split* events is more complex than others and happens in two steps. To detect the *merge* event all possible combinations of communities in the first snapshot are merged with the OR operation. After that, the combinations are compared to communities that have a similar number of nodes and appear in the second snapshot. If a community has at least fifty percent of nodes of a combination it is the *merge* event. To detect the *split* event we need to conduct the same operations but change the snapshots: combinations are detected in the second snapshot and these are compared with communities from the first snapshot. In case a combination has a similar number of nodes and a community has at least fifty percents of nodes, we can talk about the *split* event. Node events such as *disappear* and *appear* are detected by finding nodes that appear in one of the snapshots only. Other node

events such as *join* or *leave* are detected comparing nodes of a community with its follower where new nodes in the follower *join* the community while no nodes in the follower indicates that the nodes *left* the community.

3.5 Summary

Supporting learning community needs requires an appropriate process where learning context, facilities of learning media, and other players in the learning environment are taken into consideration. We recognized that to assist learning communities one-sided investigation, either technological or learning-theoretical, is not sufficient. Therefore, we appealed to the information modeling approach *i** (Yu, 1995) that is used for early requirements engineering to detect community needs. Furthermore, *i** pays attention to the usage world (Jarke et al., 1992) of an information system that allows to represent learning communities with their users appropriately.

Our composite approach involves both the knowledge of learning theories and the analysis of learning environments for modeling of learning communities using the process of community model creation. We based this process on the ATLAS methodology (Klamma et al., 2006a) that accesses community needs in information systems continuously. The process consists of four phases: *modeling*, *refinement*, *monitoring*, and *analysis*. These can be iterated to reduce continuous modeling of learning communities.

For modeling learning communities we firstly need a general model that will suit any learning community in social media. We consider learning communities as *Communities of Practice* (CoP) (Wenger, 1998) that exchange knowledge using social media. Since community models can not be created without prior investigations (a cold start problem) we described specific models that can be used by community stakeholders as a starting prototype. They can refine their communities according to the prototype and prove the effectiveness of changes by modifying models and validating them through simulations that can predict possible outcomes of the changes. During *monitoring* and *analysis* phases community states are extracted while these states are used to create community models. Based on the states and models, the stakeholders can estimate community issues and success that can be shared with other communities.

To the our best knowledge, our process of community model creation provides the only, solution for continuous learning community modeling in social media that reveals community needs and states. Considering different dimensions of CoP and in doing so connecting community vision from learning theories and data science is another contribution of our methodology. Following these dimensions we built graphs of learning communities based on their collaborations and assessed learners' social network measures (Wasserman and Faust, 1994). Moreover, we detected communities in graphs and their evolution using Louvain (Blondel et al., 2008), propinquity dynamics (Zhang et al., 2009) and event detection (Asur et al., 2009) algorithms. We further investigated community text by detecting community goals, emotions and topics. The

storage of such data is challenging since we need to execute complex queries in short time to provide rapid community modeling. Therefore, we design a data warehouse with a multidimensional model based on the Mediabase model (Klamma, 2010) that emphasizes important actors of social media.

The application of the presented methodology is explained in the next chapters. Firstly, we present the data warehouse solution, after that we discuss three case studies that apply the phases for detecting community needs, states and models. These are evaluated by different types of stakeholders.

Chapter 4

Mediabase Cube: A Data Management Solution for Learning Communities in Social Media

In the previous chapters I have shown shortcomings of existing approaches for data collection and storage. Efficient modeling of learning communities is possible with a well-designed and fast data management solution. Data warehousing allows to store and operate with historical data and Online Analytic Processing (OLAP) operations. This is possible due to data warehouse design that allows aggregation and other operations of the data that let us perform complex queries. In this chapter I firstly introduce the questions that are asked by community stakeholders and that can be solved by complex queries. After that, I present the core and design of the snowflake schema of the Mediabase cube described in Section 3.3.3.1. Later I discuss examples how the cube is used while supporting learning community needs.

4.1 Accessing Community Needs

In the time of World Wide Web we have produced tremendous amounts of data every year and the amount of produced and replicated data will grow roughly by a factor 44 in 2020 (Gantz and Reinsel). This happens since every actor in the Web is creating a data, whether it is a human or not.

While sharing data in social media peers organize online communities (Preece, 2000). We consider the communities as Communities of Practice (Wenger, 1998) since their members negotiate about common practices. Communities differ in the number of members, topics, variety of used media and other characteristics and thus each community requires a special support. Such a support can be realized by same tools that operate with data stored under the same schema. For example, all communities can be represented by graphs that can be investigated by the same implementations of the same algorithms; all communities include texts that are analyzed in the

same manner using information retrieval techniques. The Mediabase model provides application-independent and cross-media views on data (Klamma, 2010) that allows to apply similar analytical tasks to data from different sources. But we need not only efficient models but as well an efficient implementation. Therefore I choose the data warehouse solution since it can respond to complex queries in the near real-time.

In the following, I introduce a number of questions that are of community stakeholders interest. I classify the questions into phases of the process of community model creation (Figure 3.2) while the questions consider that the community data has already been analyzed with techniques described in Section 3.4.

Modeling communities

- Which users are members of a community C?
- What is the next state of a community C, i.e., does the community C have a continuation?
- What are user patterns (roles) in communities of a medium M?
- What type (pattern) does a community C belong to?

Refinement of models

- What is the highest and the lowest cognition rate for communities that have the same number of users and appear in the same medium M as a community C?
- Have communities with a given number of users and intents stay alive (merge, split, continue events for the communities)?

Monitoring communities

- Which is the most popular medium at a timepoint T?
- How often do users start threads in forums?
- What is the geographical distribution of school teachers?
- How many Wikipedia pages and revisions has each Wikipedia instances?

Analyzing communities

- Are there any communities that are active in both commenting in eTwinning and taking part in eTwinning projects?
- How many forum communities do exist in a network N?
- How did the number of communities change within the last years in Wikipedia?

- Which users possess powerful positions and connect disconnected groups together (have high betweenness centrality) in a medium A in a time interval B?
- How many users create artifacts that express negative moods?
- What is the number of intents expressed in forum posts?
- What sentiment rates do users with a pattern P express while creating threads in forums?
- What are characteristics of users with a pattern P that are coming from Germany?

All these questions can be answered by performing complex queries in the data warehouse. To conduct them efficiently, I design the snowflake schema regarding the Mediabase model. In the following, I firstly explain the data warehouse design and after it I introduce the snowflake schema with some examples of its usage.

4.2 The Core of the Data Warehouse

The Mediabase data warehouse (DW) requires several processes to be performed for its realization. Figure 4.1 depicts these processes and the DW structure. The extraction - transformation - load (ETL) process is realized by the forum, Wikipedia and eTwinning watchers (Section 3.3.4). The first two watchers are Perl scripts while the eTwinning watcher is Java-based. The watchers not only collect but as well clean data from irrelevant characters and transform it for the storage according to the relational schema (Figure 3.8). We analyze stored data using our methodology described in Section 3.4 and the outcomes are stored according to a snowflake schema in the Mediabase DW. The global DW includes all data while the recent data appears in the operational data source. A data mart provides data required for a particular set of tasks. Each application usually has a data mart that requests data from the global DW.

4.2.1 Snowflake Schema

A *star* or *snowflake* schema guarantees efficient design for an information model of a data warehouse. Such a schema consists of one or more fact tables and a set of dimension tables.

Figure 4.2 depicts the Mediabase snowflake schema that has one fact and five dimension tables. The fact table includes foreign keys relating the table to other dimension tables to facilitate exploration of the data cube. It stores as well to_user_id to enable creation of social networks between users that explicitly mention other users in their artifacts, e.g., in headers of e-mails. The *Medium* table includes the $type$ attribute that defines a specific type of media (forum, Wikipedia and eTwinning). Forums can consist of subforums that have their own topics. Wikipedia includes different instances

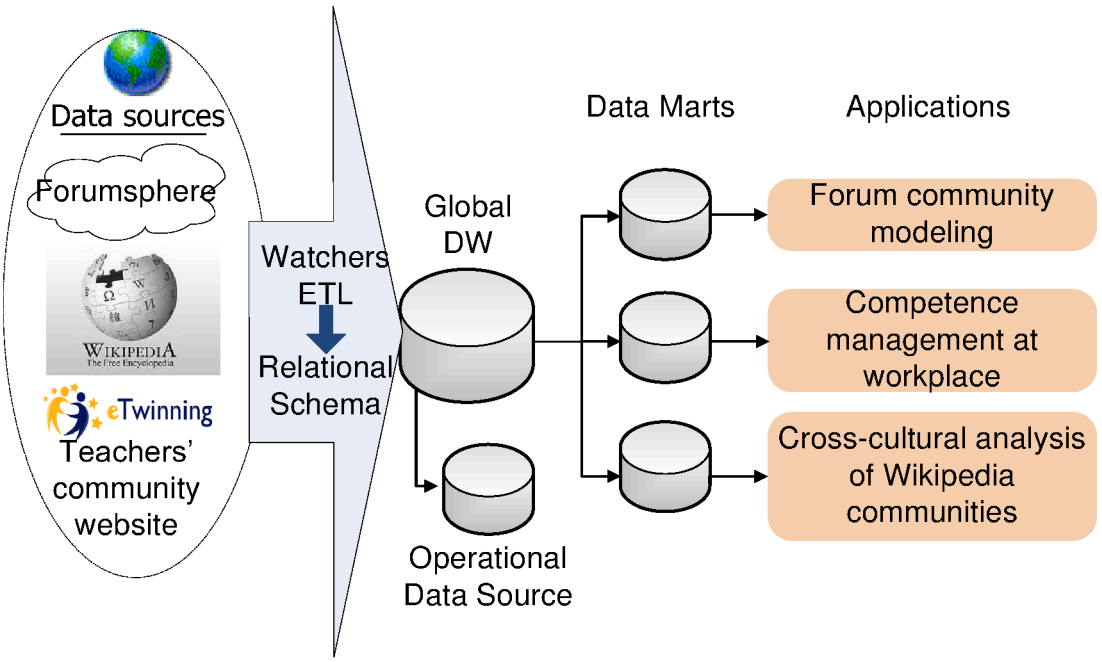


Figure 4.1: The Mediabase warehouse design and applications

that are characterized by $URLs$ and languages while eTwinning has only one instance. The *Artifact* table includes an id and a corresponding $type$ of an artifact. Using types we can request the *Artifact* tables directly, for example, we can query for all articles in Catalan Wikipedia. A $type$ in the *Process* table specifies a type of process according to transcriptivity theory (Jäger et al., 2008). The *Process* is included since it is one of main actors in the Mediabase, though it has not been used explicitly in this work because of complexity of type assignment to activities in social media. In the *User* table we can drill down to other levels of the *User* dimension with the help of $community\_id$, $network\_id$ and $school\_id$. For example, with a given $community\_id$ we can retrieve all facts devoted to users of a community and find, for example, which media they use. The $school\_id$ field is relevant for eTwinning data only, where users are teachers in schools that are active in the eTwinning portal. The eTwinning users must have a $school_{id}$ while for other media, users can be members of networks only. The *Time* table consists of years, months organized in quarters, weeks, day of week and time of day and a time point.

Artifacts can be connected to each other. A row in $to\_artifact\_id$ indicates artifacts related to a fact, e.g., a revision of a Wikipedia page is related to the Wikipedia page. Wikipedia pages refer to the *Revisions* table as the last approved revision of a Wikipedia page is the final Wikipedia page at the current time. In the *Artifact* hierarchy e-mails or forum posts refer to threads. All other artifacts refer to *URL* artifacts that have URL addresses; these are threads, posts, Wikipedia pages, e-mails in eTwinning

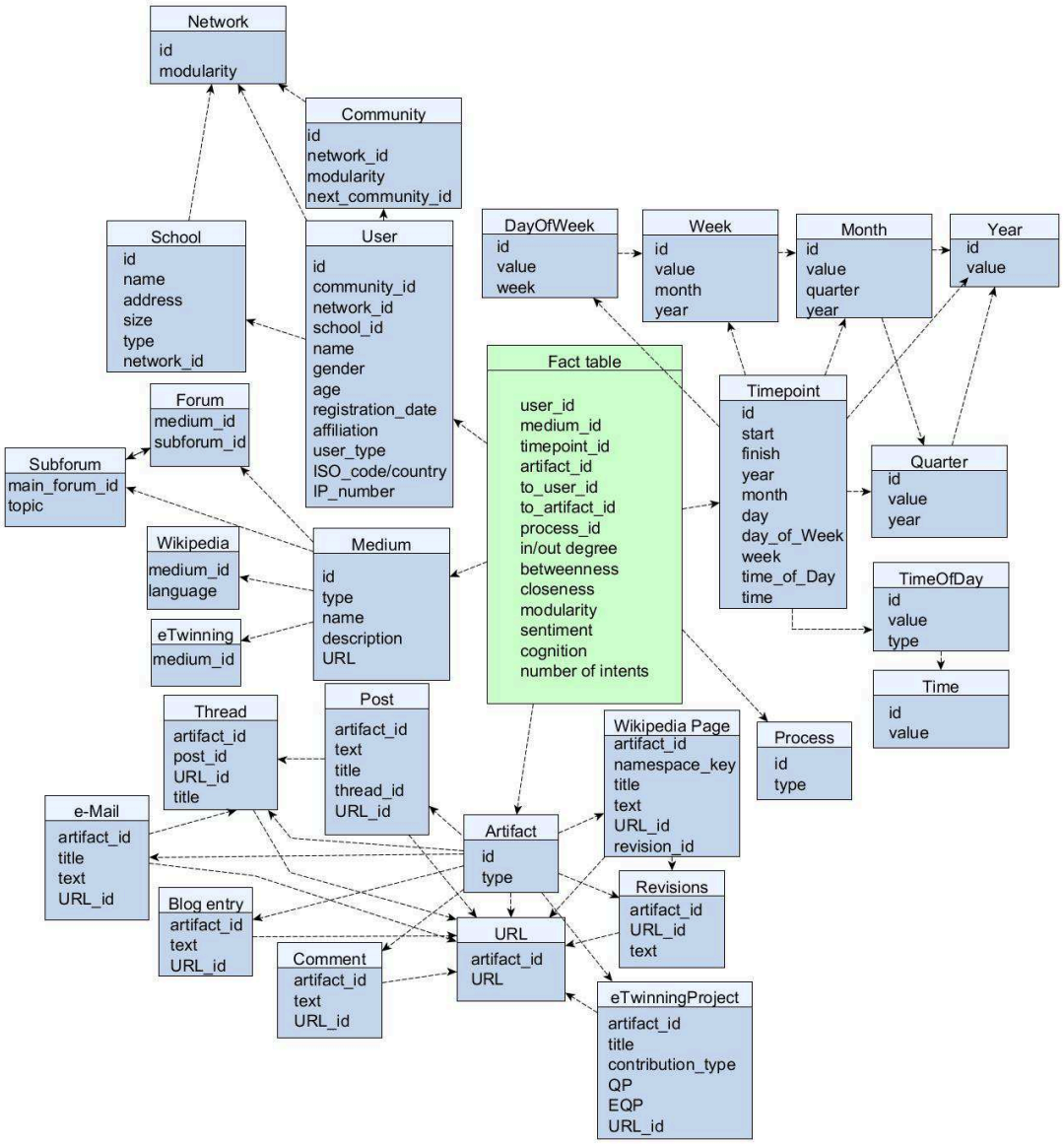


Figure 4.2: Snowflake schema of the Mediabase cube

and more. These references are as well realized through $to\_artifact\_id$.

4.3 An Example

Table 4.2 presents an example of the content in the Mediabase cube. For the sake of clarity I explain values that are behind the constant names in Table 4.1. Anna (a_1) posted a forum message (p_1, ar_1, m_1). Furthermore, she created a project (p_1 and ar_2) in eTwinning m_2. The time points of creation of these artifacts are different.

After creating a project, Anna created a post (p_1 and ar_1) in eTwinning. Dominik (a_2) created a post in the forum at the same time as Anna created the forum post. Ralf (a_3) performed two processes in the forum. He posted a post and shared (p_2) it. Mohsen

Dimension	Type	ID	Description
Agent	Human	a_1	Anna
Agent	Human	a_2	Dominik
Agent	Human	a_3	Ralf
Agent	Human	a_4	Mohsen
Process	Transcription	p_1	Create
Process	Addressing	p_2	Share
Artifact	Post	ar_1	Forum message
Artifact	Project	ar_2	eTwinning project
Medium	Forum	m_1	URCH
Medium	collaborative workspace	m_2	eTwinning

Table 4.1: The mapping between real values and their ids

(a_4) has only made a post in the forum.

Fact values in Table 4.2 refer to the described events. The values are in/out degree, betweenneess, connectiveness, sentiment and cognition scores, and number of intents. For some cells all these measures exist, while for some such as ar_2 no number of intents, sentiment and cognition scores are enabled (last three zeros in the fact set). Anna created two posts in the same forum and her measures specifying her position in a forum network (structural measures such as betweenness) slightly changed while measures classifying her posts (semantic measures such as sentiment) were different. This is different to Ralf's values where his structural measures in row 6 changed a bit while semantic measures describing his post were the same in rows 5 and 6 since he operated with the same post.

4.3.1 End-user Operations

Data cubes support a number of operations used for querying.

- **Aggregation** or **Roll up**: the data is collected from different dimensions or within dimensions. For example, if we are looking for a user who explores several media, we need to aggregate facts about the requested user.

  Table 4.3 shows the output for the query of the given *Agent* dimension. If a user name is given, the cube includes aggregated facts about the user, e.g., the user can be presented in different media or the user is active in one media creating different artifacts. If we aggregate according to a community, we get the aggregation of facts of all users that appear in the community. The same procedure is

Dimensions						Facts
Number	Process	Artifact	Medium	Time	Agent	
1	p_1	ar_1	m_1	t_1	a_1	$f_1 = \{6, 3.5, 1.6, 0.1, 0.7, 3\}$
2	p_1	ar_2	m_2	t_2	a_1	$f_2 = \{14, 71.6, 3, 0, 0, 0\}$
3	p_1	ar_1	m_2	t_3	a_1	$f_3 = \{7, 3.5, 1.6, 0.5, 0.0, 1\}$
4	p_1	ar_1	m_1	t_1	a_2	$f_5 = \{5, 48.2, 2, 0.3, 0.8, 7\}$
5	p_1	ar_1	m_1	t_4	a_3	$f_6 = \{67, 12, 6, 0.7, 0.4, 5\}$
6	p_2	ar_1	m_1	t_5	a_3	$f_8 = \{68, 11.8, 6, 0.7, 0.4, 5\}$
7	p_1	ar_1	m_1	t_1	a_4	$f_9 = \{1, 0.8, 0.3, 0.2, 0.6, 3\}$

Table 4.2: An example for a subset of cells from the Mediabase Cube

relevant for school and network dimensions. The result of the aggregations over the *User* dimension is the collection of all related facts.

User dimensions	Aggregated facts
user a_1	$\{f_1, f_2, f_3\}$
community $C_1 = \{a_1, a_2\}$	$\{f_1, f_2, f_3, f_5\}$
school $S_3 = \{a_1, a_3\}$	$\{f_1, f_2, f_3, f_6, f_8\}$
network $N_2 = \{a_1, a_2, a_4\}$	$\{f_1, f_2, f_3, f_5, f_9\}$

Table 4.3: An example for aggregation over the *Agent* dimension considering cells from Table 4.2

- **Roll down** or **Drill down**: these are queries for more fine-grained data. With their help a data cube is explored in more details. Using this operation one may query for a specific measure, i.e. move down according to the hierarchy. This operation is opposite to the aggregation operation. For example, using this operation we can query a cube that aggregate all dimensions over communities and in our query drill-down to users of communities. for communities of a user.

- **Screening** or **Filtering**: these operations set restrictions of the retrieved data based on criteria in dimensions. We can ask for all the facts that appeared last year where sentiment rates are high, e.g., > 0.8. An output will include facts with references to artifacts that include texts with positive sentiments.

- **Slicing**: single or more values of a particular dimension are specified. For example, a user A is a slicing condition. Then we query for all remaining dimensions and retrieve all facts related to the user A. Particularly, we retrieve all artifacts and processes in all media in all time periods in all media communities, schools or media networks. Such information is relevant for estimating the learning progress of user A.

- **Pivot**: this operation allows to compare facts according to a dimension that is chosen as an independent variable. Using the pivot operation stakeholders can compare facts about communities to find similarities and differences between them.

4.4 Applications

One of the applications that exploit the Mediabase cube is the i*-REST services that models learning communities (Petrushyna et al., 2014b). Resulting models presented in the next chapter and (Petrushyna et al., 2015) can be used for software developers to discover requirements of communities for social media (Hilts and Yu, 2011), for community stakeholders to develop appropriate recommendation strategies and applications (Brusilovsky, 2001), and for experienced community members to observe community situations. Semantic measures were used to provide forum users with information about learning goals (Krenge et al., 2011). Recorded collaborations of teachers and their analysis in the Mediabase Cube help to reveal patterns of their behavior and allow to compare them with behavior of other peers Chapter 6). Collected activities and Wikipedia communities in the Mediabase cube let to observe differences in collaborations of contributors coming from different countries (Chapter 7).

Here we present further findings we do by exploring data from the Mediabase cube. Community stakeholders can be interested in betweenness scores (Section 3.4.1) of community users in communities with more than 50 users but less than 100 users.

```
select betweenness from facts where user_id in
(SELECT user_id FROM users
group by community_id having count(user_id) > 50
and count(user_id) < 100)
```

Listing 4.1: Mediabase cube query for betweenness of community users in middle-sized communities

This query helps to understand the distribution of betweenness centralities in communities. Users with high betweenness centralities possess broker positions (Burt, 1992), where they connect isolated groups and therefore serve as important chains in transfer of knowledge between the groups. The query reveals if communities of a given size include users with *broker* positions. If yes, researchers can estimate a ratio of brokers to other users in the communities. Investigating community developments one can compare how the role of brokers changes in evolving communities.

```
select count(distint community_id) from users where user_id in
    (select user_id from facts where number_of_intent >5)
```

Listing 4.2: Mediabase cube query for the number of communities with users that express at least 5 intents

The query in Listing 4.2 is counting the number of communities where at least one community member expressed more than 5 intents in one message. The intentions indicate that 1) users express goals or 2) community members formulate their advice by expressing intentions.

Results of sentiment and cognition rates (Section 3.4) for users with different activities are depicted in Figure 4.3. I differentiate between *popular* users that participate in more than one hundred communities and *rare* users that participate only in four communities. Users participating less than four times are less engaged and therefore I omit them from the comparison. In the comparison I explain small differences between mean values of sentiment and cognition rates of both user types.

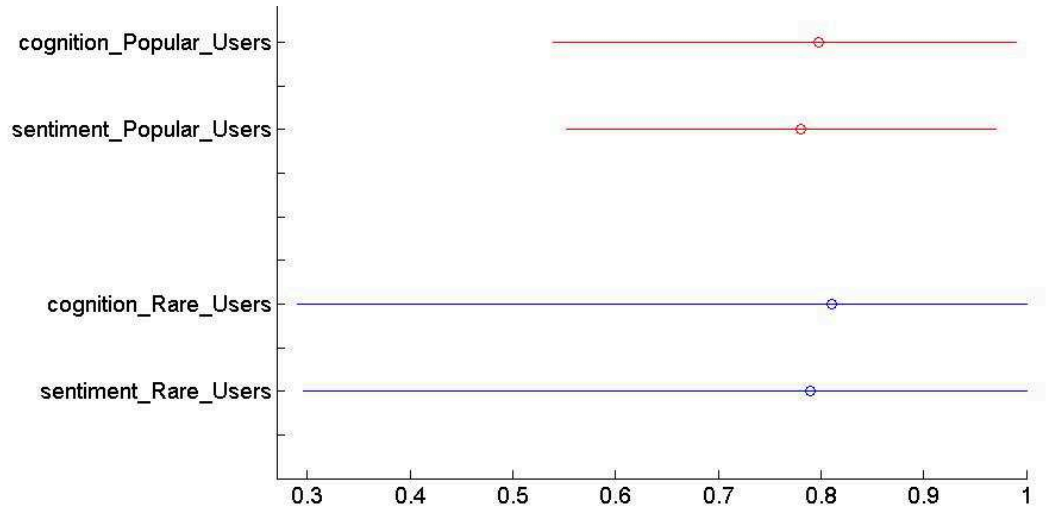


Figure 4.3: Comparison of user sentiment and cognition rates of popular and rare users.

Figure 4.4 shows the distribution of intents for communities of different size. The mean of intents per user, denoted as a point, is similar in all communities, while the range of the measure is different and relatively low for smaller communities.

I classify communities according to the number of members (2 for small and 93 for huge) and compare their closeness in Figure 4.5. Users in huge communities have the most broad range of closeness (Section 3.4.1), i.e. their members are both close to and far away from the center of a network. In other words, such communities include both new members with low closeness and active members in a network. The mean value of closeness is the highest for the huge communities where a probability is high that a member has high closeness.

The last application in Figure 4.6 compares measures of a community user that has stayed the longest time active in a forum and participated in 359 communities with

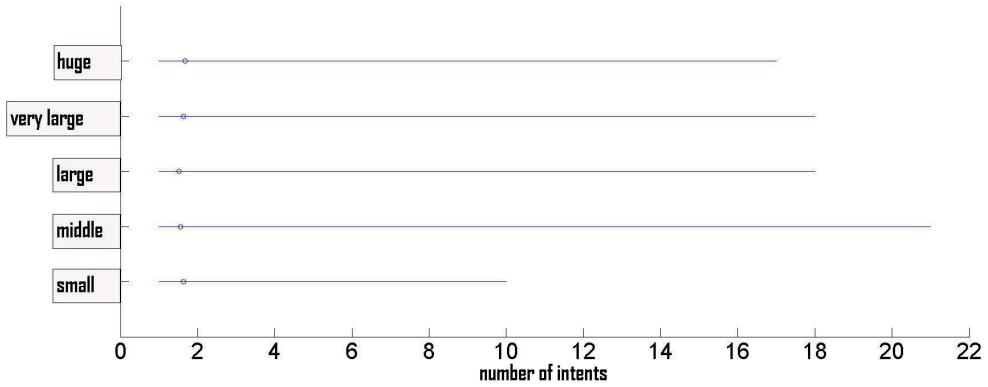


Figure 4.4: Comparison of the number of intents in communities of different size

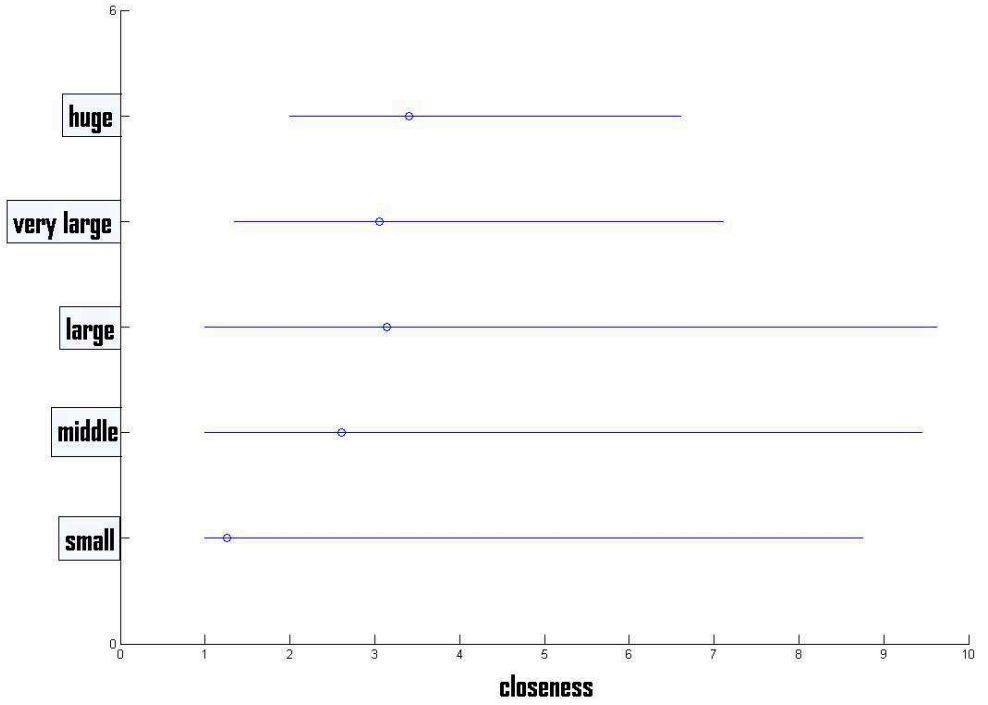


Figure 4.5: Comparison of user closeness in communities of different size.

middle-class forum users that participated in 50-100 communities in the whole period of forum observation. The values for the community users are depicted in red while the values of others are depicted in green. In the beginning of our observation (within 5000 time intervals — x axes) red dots are prevailing; just a few of the middle-class forum users appeared in the forum in this time period. Later (5000-15000 time intervals) the number of green dots is expanding extremely, i.e. the closeness and betweenness as

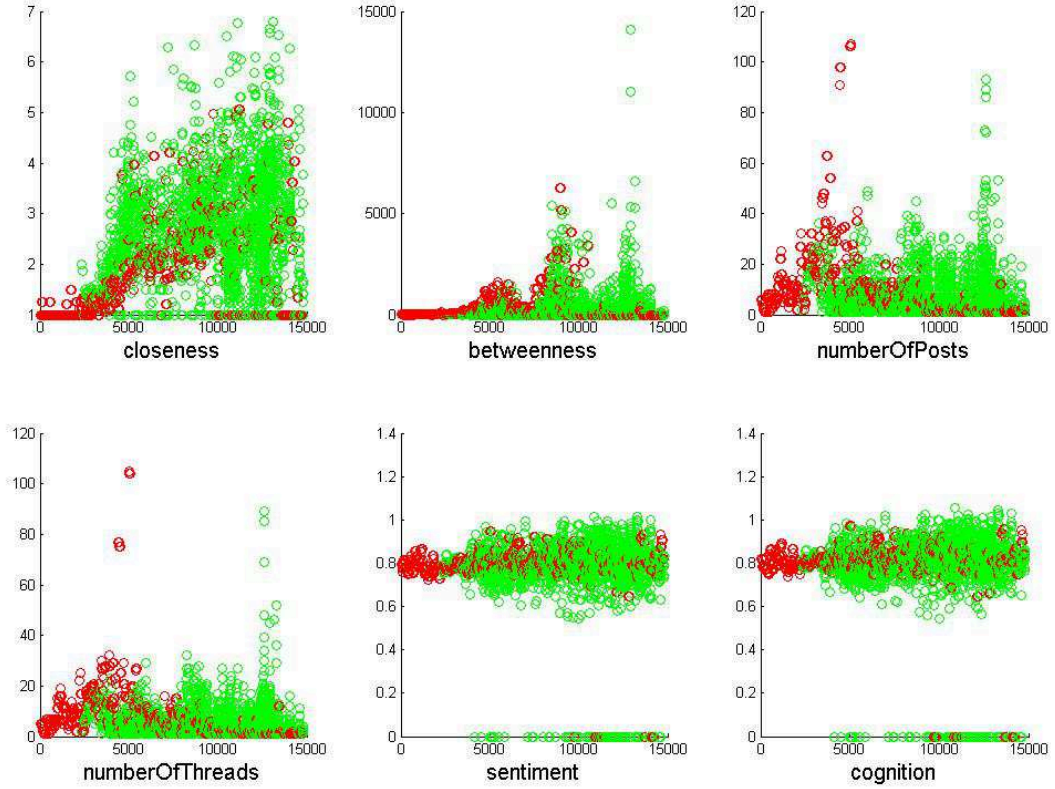


Figure 4.6: Comparison of the community monk (red dots) with middle-class forum users (green dots). X axes states for time and Y axes states for numerical values of measures

well as the number of posts and threads of the middle-class forum users are growing.

The bottom part of Figure 4.6 shows the difference in distribution of semantic measures though in this case these are highly influenced by quantity of user posts. Since the user has much less posts than the middle-class forum user, the number of sentiment and cognition words as well as the number of intents is higher for the middle-class users, though in the beginning of observation the user sentiment measures exceed other user measures.

The number of the middle-class forum users with high betweenness in Figure 4.6 can indicate the growth of the number of isolated communities that the users span. Both betweenness and closeness distributions show that in the beginning of the forum existence the most active user took a leading role in community discussions while later the middle-class forum users had an influence on forum existence and development. Statistical data about the number of posts and threads in Figure 4.6 supports the fact

that the middle-class forum users have become more active. Based on these observations we can state that after 5000 timeunits the forum became stable as it reached a number of the middle-class forum users sufficient for helping forum communities to sustain. The investigated forum has a high fluctuation of users thus the middle-class users are changing continuously but it does not influence community sustainability since the required number of supportive users has been reached.

4.5 Summary

Reviewing existing solutions for storing learning community data in the previous chapter we recognized the need for an efficient data management solution that provides application-independent and cross-media views on data. Although community needs are various we found that the tools we used for detecting community needs and states are the same. Furthermore, the tools execute complex queries and demand to get replies quickly. Thus there is a need for the schema and its realization that can be used for maintaining in an efficient way.

I started this chapter with examples of questions community stakeholders are interested in. To answer the questions complex queries need to be performed. Therefore, I have chosen to use the data warehouse technology to ensure efficient performance of queries. In particular, I replicated the Mediabase model (Klamma, 2010) to create a snowflake schema that represents the Mediabase Cube dimensions described in Section 3.3.3.1. Examples, end-user operations and applications of the Mediabase Cube were as well described in the chapter.

Storage of data and data models are usually designed together. To provide an efficient data modeling approach, it is required to select an appropriate data management solution that can deal with complex queries and context of stored data. The Mediabase Cube is the extension of the work of Pham and Klamma (2013) and it is the only solutions for social media storage considering application-independent and cross-media views. Our solution can be applied for storing data about communities that is later used for assessment of community states and needs, comparing these states with states of other communities, comparing community users, finding types of community users, extracting community topics and many other requests. In the next chapters, we show the usage of the data warehouse for modeling learning communities in forums, competence management of peers and communities in collaborative spaces and detection of culture-sensitive needs of communities.

Chapter 5

Continuous Modeling of Learning Forum Communities

In one of the previous chapters I have described the framework for supporting community needs. In this case study we implement each step of the framework. First of all, we realize a service that makes continuous modeling possible. It utilizes results of *monitoring* and *analysis* and thus can be used after these phases. Alternatively community stakeholders can observe their communities with one of classical community models (Section 3.1.2). The stakeholders can set hypothesis regarding favorable changes in the models. These can be validated with the help of simulations in the *refinement* phase. In the *monitoring* phase we collect and store data while in the *analysis* phase we perform the analysis of community data using community detection and evolution, intent analysis, emotional analysis, named entity recognition, and clustering. The outcomes of these phases are used to refine community models. In this chapter I discuss results of *modeling* and explain changes we captured in one particular community. After that, evaluation results are presented.

The contribution of the study is in continuous modeling of learning forum communities by combining results of community detection and community evolution algorithms with analysis of community user behaviors. We represent the results of *monitoring* and *analysis* as *i** models developing a modeling service and validate them by implementing a service for simulating these models. Furthermore, in this chapter we find that 40% of learners in the investigated forum follow the self-regulated learning process (Zimmerman, 1990) while others need guidance for their learning. *i** experts agree in appropriateness of usage of techniques for community analysis. The results of finding posts full of sentiments and words that evoke cognitive mechanisms are promising. Simulations of bigger communities (> 40 members) are close to results of real communities.

5.1 Forum as a Learning Community

Forums are popular between learning communities. Learners can share their knowledge, ask questions, get feedback and suggestions. Most forums include only basic functionalities like posting, quoting and creating a new thread. Some others classify users according to the amount of posts they published.

In this experiment we focus on two forum domains, language learning forums URCH[1] and forums of the Student Doctor Network[2]. Most of URCH forums' users prepare for the English tests such as TOEFL[3], GMAT[4] or GRE[5]. The users discuss exercises of the tests, share essays, write and ask for feedback from other forum users. We analyze the URCH forums where users deal with learning for the tests while we leave flame forums out of consideration.

Forums from the Student Doctor Network (StDocNet) are dedicated to all medical students where they discuss schools and admission exams, share interesting information, applications, questions, issues, and solutions. For our investigations we choose only forums where community members share learning experience, and have a purpose of learning or organizing their learning as they were preparing to specific medical exams.

5.2 Modeling

Because of absence of input data about communities we have to deal with a cold start problem in the *modeling* phase. To overcome it I propose classical models, where community stakeholders have to choose between *question-answer*, *dispute*, and *innovative* communities (described in Section 3.1.2) that exist in real forums (Wagner et al., 2012). An expected type of a community model is refined with new data and details after the monitoring and analysis phases (Figure 3.2). In the following I present the architecture of the service that creates *i** models.

[1]URCH forums `http://www.urch.com/forums/forum.php`, Last access on 20.04.2014

[2]Student Doctor Network `http://forums.studentdoctor.net/`, Last access on 20.02.2015

[3]English-language test `http://www.ets.org/toefl?WT.ac=toeflhome_why_121127`, Last access on 24.07.2014

[4]English-language test for admissions decisions into quality graduate business programs `http://www.mba.com/global`, Last access on 24.07.2014

[5]The only admissions test for graduate or business schools `http://www.ets.org/gre`, Last access on 24.07.2014

5.2.1 i*-REST Service

The i*-REST proposes creation of *i** models automatically allowing service-to-services communication. It is a couple of RESTful Web services (Fielding, 2000) that are designed to manipulate *i** models on-the-large.

Representational State Transfer or shortly RESTful services have become a standard for Web Services that provide solutions for other services online. Services are based on Resource-Oriented Architecture that consists of resources, Uniform Resource Identifiers, resource representations and linked resources. The RESTful services provide addressability (each resource is addressable), statelessness (each request is processed independently), connectedness (navigation between resources is possible), uniform interface (e.g., through usage of standard HTTP requests). The RESTful design has the following advantages: 1) machine and human readable, 2) clear structure facilitates development process, 3) statelessness allows to perform requests on different service instances and makes the use of any RESTful API easy.

The RESTful services of i*-REST allow to maintain *i** models by receiving REST requests from other applications. An *iStarML Model* i*-REST service stores models as IStarML files (Cares et al., 2011) in the XML database eXist[6] that allows to retrieve different versions of models. Moreover, an *iStarML Visualizer* i*-REST service transcribes models to SVG[7], which can be embedded in Web pages and visualized by any Web browser. Last but not least, the i*-REST services are implemented as part of an open source peer-to-peer environment, LAS2Peer[8], that hosts services for community information systems.

5.2.2 The i*-REST Infrastructure

The architecture of i*-REST depicted in Figure 5.1 allows the services to be distributed in peer-to-peer environments. The Web Connector[9] realizes the connection to outside and therefore can handle REST requests from outside and converts them into method calls of the *i**-REST services or other LAS2peer services.

The XML database serves as an *i** model repository. The eXist database allows storage of XML files and their versions efficiently, where only the difference information between old and new files and not entire files are stored. Access to models is restricted by a group management system realized both on the server of i*-REST and on the database.

[6]The eXist XML database, `http://exist-db.org/exist/apps/homepage/index.html`, Last access on 22.09.2014

[7]Scalable Vector Graphic

[8]LAS2Peer is a Java-based server framework for developing and deploying services in a distributed Peer-to-Peer (P2P) environment `https://github.com/rwth-acis/LAS2peer`, Last access on 22.09.2014

[9]Web Connector of the LAS2Peer `https://github.com/rwth-acis/LAS2peer-WebConnector/`, Last access on 4.02.2015

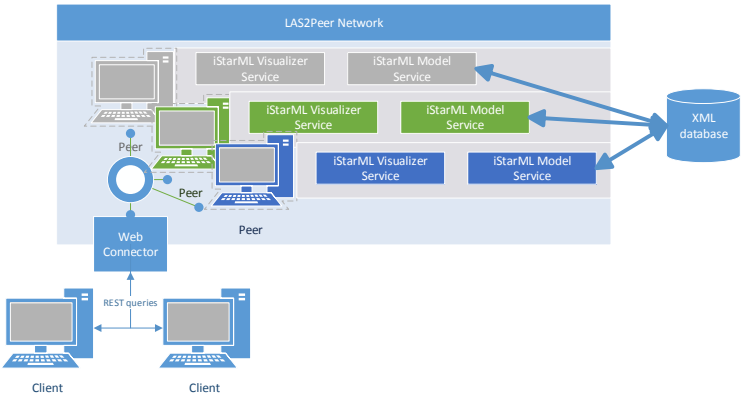


Figure 5.1: Overview of the i*-REST architecture.

The *iStarML Model Service* creates, modifies and retrieves *i** models. An example of an iStarML file is given in Listing 5.1. Each actor and intentional element node must have a unique ID to be accessible by the service (addressability for REST). Optional comment attributes are supported, which are later used in visualizations.

```
<istarml>
  <diagram name="diagram">
    <actor id="1" name="Homework" type="actor"/>
    <actor id="2" name="Teacher" type="actor"/>
  </diagram>
</istarml>
```

Listing 5.1: Initial model as an iStarML file

The RESTful API of i*-REST allows to create and modify *i** models using requests that are related to the iStarML syntax. The first request (Table 5.1) results in the code described in Listing 5.1. The goal dependency between both actors is created by requests 2-4 from Table 5.1 and will add the lines from Listing 5.2 to the initial model. Dependency links and the goal dependency are deleted by the last request. The complete list of supported REST requests can be found on the website of the i*-REST service[10]. The service keeps iStarML files always valid by rejecting invalid operations and maintaining consistency of models during modifications.

```
<ielement id="3" name="evaluate" type="goal">
    <dependency>
        <depender aref="1"/>
        <dependee aref="2"/>
```

[10]Description of the i*-REST service `http://istar.rwth-aachen.de/tiki-index.php?page=i%2A-REST`, Last access on 22.09.2014

```
        </dependency>
    </ielement>
```

Listing 5.2: An intentional element, its depender and dependee in iStarML notation

Request	Description
`GET Collection/model`	Returns model stored in a model repository.
`PUT Collection/model/ielement/3?type=goal &name=evaluate`	Creates an intentional element of type goal (ID=3) in the model.
`PUT Collection/model/ielement/3/depender/1`	Adds a dependency link between the actor (ID=1) as a depender and the intentional element (ID=3) in the model.
`PUT Collection/model/ielement/3/dependee/2`	Adds a dependency link between the actor (ID=2) as a dependee and the intentional element (ID=3) in the model.
`DELETE Collection/model/ielement/3`	Deletes the intentional element (ID=3) in the model.

Table 5.1: Examples of REST requests.

The *iStarML Visualizer Service* creates visual representations of *i** models by converting iStarML files into SVGs that can be embedded in arbitrary Web pages. The graph representation of a model is generated using the yFiles[11] library. It ensures a compact graph layout and usage of colors for nodes and labels. The color of nodes and labels can be set as optional parameters in iStarML files.

We create a web interface that allows user interactions with the model repository (Figure 5.2). The search functionality is realized by utilizing XQuery (Walmsley, 2007). A user can search for names of models, actors and dependencies.

The visualization representation can be navigated, similar to an online map, by dragging and zooming with a mouse or a keyboard. Comments specified inside an iStarML file are shown as tooltips on the top of a SVG representation. A user can download a visualized model as a SVG file and import a local iStarML file to store and visualize it.

Creation and visualization of community models is possible due to i*-REST services[12]. These services can be further used for automatic creation of *i** community

[11] yFiles for Java `https://www.yworks.com/de/products_yfiles_about.html`, Last access on 2.03.2015

[12] iStarML model service `https://github.com/rwth-acis/LAS2peer-iStarMLModel-Service` and iStarML visualization service `https://github.com/rwth-acis/LAS2peer-iStarMLVisualizer-Service`, Last access on 4.02.2015

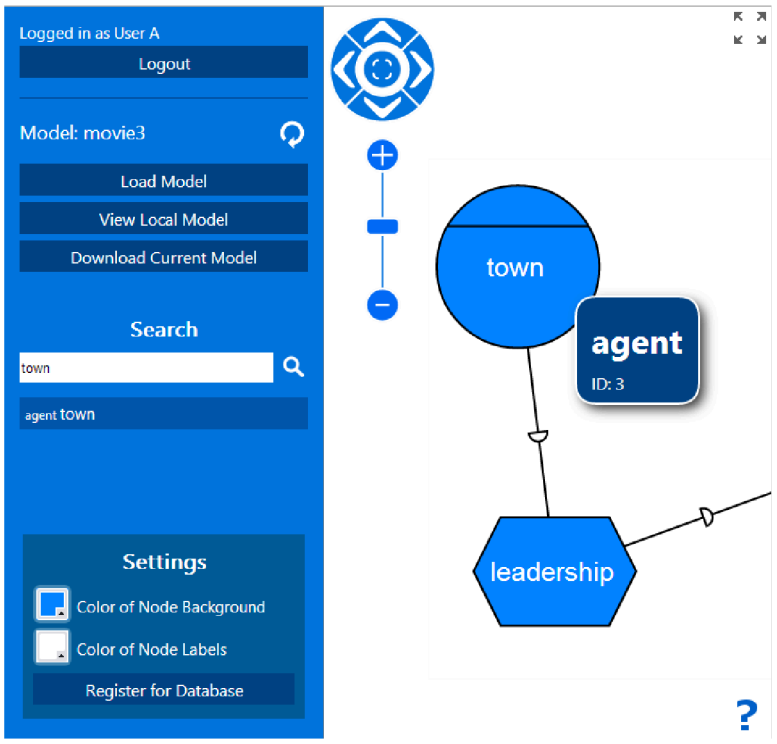


Figure 5.2: Web interface to view *i** models

models based on collected and analyzed data. At the end of the chapter I show models that are created automatically using i*-REST services.

5.3 Refinement

The phase of refinement is the phase when community stakeholders change communities in such a way that the communities function according to stakeholders' vision. Refinements that are executed by different types of stakeholders can support communities differently.

5.3.1 Simulating Learning Community Models

Effects of refinements can be predicted using simulation of models. Introduced in Section 3.2 agent-based modeling and simulation approach for the *refinement* phase is suitable for simulating social media (Ang and Zaphiris, 2009) and learning communities (Zhang and Tanniru, 2005). Similar to (Ang and Zaphiris, 2009) we focus on formation of communities in our simulation to understand when the communities will shrink or disappear and when they will grow.

*i** modeling is an agent-based approach and the existing environment for *i** models' simulation (Gans et al., 2004) is written in a situation calculus logic language *ConGoloc* that handles actions of agents but will have difficulties working with networks organized due to relations of agents with each other. Therefore, we appeal to a simulation environment, the Recursive Porous Agent Simulation Toolkit (Repast)[13].

Before we can perform simulations, *i** models need to be mapped to Repast Java-based agents. For that purpose we need to specify a formal model from Section 3.2 and define other models.

Environment Model Agents do not interact with each other directly. In learning forums they interact asynchronously through *threads* that include a set of messages. A thread has a relevant heading or a title that determines a thread's topic. Answering the thread users enter its environment.

Agent Model Agents, who represent community users, are responsible for creation and answering threads. Depending on network measures community users are classified onto *usual users*, *answering persons*, *inactive*, *questioners* and *conversationalists* (explained later in Section 5.6.2). We predefined probabilities for their actions according to user patterns.

Network Model People tend to follow socializing habits in any medium. For instance, they prefer to contact people they have already communicated with (frequently) in the past (Tsvetovat and Carley, 2004). Schnegg (2006) suggested that user communication in social networks can be resembled by a combination of *reciprocity* preference and *preferential attachment*. The former principle is based on connecting to known items or responsing to an existing connection, e.g., forum users respond to users they have already communicated with. *Preferential attachment* leads to a creation of scale-free networks (Barabási and Albert, 1999) with a power law degree distribution where rich nodes (many connections) get richer (get more connections). Such forum users will post or receive questions more often than others.

5.3.1.1 Learning Forum Community Model

The initial formal model is described in Section 3.2. We apply this definition to a learning forum community.

Every user is represented by one agent in our simulation. Referring to structural measures described in Section 3.4.1, each agent in our network has four attributes:

$$X = \{\ \textit{degree, betweenness, closeness, clustering coefficient}\ \} \tag{5.1}$$

Possible user actions Act are creating and answering threads:

$$Act = \{\ \textit{create thread, answer thread, answer community thread}\ \} \tag{5.2}$$

[13] An open-source, agent-based modeling and simulation toolkit `http://repast.sourceforge.net/`, Last access on 20.02.2015

Here *answering community thread* denotes that an agent who answers in a thread belongs to a community that consists of others agents that have participated in the thread.

Strategy functions in our model can be defined formally as:

$$S = \{\ \textit{Reciprocity, Preferential Attachment}\ \} \tag{5.3}$$

In the beginning of the simulation we operate with a graph of agents

$$G = \{A, E, Threads\} \tag{5.4}$$

A is a set of agents, E is a set of connections between agents $E \subseteq A \times A$ and $Threads$ is a set of artifacts created by agents, $Threads \subset Artifacts$. Given a graph G_t at time point t, a simulation aims to find successor states of the network G_t in the time point $t + 1$.

$$G_t \rightarrow G_{t+1} \tag{5.5}$$

Algorithm 4: The pseudocode of the multi-agent simulation

input : A set of agents and their attributes in a network G_t, number of steps n, initial Probabilities p, thresholdOfDecay d, thresholdForEdgeRemoval e

output: A set of agents, their attributes, their relations and threads in the network in the time interval G_{t+n}

```
// from the beginning agents get probabilities
// assigned for acting
agentProbabilities = initializeAgentProbabilities(p);
for time = 1..n do
    for each Thread do
        // since threads lose their popularity, we have
        // to update their attractiveness score
        threadProbabilities = updateThreadProbabilities(threads,d);
    edges = updateNetworkConnections(threads,e); for each agent do
        for each thread do
            mPA = findTheMostProbableAction(agent);
            mPABest = mPABest ¿ mPa : mPABest : mPA;
    step(MPABest);
```

Algorithm 4 describes briefly the main steps of the simulation. First of all, the function $initializeAgentProbabilities(p)$ uses the preliminary values p given for activity probabilities of agents that depend on agent roles or patterns (*usual user*, *answering person*, etc.). After that, if there exists any thread in the network we update probabilities of threads with the function $updateThreadProbabilities(threads, d)$ since

threads become less attractive for users to participate. The attractiveness is defined with the help of exponential function and the threshold d. After all probabilities of threads are recalculated, we need to detect if edges in a network G_t need to be updated ($updateNetworkConnections(threads, e)$). The edges are removed in the case they are connectind agents that participated in an old thread which decay coefficient is the same to e.

Then $findTheMostProbableAction(agent)$ finds the probability for an act of an agent: to create a thread or to answer a thread. After that the agent with the highest probability is defined in $mPABest$ and it is selected to be perfomed in the network by $step(mPABest)$.

In case of reciprocity

The function $findTheMostProbableAction(agent)$ iterates over all agents $A_1, ...A_n$ and over all existing threads $thread_1, ...thread_m$ and chooses an action of an agent based on probabilities. Such actions cause connections between agents and other agents that have performed actions connected with the threads. The connection $E^k_{i,j}(t)$ between A_i and A_j in the time point t, connected with a thread k is calculated as following:

$$E^k_{i,j}(t) = \begin{cases} 0 & \text{if } d_k \leq 0.05 \\ 1 & \text{if } \varphi_{k,reciprocity}(A_i, A_j) \text{ is the highest from } \forall A_i, A_j \in A \\ 1 & \text{if } E^k_{i,j}(t) == 1 \end{cases}$$

The connection between agents A_i and A_j according to thread k is possible if the probability of the connection between A_i and A_j is the highest between all other possible connections of agents A. The function $\varphi_{k,reciprocity}(A_i, A_j)$ estimates such a probability. Alternatively, if the connection $E^k(i, j)$ exists and the decay coefficient d_k is higher than the threshold 0.05 than the connection will sustain, otherwise the connection disappears.

$\varphi_{k,reciprocity}(A_i, A_j)$ estimates the probability of the connection $E^k(i, j)$ as follows:

$$\varphi_{k,reciprocity}(A_i, A_j) = \begin{cases} 1 \times d_k & : \text{if } A_i \in C_x, A_j \in C_x \text{ and } \gamma_k(A_j) = 1 \\ 0.01 \times d_k & : \text{if } A_j \neq initiator(k) \\ 0 & : \text{if } A_j = initiator(k) \end{cases}$$

The probability depends on a coefficient d that specifies the attractiveness of the thread k based on the age of the thread:

$$d_k = e^{-\sigma\tau}$$

where σ is a parameter defining the speed of aging and τ is the thread age.

Furthermore, the probability of the connection depends on the fact that the agent A_j participated in the last thread. It is calculated with the help of function γ_k:

$$\gamma_k(A_j) = \begin{cases} 1 & : \text{if } A_j \text{ appears in a last thread} \\ 0 & : \text{ else} \end{cases}$$

Finally, the agent A_j receives a low probability for the connection $E^k(i,j)$ if the agent is the initiator of the thread k. The function $\varphi_{reciprocity}$ calculates probabilities of connections between both agents considering the $reciprocity$ strategy while the decay coefficient d allows to simulate real situations in forums when old threads are no more attractive and the γ function filters agents and leaves only active ones.

In case of preferential attachment

The function $findTheMostProbableAction(agent)$ iterates over all agents $A_1, ...A_n$ and over all existing threads $thread_1, ...thread_m$ and estimates the probability of an action of an agent in a thread k. In contrast to the *reciprocity* strategy, the *preferential attachment* strategy emphasizes the role of an agent degree (a number of connections of the agent in the network) for estimating the probability of an agent action and a connection with others. For example, the connection between A_i and A_j is defined as following:

$$E^k_{i,j}(t) = \begin{cases} 0 & : \text{ if } d^k \leq 0.05 \\ 1 & : \text{ if } \varphi_{k,PA}(A_j) \text{ is the highest from } \forall A_j \in A \\ 1 & : \text{ if } E^k(i,j) == 1 \end{cases}$$

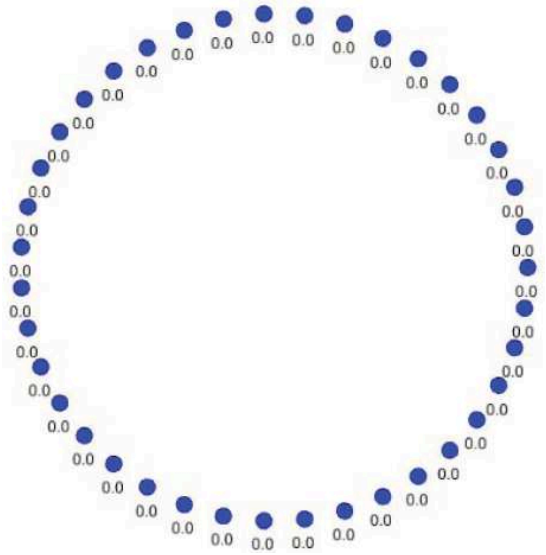
Similarly to the previous strategy, the connection between agents A_i and A_j according to thread k is possible if the probability of the connection between A_i and A_j is the highest between all other possible connections of agents A. The function $\varphi_{k,PA}(A_j)$ estimates such a probability. Alternatively, if the connection $E^k(i,j)$ exists and the decay coefficient d_k is higher than the threshold 0.05 than the connection will sustain, otherwise the connection disappears. The $\varphi_{k,PA}$ functions defines the probability of the action from the A_j agent.

$$\varphi_{k,PA}(A_j) = \begin{cases} (degree(A_j) + 0.01) \times d_k & : \text{ if } \gamma(A_j) = 1 \text{ and } A_j \neq initiator(k) \\ 0 \end{cases}$$

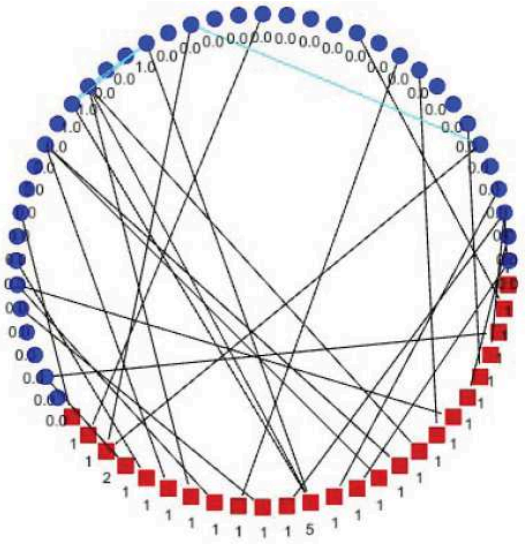
All these preparations are required to realize a simulation using *i** community models. An example of a community simulation is presented in the following section. Community stakeholders can use these results to check relevance of changes they want to apply to their communities.

5.3.1.2 Model Simulation

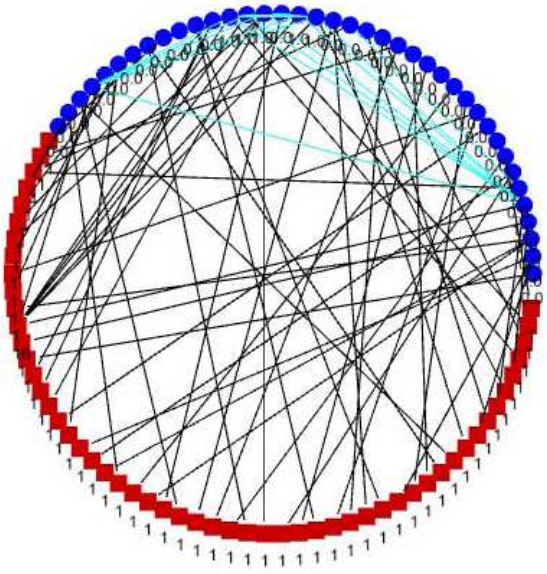
We execute simulation runs of a real forum community that is presented by an *i** model. Users (blue circles) create threads (red squares) and establish relations between one another (blue links) in Figure 5.3 (a) to (d). In the beginning (Figure 5.3 (a)) no threads exist. The number of users is defined according to the number of users in the model. Other information such as user patterns from the model influence probabilities of user actions. On the following pictures we can observe changes in the community after 10 (Figure 5.3 (b)), 20 (Figure 5.3 (c)) and 30 (Figure 5.3 (d)) days, where some of users receive more attention than others.



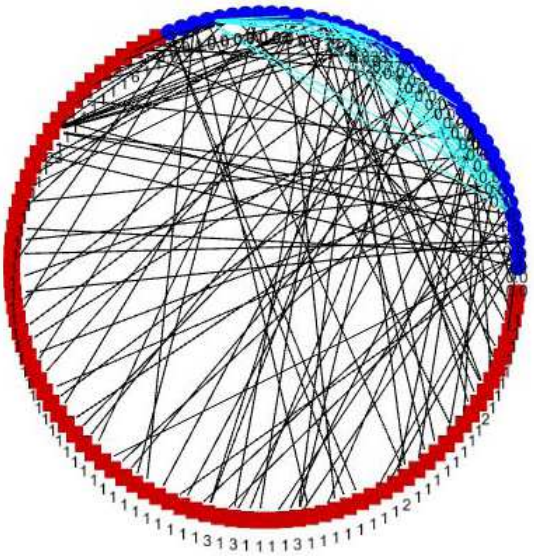
(a) Day 0. Users are denoted as blue circles



(b) Day 10. The communication between users is realized by threads presented as red squares



(c) Day 20.



(d) Day 30.

Figure 5.3: The example of a simulation execution

5.4 Monitoring

The data of two forums, the URCH and the StDocNet are collected with the Forum Watcher (Section 3.3.4.1). It includes posts for the time period of 10 years for the URCH and 13 years for the StDocNet.

Table 5.2 shows the difference between both communities: the URCH community

	URCH	**StDocNet**
Number of posts	429K	208K
Number of users	21K	25K
Number of threads	67K	8K
Users with $>$ 50 posts	2K	1K
Threads with one message	$>$ 10%	$>$ 13%
The longest depth of a thread	318	6K
The average depth of a thread	6	25

Table 5.2: Statistics of crawled data from examined forums

has more posts and threads than the StDocNet community. Nevertheless, StDocNet users participate longer time in a thread on average than URCH users so that many threads get a high depth, e.g. 6K posts were posted in the longest StDocNet thread versus 318 posts in the longest URCH thread. The URCH forum includes a number of threads (more than 10%) that can be called *single-user* threads where only one person posts. The StDocNet forum includes even more such *single-user* threads (more than 13%). The average number of posts in URCH threads indicates the question-answer model of the forum community. Such behavior is normal for communities that have relatively short-term tasks. It indicates the importance of core users that stay longer within the community and have experience with community topics. The average length of StDocNet threads is longer, i.e., 25 posts. StDocNet forum communities instead aim to support life long learning of medical students starting from pre-college period, following university, practical and working periods.

5.5 Analysis

For the analysis of community data we create the TargETLy service that is based on the Light Application Server (Klamma et al., 2006a), the GPU-based library for efficient community detection and evolution and the Matlab-based script for detecting user patterns based on outcomes for the TargETLy service.

The TargETLy service includes several modules that perform community detection and evolution, text mining, and named entity recognition. In the following I present results of the TargETLy service and GPU-based library for community detection and evolution. After that I explain outcomes of intent analysis, emotional analysis and named entity recognition used for clustering and modeling. In the end of the section the procedure of finding patterns of learners is described.

5.5.1 Community Detection and Evolution

We divide the time of our observation into time intervals. Usually a time interval can be chosen according to an event such as an exam. In the case of the URCH, users

posted exam dates and results in special threads. In the StDocNet forums, exams have fixed dates. Extracting these dates programmatically is possible, but time-consuming since user posts about exams' dates include a lot of exceptions that make the extraction complex. Therefore, in this experiment we choose a static time interval of 5 days

	URCH	**StDocNet**
N. of all communities	6949	18069
N. of unmapped communities	475	1452
N. of mapped communities	6474	16617

Table 5.3: Community statistics

length. The shift between time intervals is only 2 days, e.g., '01.01.2010 - 06.01.2010' and '03.01.2010 - 08.01.2010' are the first two time intervals. Using these intervals we create network snapshots and easily retrieve an impressive number of communities that exists in more than one snapshot, though some of them just exist in a time interval that appear in both snapshots, e.g., for the early mentioned time intervals a time period '03.01.2010 - 06.01.2010' appear in both snapshots. The rationale behind such a selection of time intervals or windows is in evolution of communities. Communities can stay alive for more than 5 days. Using sliding windows we can follow community evolution easily, for example, in the first time interval a community emerges while in the second grows. Because time intervals are overlapping it is possible to find communities that have a pause in activities, e.g., communities that are active in the first time interval, inactive in the second time interval and active in the third time interval.

Number of snapshots	**URCH**	**StDocNet**	Number of days
1	7%	8%	2
2	73,5%	67,3%	3-7
3	17,9%	23%	5-9
4	1,5%	1,5%	7-11
5	0,3%	0,1%	9-13

Table 5.4: Percentage of forum communities that are stretched over 1-5 snapshots

The results of mapped communities are presented in Table 5.3. Under mapped I understand communities that appear in at least two snapshots and their sets of users correlate with each other (as explained in Section 3.4.3). Other communities are unmapped.

Table 5.4 further differentiates between mapped communities according to a number of snapshots they appear in. I hypothesize that communities appearing in x snapshots have x phases, e.g., communities appearing in 4 snapshots have 4 phases. In the following I compare community characteristics depending on community phases. If characteristics of communities or community members fit an exponential curve, these are relevant for further investigation.

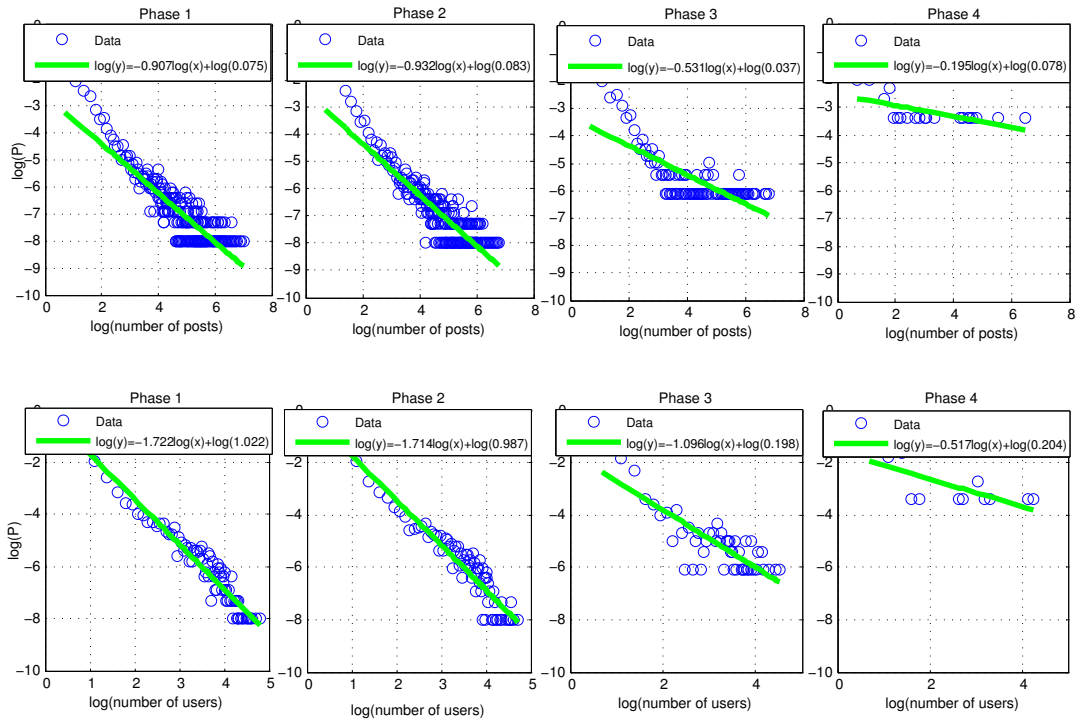


Figure 5.4: Distribution of number of posts (1st row) and number of users (2nd row) in communities that stretched up to 4 snapshots

First of all, I compare the distribution of the number of posts and users for different community phases (Figure 5.4). Both distributions have a tendency to become power-law distributions and fit the exponential curve. Such characteristics can be used to define different phases of community evolution. In contrast, adjacent nodes distribution[14] does not fit an exponential curve.

After that, I check the distribution of sentiment rate, cognition rate, number of intents, connectiveness, and betweenness of users in mapped and unmapped communities. Only number of intents and betweenness distributions fits an exponential curve but the distributions are far away from fitting the power law. In Figure 5.5 I present distributions of these characteristics for three phases of communities. They can be as well considered as potential arguments for investigations of community evolution.

[14]Here adjacent nodes are all neighbors of community members that are not in the community

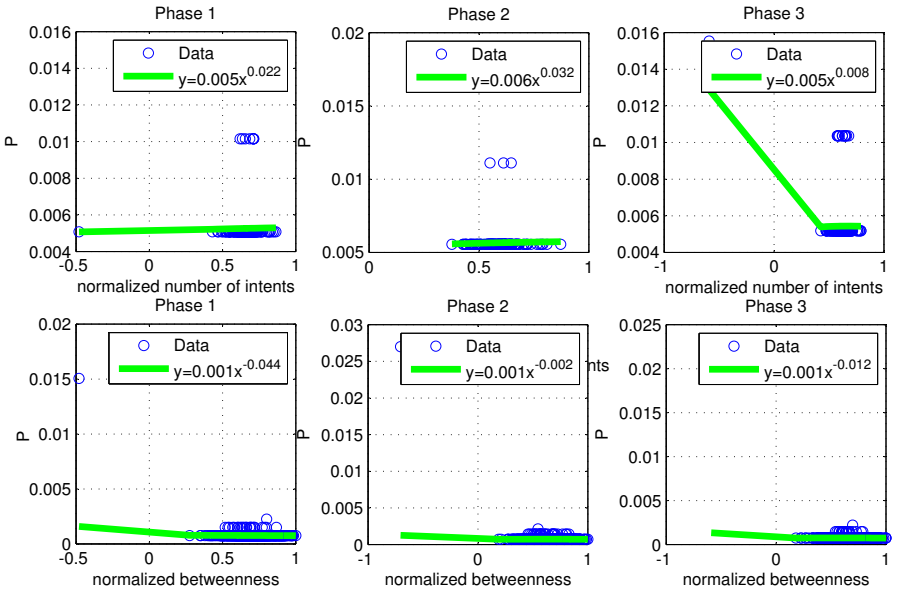


Figure 5.5: Intent phrases and betweenness distribution of users in communities that are stretched over 3 snapshots

5.5.1.1 Community Detection and Evolution Calculations in GPU

Results of the application of the GPU library on the URCH and StDocNet datasets are different in running time and in results. The library implements the accurate propinquity algorithm for community detection (Zhang et al., 2009) and complex but promising the event algorithm (Asur et al., 2009).

Name of the dataset		
	URCH (30 days)	StDocNet (30 days)
Starting timepoint	01.09.2008	01.09.2009
Final timepoint	01.10.2008	01.10.2009
Number of snapshots	1	1
Number of nodes	857	263
Number of edges	9110	1188
Propinquity threshold 4, 10		
GPU Running time	30 min	1.5 s
CPU Running time	2 h	4.0 s

Table 5.5: The running time of the community detection algorithm on a part of URCH and StDoctor datasets

In Table 5.5 we compare the results of the community detection algorithm implemented in CPU and GPU where both deploy a multithreading architecture. In all cases

the GPU implementation is quicker than the CPU one, though further experiments show that for networks with several hundreds nodes and edges the CPU implementation is quicker. The comparison of the algorithm running time executed on CPU and GPU on the datasets with more than 10K edges is presented in Table 5.6.

Name of the dataset		
	URCH	STDocNet
Number of snapshots	378	685
Number of edges	294.421	477.968
Propinquity threshold 4, 10		
GPU Running time	30 min	22 min
CPU Running time	> 4 h	3 min

Table 5.6: The comparison of the running time for the community detection algorithm

Table 5.6 includes two datasets with different number of snapshots, nodes and edges. Although the StDocNet sample has a higher number of edges and snapshots, the GPU and CPU running time is much quicker than for the URCH sample since URCH includes many small communities.

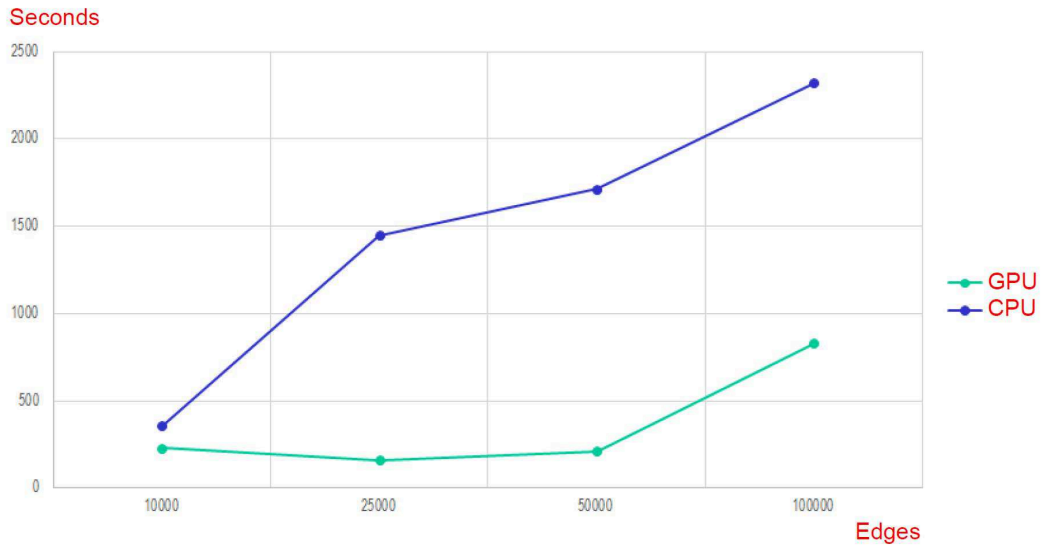


Figure 5.6: The running time for the propinquity algorithm on CPU versus GPU on the roadNet-CA dataset

In the following we test the algorithm with roadNet-CA dataset[15] (Figure 5.6). As

[15]The road network from the Standford large network dataset collection `http://snap.stanford.edu/data/`, Last access on 11.02.2015

it has no relevant time points we divide it on snapshots according to a number of edges. After that we analyze the snapshots using CPU and GPU. As a result we gain at least sixfold advantage in running the algorithm using GPU. Even though, both CPU and GPU have their limits and therefore at some point (for 100K edges) the running time for GPU increases drastically.

Snapshots		Community events					Node events			
Snapshot 1	Snapshot 2	Form	DIssolve	Continue	Merge	Split	Join	Leave	Appear	Dissappear
200	200	6	1	0	0	0	19	5	55	8
201	202	7	3	0	0	1	36	75	96	51
202	203	7	7	1	2	0	109	38	57	68
203	204	8	9	1	0	0	4	9	31	93
204	205	10	9	1	0	1	12	11	47	33
205	206	2	11	0	1	0	36	5	22	42
206	207	10	1	0	0	0	21	22	88	25
207	208	12	8	0	0	1	140	33	208	63
208	209	11	10	1	5	0	411	75	136	133
209	210	7	9	1	1	4	106	378	75	142

Table 5.7: The example of events in StDocNet snapshots

Furthermore, we define events of communities implementing the event algorithm from (Asur et al., 2009) where we differentiate between form, dissolve, continue, merge and split events for communities and join, leave, appear, disappear events for users. Similar to the propinquity algorithm we compare CPU and GPU running time and find that for our cases difference between both implementations is not so extreme as for the community detection algorithm. For example, for the StDocNet sample with 220 snapshots the CPU takes 53 seconds while the GPU only 21 seconds. In Table 5.7 results of the event detection of some snapshots are presented. It is noticeable that many communities exist only in one snapshot and just a few continue to live, merge or split. Furthermore, user events provide interesting information about traffic of new and leaving nodes that can be interested for understanding community and community media, in our case, forum success.

Outcomes from the community detection and evolution both give important information about communities in forum networks. Using data about community users we can start to model communities as we know their boundaries. Furthermore, we can investigate how and why communities sustain if we investigate communities that continue to exist and design their models. Moreover, models of communities that split, merge and dissolve help to find why such events are happening to the communities. Loss of users or flow of new users should be interesting for community stakeholders

to find reasons for such events.

5.5.2 Emotional Analysis

The emotional analysis was performed for the URCH forums only. We select relevant data for the training sets for both classifiers, sentiment and cognition, that include 150,000 sentences. After that further 1,700K sentences from the URCH forum are classified. Table 5.8 and 5.9 include some of the results.

Category	Sentence
Neutral	Solar systems would't have geological and climatic whereas planets would.
	This triangle will have a base of 8 and a height almost equal to zero.
	I think it is D.
Emotional	And this one as well, damn I just suck at prob and these type of questions.
	Screw ETS!!!
	Good luck in your studies, and good luck in your exams!
	Anyway, thank u for sharing with us such a nice essay.
	Overall I feel pretty satisfied and happy with my results.
	I'm sorry for your loss and for all of those who lost someone in this tragedy.

Table 5.8: Examples of the classification of the URCH post sentences according to their sentiments

5.5.3 Intent Analysis

We detect intents in texts of communities using the TargETLy service. We detect more than 132K intent phrases in the URCH forum. In average each fourth post includes an intent.

Investigating intents with most popular patterns $VB_1\_to\_VB_2$ (verb to verb) and WRB_to_VB (wh-adverb to verb), we pick 10% of all detected intents and find the most popular expressions of intents (Figure 5.7). The most popular keyverb is *know* that shows a learning character of a community.

Next we consider words that are following after the language patterns. Analyzing most popular of them in Figure 5.8 it becomes clear what are community purposes: "tests", "exams", "GRE", "TOEFL", "GMAT".

Since I operate with agent-based and goal-oriented modeling to model communities, mining of goals is pivotal. Then community stakeholders can get a better idea

Category	Sentence
Neutral	That was when things started getting interesting.
	I had very few sums where calculations were the only way to an answer.
	This form of energy is called wind energy.
CogMech	Please correct me in case I made any mistakes, wrong assumptions...
	How can you determine that the two areas are equal?
	I just can't understand these answers.
	I wonder if ETS goes through the Barrons list and creates questions with words that are not on the list.

Table 5.9: Examples of the classification of the URCH post sentences according to the usage of words showing cognition

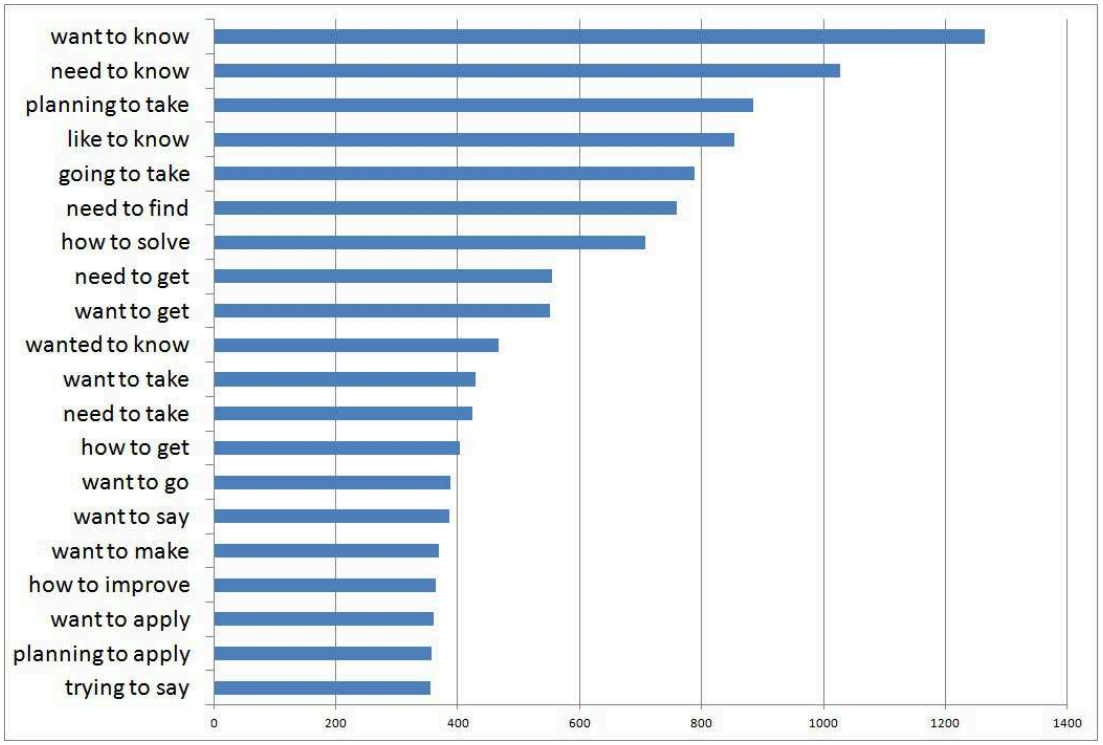


Figure 5.7: The 20 most occurring expressions of intents in URCH forums (Krenge et al., 2011)

about community goals observing intents of community users though using intent analysis we cover only a part of expressed goals thus there is a space for further investigations of learning goals.

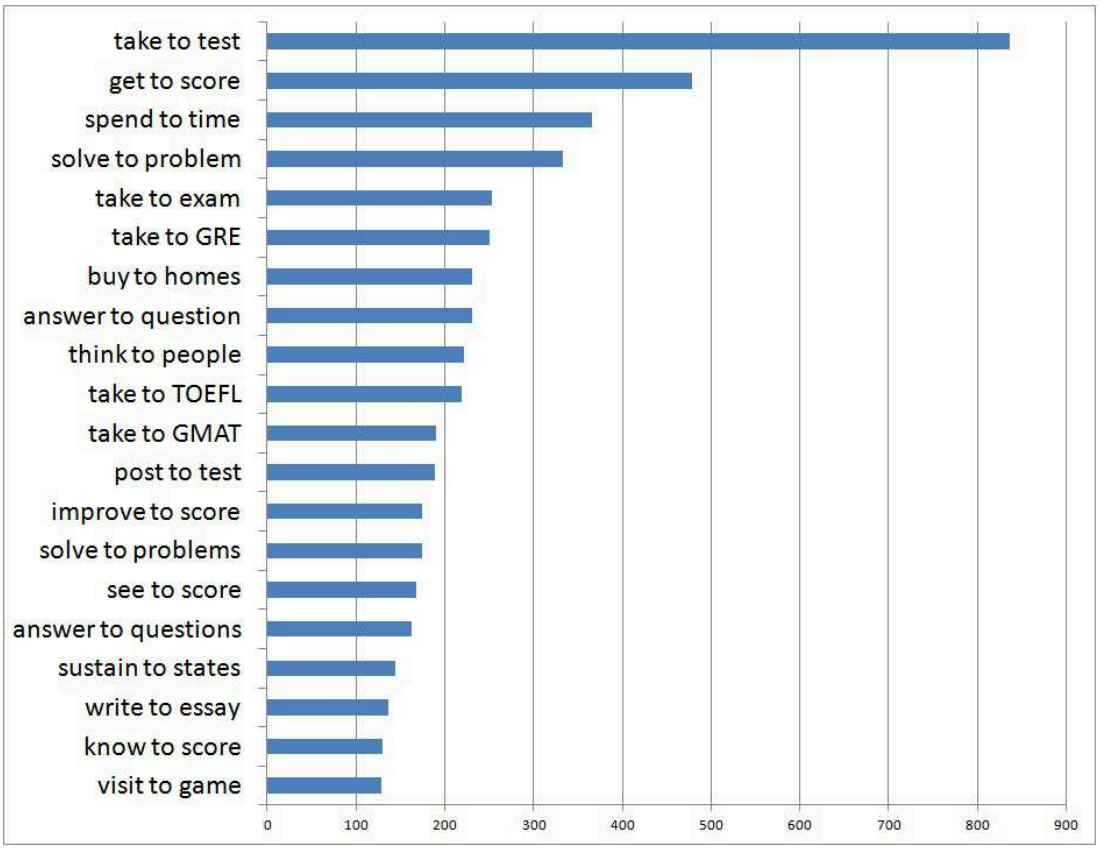


Figure 5.8: The 20 most occurring intent expressions and following nouns in URCH forums (Krenge et al., 2011)

5.5.4 Learning Concepts and Topics

Texts of communities provide a ground for extracting semantic data from learning resources using Named Entity Recognition (NER). We choose Open Calais[16] as it concentrates on extracting relevant concepts from social media. Open Calais detects NER, topics, categories of texts and tags that can be used for texts. It outputs machine-readable RDF (Resource Description Framework) files with all the information about requested texts that are stored in the RDF repository, Sesame[17]. Open Calais requires no labeled corpus since, most probably[18] it uses unsupervised learning algorithms. The ontology of Open Calais is connected with DBpedia, Wikipedia, Freebase,

[16]The home page of Open Calais `http://www.opencalais.com/`, Last access on 06.10.2014

[17]Java framework for processing and handling RDF data `http://rdf4j.org/`, Last access on 20.02.2015

[18]it is not clear how Open Calais works as it is a commercial system with the open access for academy

Reuters.com, GeoNames, Shopping.com, IMDB, and LinkedMDB[19]. The Open Calais web service recognizes entities like *Person*, *Position*, *Company*, *Organization*, *Country*, *ProvinceOrState*, *City*, *CareerType* and many others. Moreover, it classifies and estimates a probability of such a recognition.

Category	Topics covered	Number of Posts
Business_Finance	financial achievements, prices and markets	2.1K
Education	knowledge acquisition	2.5K
Entertainment_Culture	music, celebrities, Internet culture	1.7K
Environment	natural disasters, protection of the Earth	1.9K
Health_Medical_Pharma	hospitals and healthcare, medical research	2K
Hospitality_Recreation	travel, leisure, relaxation activities	2K
Human Interest	general interest for humans	4.8K
Law_Crime	enforcement of rules of behavior in society, law firms, legal practice and lawsuits	1.5K
Politics	policies and actions of politicians, elections	1K
Religion_Belief	theology, philosophy, ethics and spirituality	1K
Social Issues	behavior of humans affecting the quality of life	1.2K
Technology_Internet	technological innovations and companies, products	2.1K

Table 5.10: Categories of documents found in language learning communities

Analyzing the URCH forum data results in ≈4M triples, ≈2M statements and 3.2M entries. In Table 5.10 I present topics and categories of URCH posts. Post topics are very broad as users share essays that are not dedicated to one topic. Therefore, some texts have no connections with the *Education* category but are classified as other categories such as *Politics* or *Environment*. These texts are noise for intent analysis and can be avoided during the content analysis to achieve better results.

Named entities appeared in URCH forums are represented in Table 5.11. We can retrieve particular entities that appear in posts by querying for objects in RDF files with different types (Appendix A).

[19]OpenCalais Linked Data - Entities, `http://www.opencalais.com/documentation/linked-data-entities`, Last access on 16.05.2014

Entities	Number of Entities
City	22K
Company	20K
Continent	11K
Country	462K
Currency	29K
EmailAddress	1.9K
Facility	3K
Holiday	1.9K
IndustryTerm	71K
MedicalCondition	9K
NaturalFeature	2K
Organization	23K
Person	16K
Position	230K
Product	1.7K
ProgrammingLanguage	15K
ProvinceOrState	57K
PublishedMedium	2.6K
Region	1K
Technology	771K

Table 5.11: Entities in language learning communities of URCH

Results of analysis described in this subsection is useful to give an idea about topics and concepts of communities and their users. Such an information can enrich community models with relevant data about community shared repertoire (Wenger, 1998).

5.6 Results

In the following section I describe applications of analysis techniques from the previous chapter. As a result we explore users in communities and provide information about user patterns and learning phases to community stakeholders for getting a clearer picture about communities' states.

5.6.1 Phases of Learning

Learners in forums are self-regulated learners (Zimmerman, 1990). The Psycho-Pedagogical Integration Model (PPIM) of Nussbaumer et al. (2011) introduces a loop with four phases of learning showing the maturing of a learner and her progress. Learners can

self-reflect (Section 2.1.2) basing on estimations of their progress. They can initiate activities in their communities to refine their statuses.

In the first phase of the PPIM learners create or update profiles of social media, include their preferences, knowledge and purpose if possible. This phase helps learners to plan their learning. We hypothesize (Figure 5.9) that in the first phase of the PPIM a learner has a low activity as she takes care about her profile or plans her learning but she is not very active in a medium.

In the second and third phases peers learn. In the second phase the learner increases *help-seeking* activity and her texts include more sentiments. Her *cognition* rate is not significant as the learner asks simple and short questions. We estimate the *help-seeking* and *activity* according to thread creation and thread participation: in the beginning of learning process learners ask a lot of questions, i.e. start threads.

In the third phase the learner is more concerned about discussions and participates actively in them. Her questions and answers indicate understanding of learning topics thus comparing to other phases *sentiment* and *cognition* scores increase. The *help-seeking* decreases while the *activity* is increasing, i.e. participating in threads of others.

In the last, fourth phase, the learner can reflect according to her results and learning process. In the fourth phase she estimates her learning process and outcomes by expressing her feelings about it (*sentiment* is high) while other indicators decrease.

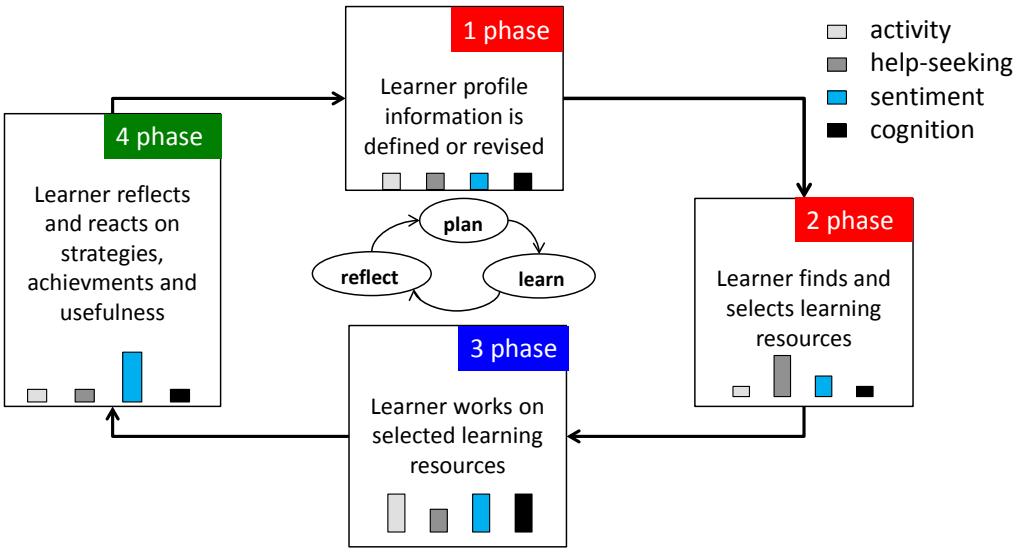


Figure 5.9: Phases of learning with their indicators (Krenge et al., 2011)

5.6.1.1 Realization

We observe small communities with 6-10 users where at least 75% of users from the communities have to appear in community threads. In Figure 5.10 we depict footprints of 4 communities that exist for 7-8 weeks while the average life span of found communities is 7 weeks. Figure 5.10 depicts the most active community users.

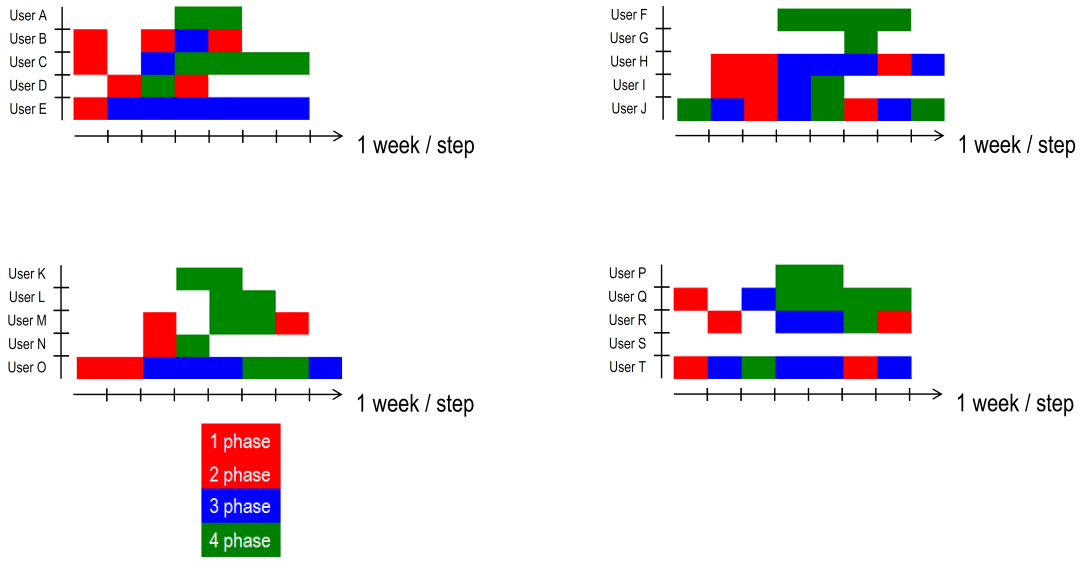


Figure 5.10: Example of footprints of 4 communities

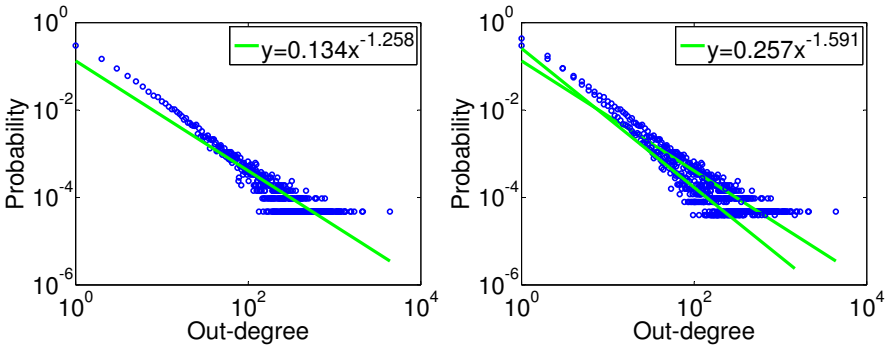
Depicted communities include different types of users: supporters appear mostly in phase 4 and recommend or give advices to others. Other users change their learning behaviors from first-second phase to third and fourth. In some communities consequences of phases is different to our hypothesis in Figure 5.9. Users of 38.6% of found communities follow the PPIM while users of 11.8% communities show exactly contrary behavior, i.e. phase 4 comes first, following by phase 3 and 2 and 1.

Results of such an analysis can be used by learners themselves to refine their learning processes. Community stakeholders can profit from such analysis since they have a view on community learners' phases and indicators that can help them to estimate communities and decide to support communities by attracting experts or giving interesting for communities tasks.

5.6.2 User Patterns

First of all, we check how *scale-free* are the out-degree distributions of both forums in Figure 5.11. A network is scale-free if its degree distribution follows a power law. The distribution shows that the network includes just a few nodes with high degrees, i.e. high number of incoming and outgoing edges and many nodes with other degrees. Mathematically the power-law degree distribution is explained as following $P(k) \approx$

k^{γ}, where k states for a degree frequency and γ is the exponent of the power law distribution. γ should be between 2 and 3 to represent a scale-free network.



(a) URCH degree distribution (b) STDOCTOR degree distribution

Figure 5.11: Out-degree distributions

Both forum networks are still not *scale-free* networks. Both networks have $\gamma < 2$ which means that the average degrees of the networks diverge. In other words, differences between frequencies of different degree distributions is not so high than for scale-free networks. Although high degree nodes exist, their degree distributions are not extremely higher comparing to others as well more nodes exist with degree values that differ by a smaller variance than it is in a scale-free network. It means that differences between users in forum networks using degrees can not be so clearly defined as in scale-free networks and therefore other features for detecting user patterns are required.

Patterns are initially defined as repeating situations (Alexander, 1978). In social media I define patterns of users as user states depending on their behavior. These characteristics are structural and semantic measures we collected in the Mediabase Cube. For defining patterns of users I apply k-means clustering algorithm (Han and Kamber, 2006) to find groups of users with similar measures. K-means uses unsupervised learning, therefore no labeled data is required. Using k-means clustering I expect to get clusters with similar amount of entities. Moreover, I use the silhouette function (Han and Kamber, 2006) to estimate the accuracy of clusters. Implementing k-means algorithm in Matlab we extend it by providing weights for given measures that define states of users. The measures include *connectiveness*, *betweenness*, *number of intents*, *sentiment* and *cognition* rates. We range weights between 0 and 5 for all 5 measures by firstly normalizing weights and then multiplying the measure values with the normalized weights' values. The weights help to classify measures into more and less influential.

For the clustering we use a small set of users (1262) that are members of mapped communities (introduced in Section 5.5.1). We perform clustering using a combination of 5 features and 5 possible weights for each feature. As it is already shown in results

of Section 5.5.1 cognition and sentiment scores are not relevant to classify different states of communities during their evolution. Furthermore, we find that these scores have no influence on building of user clusters as can be seen from Figure 5.13, where sentiment scores of different clusters do not differentiate from each other. Such a result is highly influenced by a procedure of sentiment/cognition scores' calculation as well as by items of investigation, in our case forum users' texts.

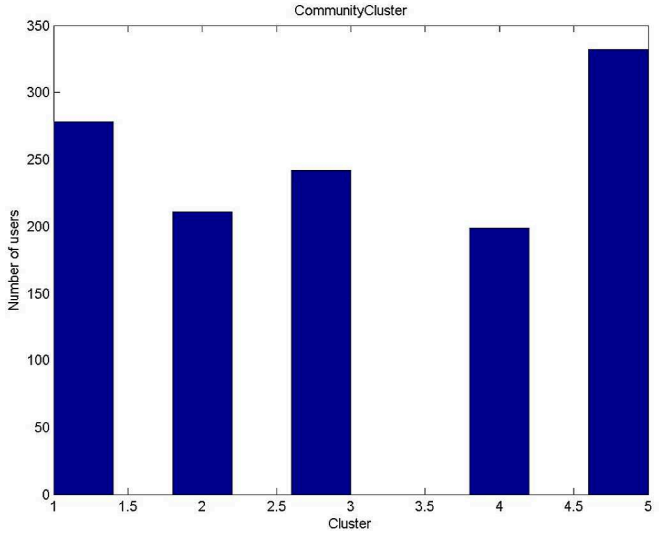


Figure 5.12: The distribution of users over 5 clusters. Only connectiveness (weight=5) and betweenness (weight=2) were considered for the calculation. The average silhouette value is 0.76523.

One of the best results of clustering is depicted in Figure 5.12 where only connectiveness and betweenness measures are used for clustering. Figure 5.14 illustrates median, 25th and 75th percentiles (top and bottom of boxes), and outlier values (pluses) of connectiveness of the defined clusters. Starting with connectiveness values I define patterns of users that belong to the clusters. I hypothesize that first and second clusters include users that are not central in a network, e.g. their connectiveness is low. The third cluster consists of *newbies* with very low closeness. More central users like *conversationalists* are in the fourth and fifth clusters.

Figure 5.15 depicts differences between betweenness rates of users in the clusters. The first and second clusters differ in betweenness: the first cluster includes *usual users* with low closeness and low betweenness; users in the second cluster are *questioners* as they have high betweenness and get replies from members of different connected groups but are on the periphery of the network (low closeness). Users in the fourth cluster with high betweenness and high closeness are *answering persons* as they are relatively central and contribute to different communities (betweenness). Users from the fifth cluster are *conversationalists* as their positions in the network are central but their betweenness is low. They prefer to answer community peers, e.g., the peers they

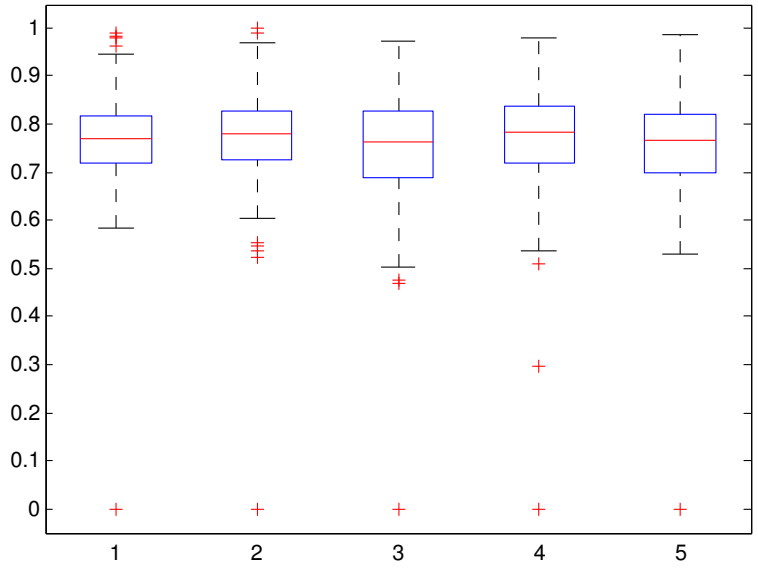


Figure 5.13: The statistics of sentiment values of our training set. On the x axis are clusters, on the y axis are values of sentiment scores.

have already communicated with. Such users can contribute to many threads but do not connect isolated groups or users to their communities.

Presented patterns are used in the following community models to specify community member roles. Later these roles have an impact on user behavior that is required for an appropriate simulation of models.

5.7 Learning Community *i** Models

According to the process of community model creation in Figure 3.2 either we start with classical models as described in Section 3.1.2 and allow stakeholders to think about community models, make refinements and check their efficiency using simulations. Alternatively the results from *monitoring* and *analysis* are used for the actual modeling. Automatic modeling of communities using i*-REST (Section 5.2.1) is conducted using the data and results are discussed in this section.

The following models represent communities detected in URCH learning forums where community users are parts of the *Community* actor. Figure 5.16 shows a community with 11 users. In the analysis I find only three users that *play answering persons* and one user that *plays conversationalist* in the community. Other dependencies show important actors of the forum community. The *Community post* in the *Forum*, the *Threads keep* a *Forum* alive, and the *Named Entities appear* in content of the *Threads* because the *Community talks about* or *use* them.

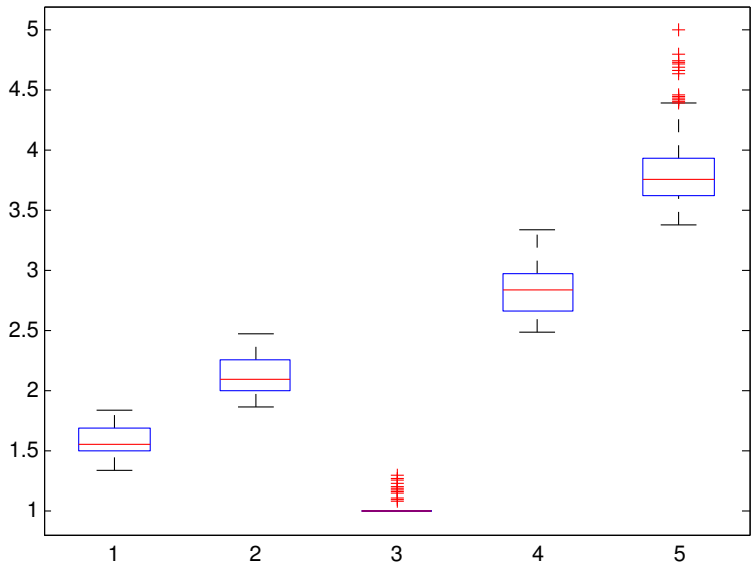


Figure 5.14: The statistics of connectiveness values for different clusters. On the x axis are clusters, on the y axis are values of connectiveness.

Figure 5.17 presents a part of a model of another community where one can view learning topics and concepts of the community (Section 5.5.4). For this community some geographical US items were extracted as well *diagnostic tests* that is a topic of tasks users are talking about in the community. Figure 5.18 includes another part of the model with 1) titles of *Threads* and 2) a set of intentional phrases for one of users. Both the titles and intentional phrases help community stakeholders to understand topics and goals of communities. The titles include topics devoted to reported scores in tests as well as test questions. So community topics are distributed as the titles are not following one topic direction. The user intents include only phrases that are located in sentences with special grammatical constructions (check Section 3.4.2.3). Some users express many intents while others just a few or none at all, combining all of these stakeholders can get an impression of learning community goals.

Stakeholders are interested not only to view a current state of a community but as well follow its evolution. In Figure 5.19 I present two models of a community. One model represents the state of the community between December, 1 and December, 10, 2004 while the other depicts the community between December, 08 and December, 17, 2004. The community models have some trivial statistical data like the number of posts, the number of users and the number of adjacent nodes that clarify how central the communities are. The number of posts and users decreases so the activity diminishes in the second period as well as the number of adjacent nodes is descreasing.

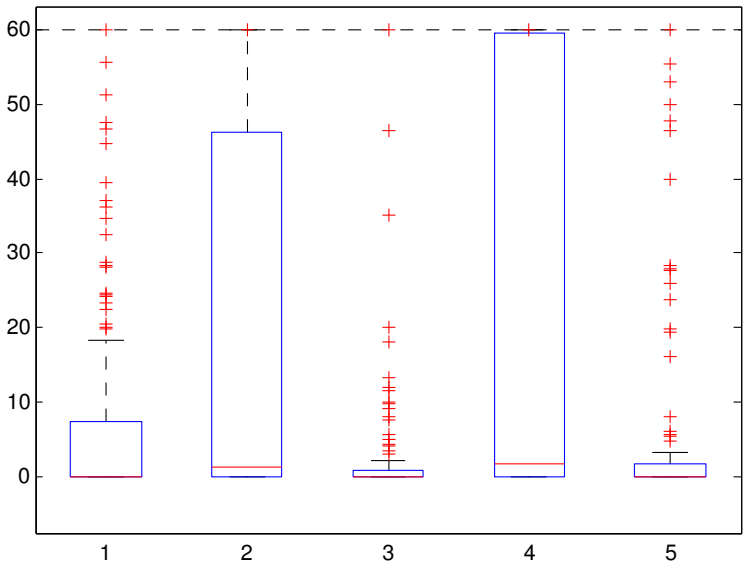


Figure 5.15: The statistics of betweenness values for different clusters. On the x axis are clusters, on the y axis are values of betweenness.

Word clouds for both communities include the words appearing in thread titles of the community. For both time intervals topics are devoted to math tasks. The phrases we discover during the intent analysis are on the bottom of Figure 5.19. All phrases define the desire of users to learn while some phrases can help to identify more precise information about goals by investigating sentences where learning phrases appear, e.g., how to answer (what?), need to learn (what?), (what?) take to solve (what?).

5.8 Evaluation

In the following, I present evaluation of sentiment measures, models and model simulation we have done for this case study. Structural measures are based on well-established properties of nodes in graphs while patterns are evaluated implicitly since they are used in models and model simulations.

5.8.1 Sentiment Measures

We evaluate the retrieved intent phrases, sentiments' and cognitions' scores in the URCH forum by surveying 18 active forum users. The URCH forums includes 21K users. Therefore at first glance 18 users is just a tiny amount of forum representatives. On the other hand, considering usual behavior of most URCH forums it becomes clear

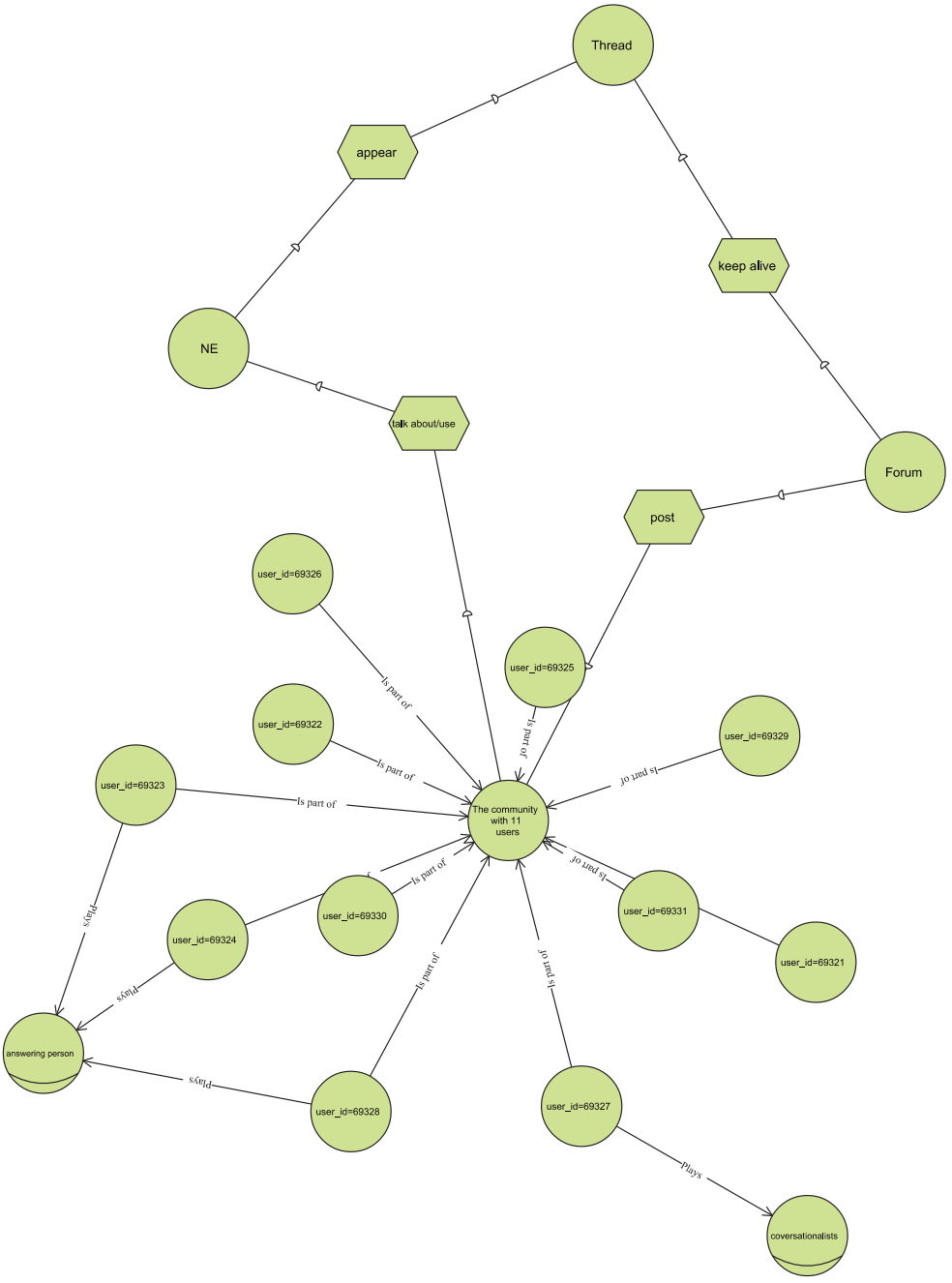


Figure 5.16: A model of the community with 11 users

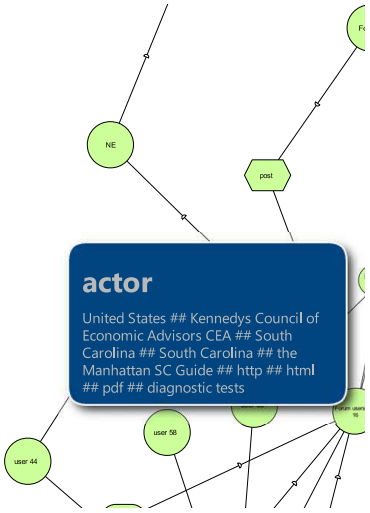
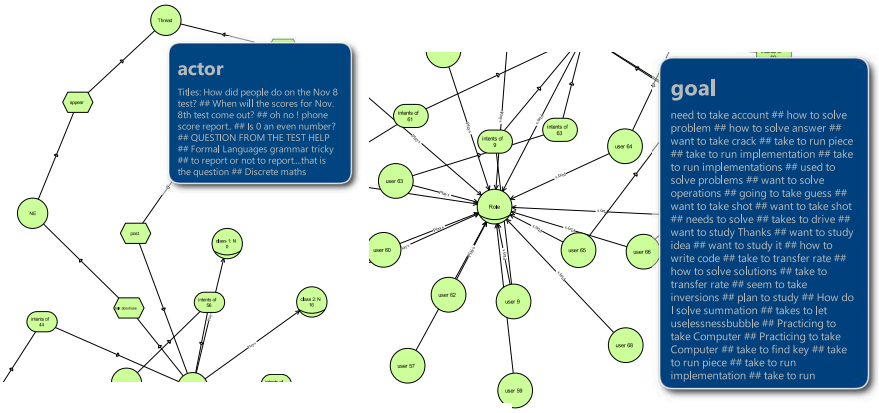


Figure 5.17: The named entities extracted in threads of the community



(a) The titles of community threads (b) Intents of one of users from a community

Figure 5.18: Details of community models

that most of them are leaving the forum as soon as their goals are achieved. Thus just a few of members belong to active users. We surveyed these users as they are interested to improve the forum.

We let the forum users review verbs of intent phrases we retrieved from the sentences they wrote. In the questionnaire we showed not only the phrases but the whole sentences where we found user-specific intents. The questionnaire shows that *know*, *solve*, *analyze*, and *work* are words that are relevant for expressions of intents. While other words like *find* are not reasonable for intent expressions in our context (Figure 5.20).

Also the users evaluated 2 posts themselves: one with/without sentiments and the

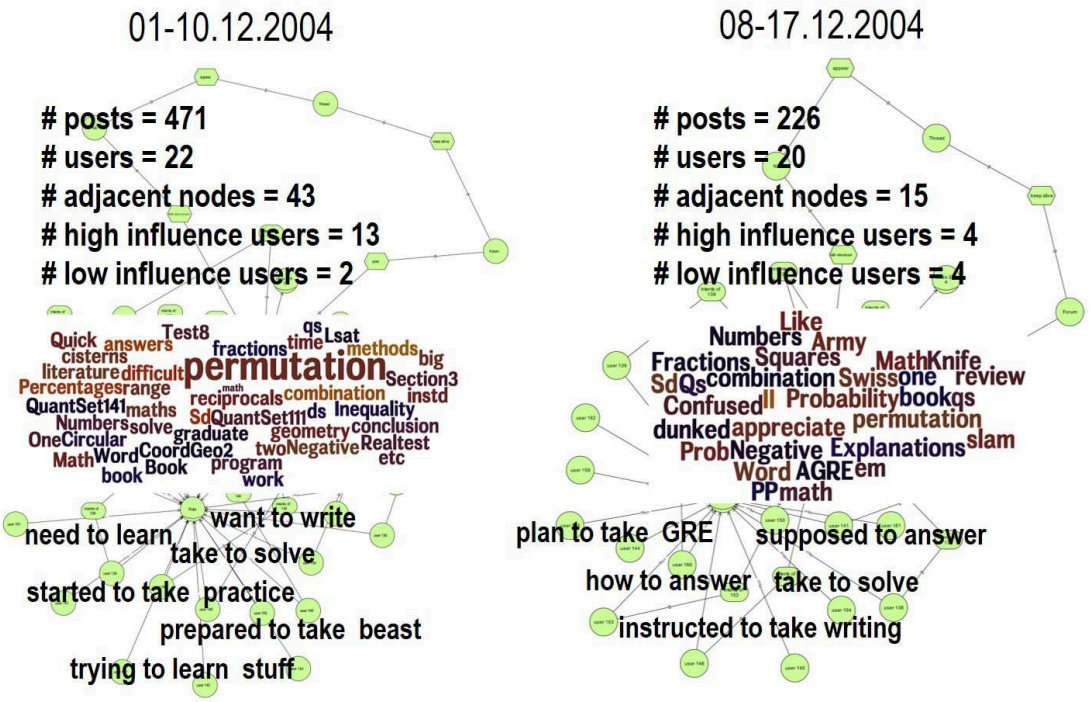


Figure 5.19: The example of 2 models of the evolving community

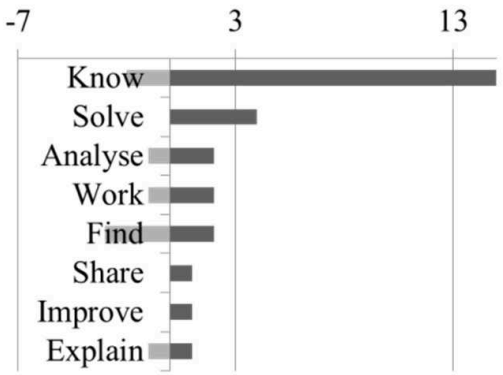


Figure 5.20: Significance of keyverbs in intents based on user-specific evaluation in URCH forums (Krenge et al., 2011)

other with/without cognition words. The results of sentiment/cognition rates are estimated using precision and recall and included in Table 5.12.

The users agreed that all posts we tagged as emotional are emotional from users' point of view. Our classifier recognizes short posts in many cases as emotional, while users find them neutral. The recall value 0.67 indicates that we should increase the value by more precise evaluation of sentences. Table 5.13 shows examples of sentences that users evaluated.

	Precision	Recall
sentiment	0.67	1
cognitive mechanisms	0.73	0.53

Table 5.12: Evaluation of sentiment and cognition rates by users

Classification	Sentence
true positive (emotional)	apoovra, Yes, I compare crimes rather than difficulty in both cases. oh... my weakest part Thanks.
false positive (wrongly defined as emotional)	*[...]* see I told you I am poor at acronyms! Thanks for clarifying though!
	That would require studying to optimize my score, which i am not going to do. The University of Florida Warringtom is allowing me to take the GRE *[...]* I am just wondering how I can estimate my score on the GMAT using my GRE score. I have found a site that estimates the SAT score and IQ, from the GRE but not the GMAT.

Table 5.13: Examples of evaluated by users posts

The evaluation of sentences defining activation of cognitive mechanisms shows a precision value of 0.73. While 7 from 18 posts were wrongly defined as neutral and, therefore, the recall is very low. Table 5.14 shows examples of sentences estimated by the classifier and the forum users if the sentences cause cognitive processes of the users or not.

The limited vocabulary for emotional analysis, the specific content domain, the language models and the small amount of evolution items can influence the results tremendously.

5.8.2 *i** Models

I performed the evaluation devoted to the process of community model creation within the seventh *i** workshop[20] in Thessaloniki (Dalpiaz and Horkoff, 2014). Twenty one experts in *i** modeling answered a survey that aims, among other things, to evaluate techniques that are used to support community needs. The results of acceptance of Social Network Analysis, Community Detection and Evolution, Goal Mining (Intent analysis) and Named Entity Recognition is depicted in Figure 5.21.

The experts find mentioned techniques in average *more or less relevant* for the *i** model generation. Some of participants choose *Goal Mining* as a relevant or more or less irrelevant technique, but in average the technique is found as more or less

[20] *i** workshop in 2014 `http://istar14.wordpress.com/`, Last access on 8.10.2014

Classification	Sentence
true positive (cognitive mechanisms are presented)	The forums are still working. Ill look into it a bit later.
false positive (cognitive mechanisms are wrongly defined)	Just a little FYI, the real test is a lot harder than the practice test. You will encounter new words you have not studied. The best way to prepare is practice and more practice. As for reading comprehension, I'm struggling with that too.
false negative (cognitive mechanisms exist but are not defined)	Alrite guys. What do you feel about the elimination reasons I have used. Do you think that's correct?.

Table 5.14: Examples of evaluated by users posts

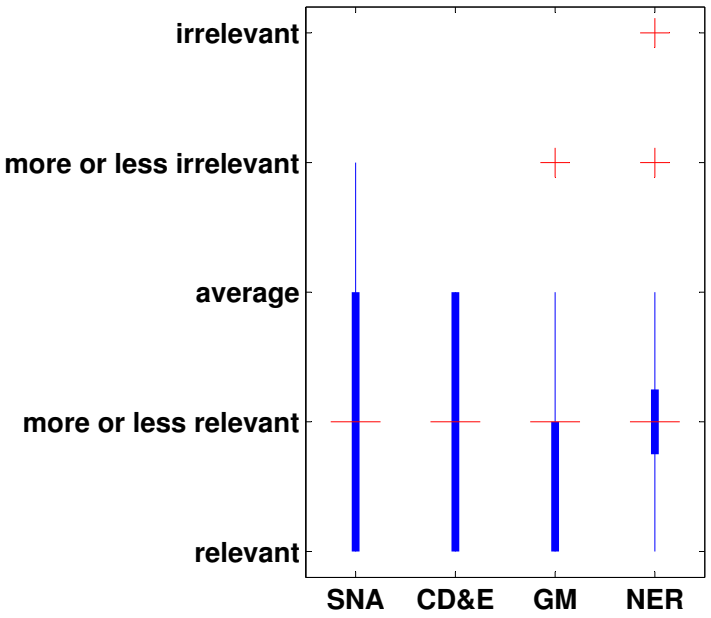
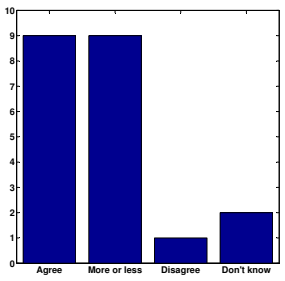
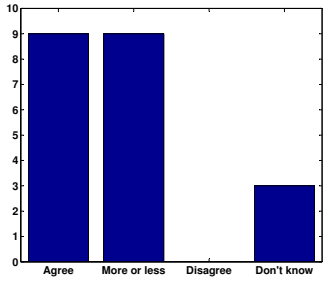


Figure 5.21: Answers to the question about relevance of techniques for *i** model generation. SNA stands for Social Network Analysis; CD&E for Community Detection and Evolution; GM for Goal Mining; NER for Named Entity Recognition. Median values are presented by red lines. Bottoms of boxes are the 25th percentiles while tops are 75th percentiles. Whiskers, who define other answers as a majority, are connected with tops or bottoms of boxes using lines. The outliers plotted as red pluses define unique values.

relevant. The rates for *Social Network Analysis* and *Community Detection & Evolution* are more distributed between relevant and average rates. Most participants agree that *Named Entity Recognition* technique is *more or less relevant*. The survey shows the acceptance of the techniques used in the thesis for community modeling.



(a) Answers to the question about appropriateness of *i** models to use for community stakeholders to understand communities



(b) Answers to the question about developer work facilitation by automatic creation of *i** models that developers can use for understanding community requirements and extending information systems based on the requirements

Figure 5.22: Results of answers on questions

Another question of the survey pursuits to estimate if *community stakeholders* - teachers, instructors, community managers of learning communities - *can follow changes in communities and recognize problems and conflicts just by following i* models*. Experts divide into those who agree with the statement and those who more or less agree with the statement (check Figure 5.22 (a)). Some experts commented that *i** models can be *abstract* and *not straightforward*. Therefore, *training is required* before community stakeholders can use models for analyzing their communities.

The last question I consider here investigates if *i** models help in developing (community) information systems. Most experts agree or more or less agree that *i* models facilitate work of developers in extending information systems*. Figure 5.22 (b) presents answers to the question.

5.8.3 Simulation Validation

Model validation shows if *the expression of the simulation in terms of outcomes is faithful to the relevant social phenomena* (David, 2009).

We compare our simulation results in terms of different factors such as user degree, betweenness, closeness and local clustering coefficient with factors of communities in

real life. We apply the Kolmogorov-Smirnov test (K-S test) that is a nonparametric test with help of which we can compare two probability distributions (Lin et al., 2010).

Performing simulations we consider user patterns to choose appropriate probabilities of acts. In Table 5.15 we define several maps with probabilities vectors for each user pattern. *Answer thread* means to answer an arbitrary existing thread while *answer community* defines a case when an initiator of an answer is a part of the community she answers. The presented probabilities are used for the $initializeAgentProbabilities()$ function in Algorithm 4.

Role	Behaviour	map1	map2	map3
usual user	answer thread	0.01	0.03	
	answer community	0.08	0.06	
	create thread	0.04		0.06
answering person	answer thread	0.01	0.09	
	answer community	0.12	0.04	
	create thread	0.06		0.1
questioner	answer thread	0.01	0.07	
	answer community	0.13	0.07	
	create thread	0.07		0.1
inactive	answer thread	0.01	0.21	
	answer community	0.2	0.0	
	create thread	0.04		0.15
conversationalist	answer thread	0.01	0.05	
	answer community	0.08	0.04	
	create thread	0.07		0.1

Table 5.15: Probabilities of activities considering user patterns

We achieve best results in similarity between simulated and real communities for communities starting from thirty nine members. Figure 5.23 shows the results of the K-S test. The value of the K-S statistic is on the y axes that shows the difference between factors of simulated and real communities. Therefore, lines that tend to have low y value represent factors that converge. Furthermore, we differentiate between different maps from Table 5.15 as well as between different strategies. We simulate considering that community members collaborate only according to reciprocity strategy (first row), a combination of reciprocity and preferential attachment where reciprocity is prevailing (second row) and a balanced usage (50/50) of reciprocity and preferential attachment.

In general, simulations under the reciprocity show the best convergence of factors, e.g., clustering coefficient (CC) values of a simulated community are close to CC values of a real community. Results of simulations are far away from being ideal and thus these have to be refined by defining better starting probabilities and calculation of these according to network strategies. We operate with communities detected using

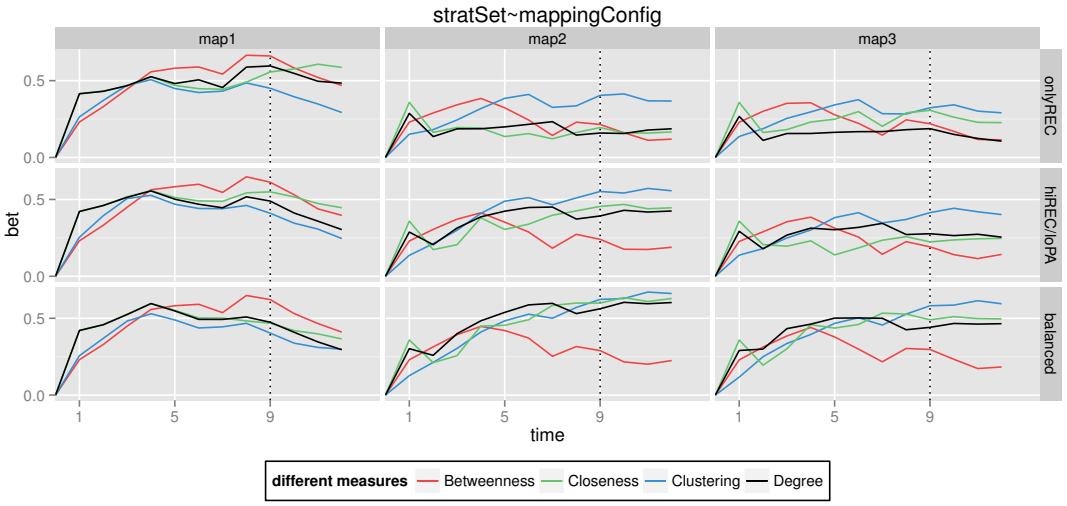


Figure 5.23: Validation results of simulation using K-S test

a Louvain algorithm (Blondel et al., 2008) and mapped communities using our naive approach described in Section 3.4.3 while usage of more efficient algorithms for community detection and evolution can help to refine input for simulations. Moreover, operating with all network members in simulation environment and not only members with peers from a single community can help to specify rules of community participation more precise and make community simulation more realistic.

Anyway after simulating communities based on *i** models community stakeholders can estimate development of their communities according to patterns of community members and network strategies they follow.

5.9 Summary

The community model creation process has been validated in this chapter. We investigated learning communities in a language learning forum and a medical student forum. Firstly, we implemented i*-REST services (Petrushyna et al., 2014b) that let us create *i** models automatically using REST requests (Fielding, 2000). We specified the *refinement* phase of the process by introducing learning forum community model for multi-agent simulation. *i** models of communities are mapped to Java classes where communities as well as community members are represented by objects. We defined two strategies for community simulation: *reciprocity* and *preferential attachment* (Schnegg, 2006). Learners in communities prefer to communicate with their peers in case of the reciprocity strategy while in case of the preferential attachment strategy learners are communicating more often if they have high degrees. Data of the forums were collected with the help of the Forum Watcher described in Section 3.3.4.1

while the analysis has been conducted using techniques described in Section 3.4. Particularly, communities were detected using Blondel et al. (2008) and Zhang et al. (2009) algorithms. Events in community evolution such as *merge* or *split* were detected with the help of the Asur et al. (2009) algorithm. We adapted both Zhang et al. (2009) and Asur et al. (2009) algorithms in a distributed environment, GPU, to ensure competitive running time for large graphs. The community detection implementation showed the sixfold advantage in running time while computing communities in networks with more than 1K edges. Further applied techniques conducted structural analysis of communities and their members and emotional and intent analysis of forum texts. We classified 1,700K sentences and detected 132K intent phrases. Moreover, we extracted learning topics and concepts of communities using named entity recognition and data from the Linked Open Data Cloud. We used the results for investigating learning phases of users in communities, detecting learner patterns and modeling learning communities as *i** models (Petrushyna et al., 2015). Results of emotional and intent analysis as well as *i** models and their simulations were successfully evaluated.

With this case study we realized the only solution that creates *i** using RESTful queries. Using the TargETLy and i*-REST services one can create and maintain *i** models automatically. Since methods used for *analysis* are approved by *i** experts and the methods realize requirements of dimensions of Community of Practice (CoP) (Wenger, 1998), they can be used for analysis of any CoP. The combination of structural and semantic analysis for community analysis is rare and for community modeling it is unique. Furthermore, we created the only solution that allows to simulate learning communities presented as *i** models considering different network strategies into consideration. Our mapping service can be used for simulation of other *i** community models as well.

Modeling of learning communities have been rarely touched in research works since it requires to work with large amounts of data and conduct a comprehensive analysis. This case study presents the first approach of modeling learning communities as i* models automatically after monitoring and analyzing their data. Similar to user modeling, community modeling gives us important hints about requirements of communities that can be used for adopting information systems according to community needs.

Chapter 6

Competence Management of Learning Communities

In the previous chapter we find that learners of about 40% communities in forums follow the Psycho-pedagogical Integration Model (Nussbaumer et al., 2011) and therefore are competent self-regulated learners in digital learning media. Other learners require a support not only in learning processes but as well in their competences to learn without assistance. Even the desire to guide learning independently it not enough. Self-monitoring and self-evaluation are defined as some of key activities for Life Long Learning (LLL) (Kitsantas, 2002; Kitsantas and Dabbagh, 2004) and applications that support learners in these activities can help them to acquire or refine their abilities to self-regulate learning.

In this chapter I operate with the *monitoring* and *analysis* phases of our methodology since community stakeholders are interested, among others, in community states and states of community users. Furthermore, we operate with data from the collaborative workspace eTwinning [1] and its initiator, the European Commission, emphasizes the role of competences, especially a role of self-reflection. Therefore, we handle with the output of the monitoring and analysis phases to perform competence management in eTwinning. Firstly, I introduce the meaning of competence and the related work in competence modeling and management in Technology Enhanced Learning. After that, I explain how we model competences in eTwinning using Social Network Analysis and Visualization. Then I describe the Competence Analyst for eTwinning application and its outcomes together with further investigations of eTwinning networks. Later the evaluation of the application is discussed.

In this chapter we monitor and analyze communities emerged due to collaborations of teachers in eTwinning. We design and implement a competence management application based on outcomes of the monitoring and analysis. Using the outcomes we support teachers in extending self-monitoring and self-reflection competences since

[1] eTwinning European teacher network `http://www.etwinning.net/en/pub/index.htm`, Last access on 13.08.2014

they are informed about their competences and competences of their peers. We find that teachers are interested in these and therefore can use the application for acquiring or extending of their LLL competences though lack of understanding of social network visualizations and measures should be solved by a better support and user-friendly design.

6.1 Learning Communities in TeLLNet

In first term I describe a project that was a trigger for the competence management application for eTwinning. The TeLLNet[2] project aimed to study the eTwinning network together with other universities using social network analysis and visualization techniques to define main actors and the reason why some teachers are interested in social networking while others are not.

eTwinning was founded in 2005 by European Commission with the purpose to facilitate collaborations among European schools and teachers. Therefore, eTwinning proposes following artifacts that enable collaborations between teachers: contact lists, e-mails, projects, blogs, guest books and prize comments. Although many artifacts enable collaborations of teachers, projects provide collaborative workspace where teachers practicing the same thema in different schools and therefore organized communities can be seen as communities of practice. Due to knowledge fluctuation and meeting of CoP boundaries teachers are learning new skills and experiencing new techniques thus we consider teacher communities in eTwinning as learning communities.

At least two teachers from different schools that are located in different European countries can create a project that has to be accepted by the National Support Service (NSS). After the approvement, the project gets its space with different services that allow to create a project blog or a guest book. The NSS performs the evaluation of projects and teachers and grant the National Quality Labels (NQL) to teachers while the Central Support Service grants European Quality Labels to projects that have at least two teachers with NQL. Further prizes in different categories are awarded each year.

6.2 Related Work

In this section I clarify the meaning of competence, competence management and discuss works that maintain competences in learning communities.

[2]Teachers' Lifelong Learning Network project `http://www.tellnet.eun.org/web/tellnet`, Last access 27.02.2015

6.2.1 Competences, Their Modeling and Usage

The Latin word 'Competere' or 'Competentia' means to be suitable. McClelland (1973) introduced the concept of competence into Human Resources Management. He focused on *the knowledge, skills, traits, attitudes, self-concepts, values, and motives* that are directly related to tasks. Competences are estimations that help to find professionals that possess skills required to accomplish tasks.

Meta-competences are high-order competences (Brown and McCartney, 1995) that are responsible for extensions and refinements of other competencies (Cheetham and Chivers, 2005). Some processes like goal setting, self-monitoring, task strategies, help seeking, and time management can invoke meta-competences (Kitsantas, 2002; Kitsantas and Dabbagh, 2004). For example, by self-monitoring learners investigate their activities and discover evidence that can help themselves to enhance their competences.

Competence modeling deals with connecting information about a behavior and experience to a skill, e.g., the PALO (Najjar et al., 2010) and EQF[3] competence models are popular models for recent applications. Competence assessment estimates abilities using direct observations, simulations, video observations, interviews, examinations of related documents, and many other approaches (Cheetham and Chivers, 2005). Explicit competence assessment methods estimate competences by asking people directly or indirectly, whereas implicit methods monitor behavior of people, competence-related events and other information to assess competences.

6.2.2 Competences in Technology Enhanced Learning

Attention to competences in *Technology Enhanced Learning* triggers the creation of the TENCompetence project[4], within which a personal competence manager was created (Vogten et al., 2008). The manager was successfully evaluated by teachers in their competence development experiment where they acquire competences under the guide of the manager (Schoonenboom et al., 2008). Tabuenca et al. (2015) support pupils and adults to foster life long learning competences using mobile phones notifications. Florian et al. (2011) used learning analytics for competence assessment tasks where they observed activities of learners and defined how actions of learners relate to competences defined in their competence models. The authors argued that students and teachers benefit from different indicators showing if a certain competence is achieved or not.

In the former research of eTwinning Vuorikari and Scimeca (2013) emphasized that only one from six teachers stayed active on the website while only one third of

[3]European Qualification Framework `http://ec.europa.eu/ploteus/search/site?f[0]=im_field_entity_type%3A97`, Last access on 22.08.2014

[4]The home page of TENCompetence `http://tencompetence-project.bolton.ac.uk/`,`Lastaccesson22.08.2014`

teachers was not engaged in project collaborations. Social network analysis and visualization (Breuer et al., 2009) can be used to motivate teachers to collaborate though the lack of understanding of visualizations need to be considered in the future design. In the following section we present a novel approach for modeling competences based on results of community analysis. We use results to foster competence development teachers' life long learning competences.

6.3 Competence Modeling in eTwinning

In this section, we discuss a competence structure in the eTwinning network that we used later. We consider following teachers' competences: *professional competences*, *social competences*, and *meta-competences* as depicted in Figure 6.1. Meta-competences encourage improvement of other competences. Self-monitoring is one crucial meta-competence in the context of LLL (Cheetham and Chivers, 2005) that describes a learner ability to take advantage investigating her activities. For example, eTwinning teachers can monitor their activities and conclude changes in their professional and social competences. Moreover, comparing with achievements of other peers can motivate the teachers to enhance their competences while information about competence of teachers in projects and schools gives a clue to community stakeholders about community states.

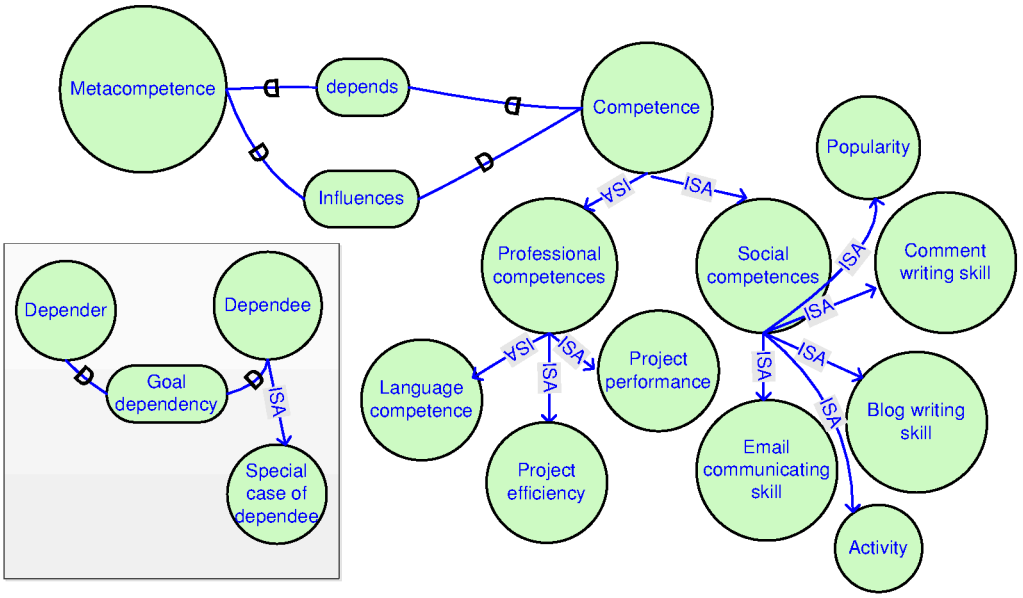


Figure 6.1: Competence Structure in eTwinning

Professional Competences in eTwinning

Professional competences involve all abilities or skills of teachers, which are necessary for performing professional tasks in eTwinning. For example, knowledge of a project subject, good idea and the same language for collaboration is essential for organizing a successful consortium. In the scope of eTwinning we can operate with media activities of teachers and thus use them for estimating professional competences.

We define two indicators for estimation of teachers' professional competences, namely *project performance* and *project efficiency*. The *project performance* describes how a teacher performs in projects in eTwinning. It depends on the number of projects that the teacher has participated in and the number of awards (NQL,EQL) that she or her projects received. The *project efficiency* is a normalized value of the project performance according to the number of projects that a teacher has participated in. Both indicators enable teachers to monitor their achievements in projects and compare their achievements with achievements of others.

Social Competences

We estimate social competence, a key competence for LLL according to European Parlament and the Council (2006), by observing activities in eTwinning as well. We consider following indicators to estimate the social competence:

- *e-mail communication skill* estimates how a teacher communicates with others using the eTwinning e-mail tool. The estimation is based on closeness centrality (Section 3.4.1) of a teacher that is represented as a node in a e-mail network where teachers are nodes that are connected if one of them writes an e-mail to another. Teachers with high closeness centrality possess hub positions thus get and spread information quickly (Newman, 2004).

- *blog writing skill* defines if a teacher can write blog posts and if these posts are popular, i.e. received comments.

- *comment writing skill* depends on the amount of comments a teacher has written and shows the readiness of the teacher to interact.

- *activity* estimates activities of a teacher in eTwinning such as sending emails, writing in a blog and commenting on blog posts, leaving messages in guest books, and writing comments devoted to eTwinning prizes or projects.

- *popularity* ranks how a teacher attracts attention of others. For example, teachers with high scores of popularity receive many emails and comments on their blog posts.

6.3.1 Modeling Assessment of Competences in eTwinning

In the previous section I introduce indicators that we use for competence assessment. We calculate the indicators using a set of factors such as network centralities (Section 3.4.1) of teachers in project or e-mail networks, the number of written blog posts and others factors mentioned in Table 6.1.

SNA Factors	
I_{em}	Indegree centrality in the e-mail network
O_{em}	Outdegree centrality in the e-mail network
C_{em}	Closeness centrality in the e-mail network
B_{em}	Betweenness centrality in the e-mail network
I_{bl}	Indegree centrality in the blog network
Statistical Factors	
PRJ	Amount of projects a teacher has participated in
QL	Amount of NQLs a teacher has gained
EQL	Amount of EQLs a teacher's project has gained
PRI	Amount of prizes a teacher has gained
QLE	QL efficiency[5]
EQLE	EQL efficiency
PRIE	Prize efficiency
EM_{out}	Amount of sent emails
EM_{in}	Amount of written blog posts
PBP	Amount of projects where a teacher has written blog posts
BC_{out}	Amount of written blog comments
BC_{in}	Amount of received blog comments
PC	Amount of written prize comments
PRC	Amount of written project comments

Table 6.1: The set of factors

Before using factors, we normalize them using the so-called z-score (Larsen and Marx, 2000; Woolf et al., 2004). A value of an indicator is then calculated as:

$$I = \sum_{f \in F} w_f \cdot Norm(f) \ ,$$

where w_f is a factor weight, F stands for a factor set related to an indicator I, f denotes a single factor in the set F, and $Norm(f)$ is a normalization value of f.

We design an extendable template for defining indicators and factors. A factor consists of following attributes:

- *identifier*: a globally unique label that identifies the factor definition
- *name*: a single mandatory text label for the factor. This is a short human-readable name for the factor

- *description*: an optional human-readable detailed description of the factor
- *assessment*: an assessment method as program code.

For the specification of an indicator, the following attributes are needed:

- *identifier*: a globally unique label
- *name*: a mandatory text label
- *category*: this mandatory attribute defines a competence type the indicator contributes to
- *description*: an optional human-readable detailed description of the indicator
- *assessment*: it involves a set of factor-weight pairs. A factor-weight pair consists of an *identifier* of a related factor, and a corresponding weight for the factor.

Examples of the factor-weight pairs are presented in Table 6.2.

6.4 Monitoring and Analysis

In the following, I describe how the *monitoring* and *analysis* phases are realized for eTwinning in a competence management tool. The Competence Analyst for eTwinning (CAfe) uses data dumps from the eTwinning project, provided by the European Schoolnet[6]. All data in the dumps is anonymous and include teacher interactions over 2 years. Our findings are based on 133K teachers, 72K institutions, 32K blogposts, and 17K projects.

Next I introduce briefly the system architecture of the CAfe that is used for monitoring and analysis of teacher activities, defining community states and other facts interesting for community stakeholders.

6.4.1 Competence Analyst for eTwinning

The system architecture of the CAfe consists of four modules. The *database* module is the interface connecting the data warehouse and other modules. The *network* module constructs collaborative networks where nodes are teachers and edges of the nodes are defined by project collaborations, e-mail communications, blog writing activities or comment writing activities. The *competence* module estimates teacher competences. The *visualization* and Graphical User Interface module visualizes networks and presents results of analysis such as competences, patterns in networks and comparison of competences in communities.

[6]the European Schoolnet `http://www.eun.org`, Last access on 08.10.2014

CAfe supports needs of two types of stakeholders: teachers and researchers or managers. CAfe visualizes collaborative networks of teachers and varies these networks depending on types of connections between teachers. A teacher network in Figure 6.2 can be changed selecting only teachers that register on a particular time interval in the eTwinning portal or only interactions that happened in a particular time interval. Furthermore, we can change visualizations by selecting edge types that are defined through project collaborations, e-mails, blog comments and prize comments. Moreover, we can visualize nodes or edges depending on their properties using node size or color and edge color or thickness. Node (teacher) properties are a country, an occupation, a registration year, and social network measures (described in Section 3.4.1) while edge properties are a collaboration occurrence frequency and types of collaborations.

CAfe provides the monitoring of teacher competences by visualizing teacher competence values and their development. To estimate competences we use weights and factors from Table 6.2 that help to calculate competence indicators. Figure 6.3(a) depicts a competence teacher report that shows the current state of teacher competences. One can check the development of competences for any teacher by choosing a node from a network visualization. CAfe users can navigate from teacher competence reports to community competence reports.

Competence Indicators	Factors*weights
Project Performance	$PRJ * 0.5 + QL + EQL * 1.5 + PRI * 2$
Project Efficiency	$QLE + EQLE + PRIE$
E-mail Communication skill	$I_{em} * 0.5 + O_{em} * 0.5 + C_{em} * 0.5 +$ $+B_{em} * 0.5 + EM_{out} + EM_{in}$
Blog writing skill	$BP + PBP * 0.5 + I_{bl} + BC_{in} * 0.5$
Comment writing skill	$PRC + BC_{out} + PC$
Activity	$PRJ + EM_{out} * 0.5 + O_{em} * 0.5 + BP * 0.5 +$ $+PBP * 0.5 + BC_{out} * 0.5 + PC * 0.5 + PRC * 0.5$
Notability	$EM_{in} * 0.5 + I_{em} + BC_{in} * 0.5 + I_{bl}$

Table 6.2: Factors and weights used for competence indicators

Communities of the competence reports are defined by projects or schools. An example of such a competence report on Figure 6.3(b) shows the development of project efficiency over the whole period of community existence in the eTwinning portal. Another example of a community competence report in Figure 6.4 depicts indicators of all project peers in a bar chart that allows to compare competences easily. Competence reports for communities are as well customizable: competence indicators, time frames and graph types can be chosen.

The CAfe tool differentiates between learners using learning patterns such as *project performance star, project efficiency star, email communicator, blog writer, comment*

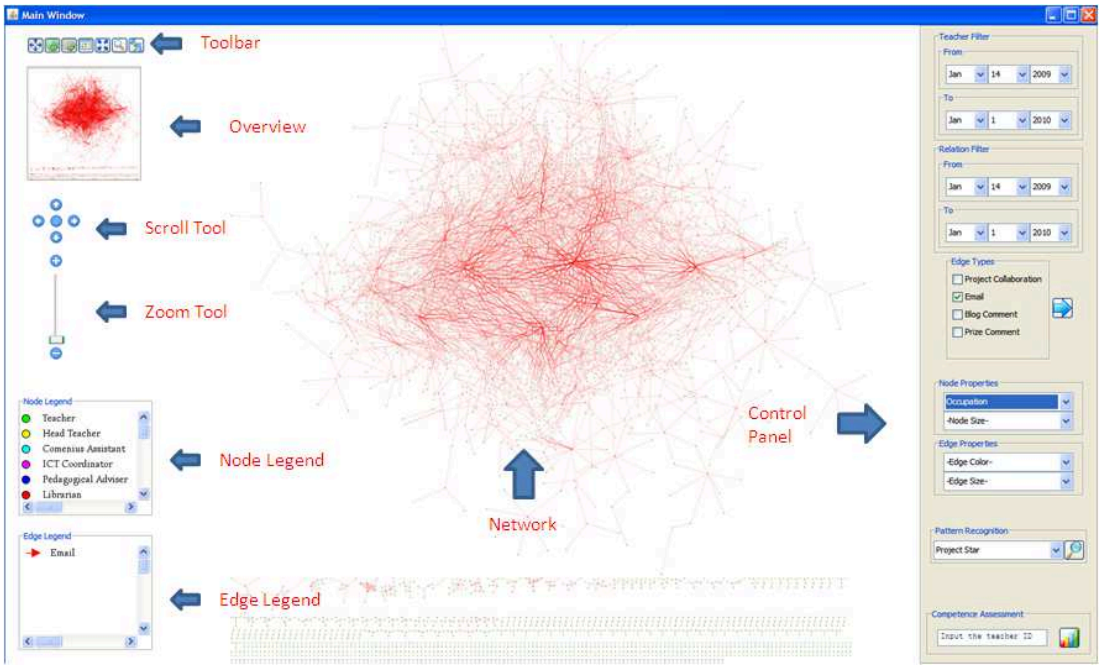
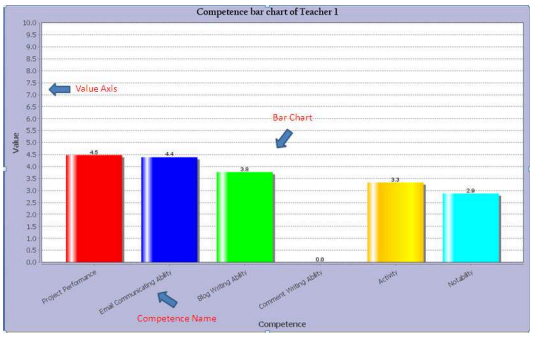
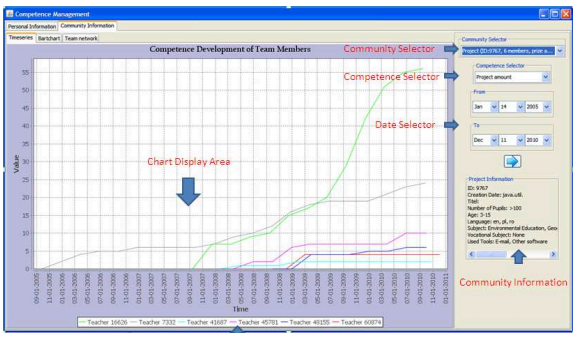


Figure 6.2: eTwinning teacher project network



(a) Competence Report of a Teacher



(b) Competence Development of a Teacher community

Figure 6.3: Examples of competence reports on individual (a) and community (b) level

writer, activist and *notable teacher*. The patterns are defined using competence indicators that can be easily changed or added in a pattern description. *Project stars* are teachers with the highest amount of projects while *project efficiency stars* are remarkably successful in projects and therefore received awards for the majority of these. An *email communicator*, a *blog writer* and a *comment writer* are exceptionally active in e-mail, blog and comment writing correspondingly. An *activist* and *notable teacher* have high scores in activity and popularity indicators respectively.

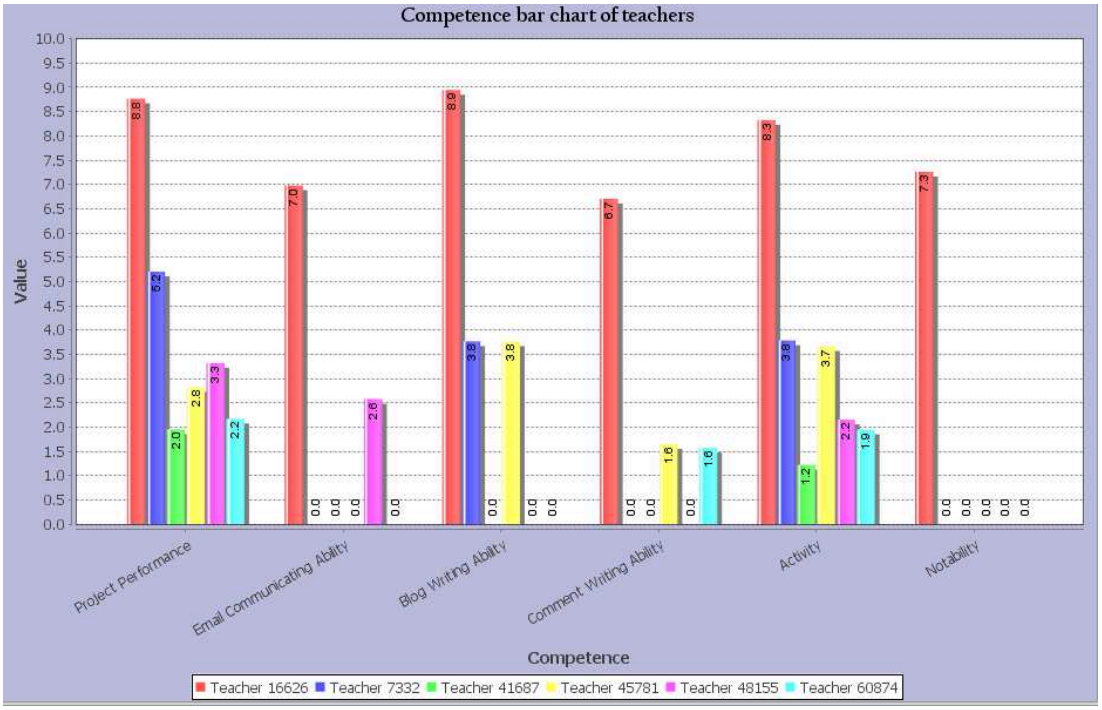


Figure 6.4: An example of a community competence report

6.4.2 Network Analysis of eTwinning

We investigate degree distribution of teachers in the network of project collaborations (Figure 6.5) and find that it follows the power law (Pham et al., 2012). In complex network theory, the power law degree distribution indicates that super connectors (or hubs) exist. These refer to nodes that connect many isolated nodes or communities. They play an important role to ensure the connectivity, the information spreading, and behavior cascading in networks. They also have more power and control over a network than other nodes.

Since teachers' performance is recognized by prizes we take quality labels as a performance and reputation indicator and find the correlation between the performance and teachers' positions in the eTwinning projects' network. We compute betweenness and clustering coefficient (check Section 3.4.1) as functions of the number of quality labels and depict results in Figure 6.6. Nodes (eTwinners) with a high number of quality labels have very high betweenness and low local clustering coefficient while nodes with a low number of quality labels have low betweenness and very high local clustering coefficient[7]. For community managers a position of a teacher in a project network can, thus, be an indicator or a predictor for a notable teacher's performance.

[7]the high local clustering coefficient shows that a node is located within communities and not in between

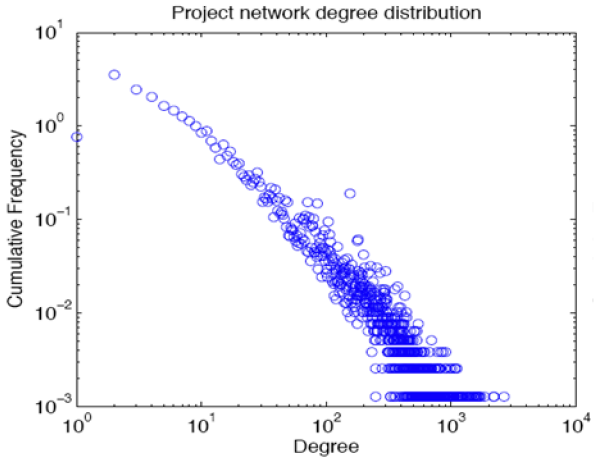


Figure 6.5: The degree distribution in the eTwinning network of project collaborations

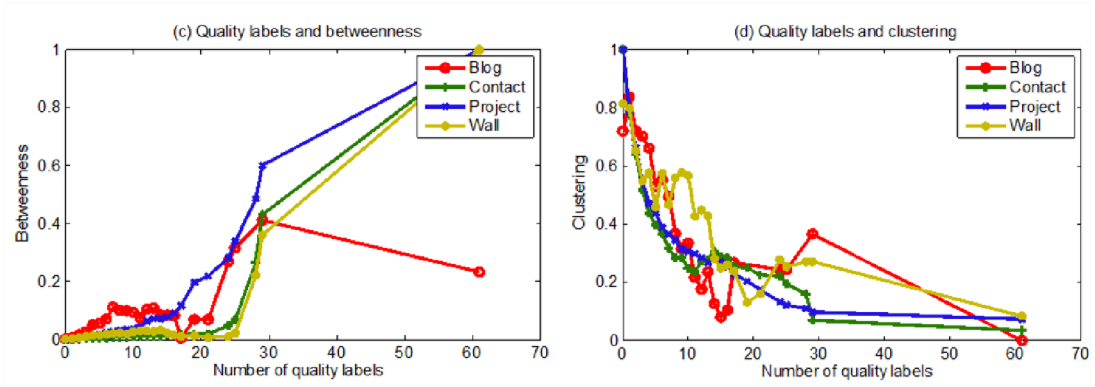


Figure 6.6: Dependencies of the quality labels from betweenness and local clustering coefficient

6.5 Evaluation

We evaluate CAfe by surveying teachers, researchers, and students. An on-the-spot evaluation of CAfe was organized in the project meeting of the TeLLNet[8] project. Participants of the meeting, experienced researchers, used CAfe and expressed their impression in interviews and questionnaires. All participants reacted to all CAfe functionalities in average positively (check Figure 6.7) though a few of the survey participants found pattern detection (2), competence development reports (1) and community reports (1) not interesting and relevant for a such kind of tool as CAfe.

Besides the researchers we evaluated CAfe with computer science students that have no experience with eTwinning and no understanding of network visualization and meaning of competences. Most users found that CAfe is relevant as a source of

[8]Teachers' Lifelong Learning Network - LifeLong Learning Project `http://www.tellnet.eun.org/web/tellnet`, Last access on 15.10.2014

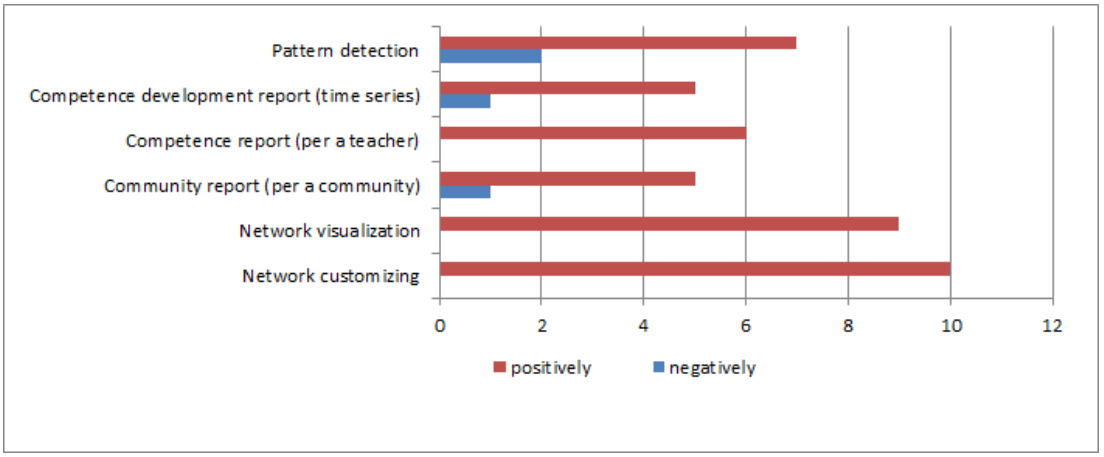


Figure 6.7: Number of users estimating CAfe functionalities in the project meeting

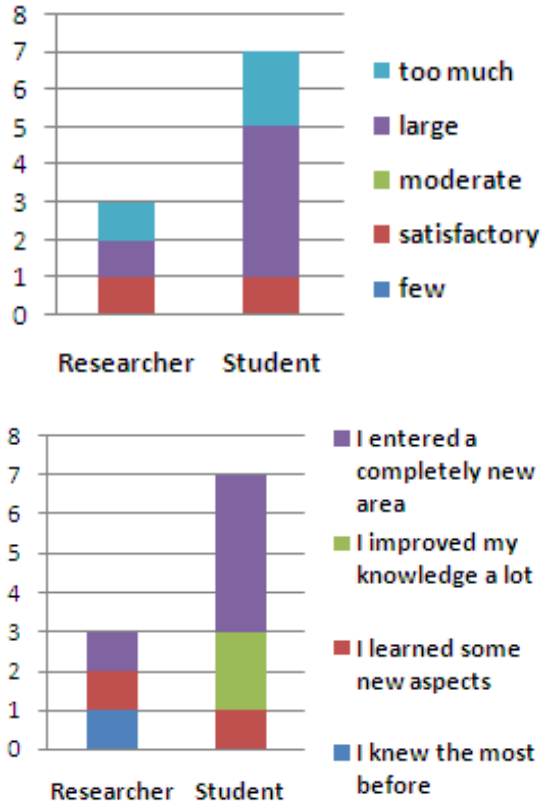


Figure 6.8: Evaluation of information CAfe provides

additional information (check Figure 6.8) thus can be used by community stakeholders for self- and community monitoring. Furthermore, many users agreed that CAfe improved their knowledge or introduced them to a new topic of monitoring activities using network visualizations (check Figure 6.5). The major drawback we find from

interviews and comments is a lack of understanding of social networks. Therefore many users requested for help functionalities that explain competence reports better (Figure 6.9).

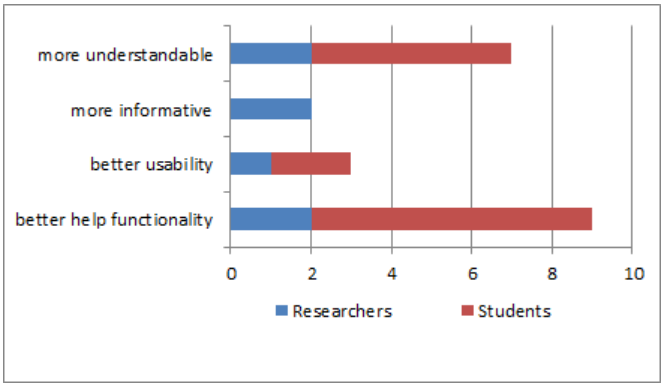


Figure 6.9: Requirements of CAfe users

Furthermore, 16 of 20 most active teachers of eTwinning in Germany showed their interest in observing their networks of collaborations, development of their competence indicators and indicators of others. It is a promising result, since the teachers had no additional incentive in examining their activities, competence indicators and community competence indicators. It proves the need of the application such as CAfe that foster competence development of teachers.

6.6 Summary

In this chapter, we applied our methodology for *monitoring* and *analysis* of European teacher networks that organize projects with shared topics and collaborative spaces in eTwinning. Since teacher professional development is more successful in informal settings social networking can serve as one way of competence development. Therefore, we monitored and analyzed teacher and community activities. We presented the only solution to teacher competence assessment using social network analysis and visualization. Using CAfe, the competence analyst for eTwinning, teachers can monitor their activities and compare them with activities of teachers in their networks consisting of peers from projects and schools. Our application aims not only to inform teachers but as well to trigger development of their metacompetences, such as self-monitoring and self-reflection that are pivotal for life long learning (Kitsantas, 2002; Kitsantas and Dabbagh, 2004). Furthermore, CAfe presents information for other stakeholders such as researchers or eTwinning managers that can estimate failures or success of teachers and communities viewing CAfe visualizations. To prove the relevance of network analysis we showed that eTwinning network based on project collaborations is a complex network and awards of teachers correlate with betweenness scores positively and local clustering negatively.

Our experience in modeling competences of teachers based on results monitoring and analysis, particularly competence factors and indicators, can be applied for investigation of communities in collaborative workspaces. Furthermore, results of network analysis indicate the approach for detecting efficient collaborators in workspaces.

Chapter 7

Cultural Analysis of Wikipedia Communities

In the previous chapter we have observed activities of users in a collaborative environment, eTwinning. eTwinning target is to provide European teachers to collaborate with each other without boundaries created by borders or cultures. Test beds of the ROLE project have been utilized in different countries as well. To satisfy individual and community needs (Giovanella et al., 2014) designed frameworks and media have to consider differences in learning caused by learners' origins (Uzuner, 2009) similar to McLoughlin and Oliver (2000) and Gunawardena et al. (2003) that used results of their studies to propose instructional design rules for creating culture-sensitive online learning courses.

To study differences in learning collaborations we take 13 Wikipedia instances as an example since their data is freely available and its amount is relevant to retrieve statistically significant outcomes.

The description of this chapter is organized as following. First of all, I give an insight into cultural theories and studies devoted to Wikipedia. After that i observe studies of Wikipedia networks that focus on cultural differences and network analysis of Wikipedia networks. Later in this chapter i describe monitoring and analysis of Wikipedia that help to detect cultural difference. Outcomes are described later as requirements for digital learning environments that can be used by the ROLE and eTwinning projects.

The contribution of the approach is in an innovative way of measuring differences between cultures by detecting collaboration patterns in communities and defining their diversity. Results of this study show that differences in collaborations of users rarely correlate with differences denoted in other estimations of cultural differences that have done using surveys (Hofstede, 1991; Schwartz, 2008). Defined differences in communities highlight some requirements that can help designers of media to satisfy learning communities' needs where communities' peers belong to one of investigated in the study countries.

First of all, I describe cultural theories used for the case study and works that

investigated cultural differences in Wikipedia.

Cultural Theories and Studies

Representatives of different cultures were characterized by their beliefs and values that were extracted based on surveys of culture representatives (Hall, 1976, 1983; Hofstede, 1991; Kluckhohn and Strodbeck, 1961; Trompenaars and Hampden-Turner, 1998). The seminal work of Hofstede (1991) proposed cultural dimensions estimated by surveying IBM workers in over 50 countries. Results of the work explained and estimated in numbers differences between cultures. But the work was conducted only in one company and therefore only a specific kind of representatives was surveyed. In the following i present dimensions defined by Hofstede as I use them later to explain differences between user activities coming from different cultures.

The *power distance* (PD) dimension ranks a relation to social inequality. The acceptability and expectancy of the power of members within societal institutions like family, school, or a community at work defines the PD. Another, important for this study, dimension identifies if everyone is responsible for him/herself (individualistic cultures) or groups are responsible for their members (collectivistic cultures). *Individualism* describes the situation when a person thinks about her own interests first while in *collectivism* a person thinks about group interests first.

In contrast, Schwartz (2012)'s cultural values were created on surveys of culture representatives belonging to different groups. Furthermore, Schwartz explicitly considered collaborative situations of respondents to estimate their values. We pay attention to *egalitarism*, *hierarchy*, *embeddedness* and *autonomy* values. *Egalitarism* states for the desire to cooperate with others avoiding negative outcomes and enhancing the welfare of all people; while *hierarchy* advocates the respect for the social power and authority. *Embeddedness* states for the respect of social relationships associating people as parts of a group while *autonomy* emphasizes self-direction, creativity and exciting life. Both cultural dimensions and values can give an insight onto differences of learners while collaborating in online communities.

Studies of cultural differences between Wikipedia collaborators investigated correlations of activities in Wikipedia and Hofstedes' dimensions. Hara et al. (2010) analyzed four Wikipedia of various sizes and different cultures, two Wikipedia belong to eastern culture (Japanese and Malay) and two belong to western culture (English and Hebrew). Courtesy behaviors in the Wikipedia of eastern countries were explained by greater respect of hierarchical structure in society and preferences of working collectively (Hofstede, 1991). While authors from Wikipedia of western countries disagree more often due to shorter power distances in these countries.

The activities around the article "game" in four different Wikipedia were measured as well using Hofstede dimensions. Pfeil et al. (2006) analyzed the article from French, German, Japanese and Dutch Wikipedia. The authors found correlations between some dimensions and activities and in doing so proved that Wikipedia is a culturally dependent place.

The described studies in Wikipedia analyzed just one particular article (Pfeil et al., 2006) or a few of Wikipedia (Hara et al., 2010). Furthermore, the studies uses Hofstede dimensions which correctness is under debate. Both works indicated the need for the further research with the use of many Wikipedia instances and for more cross-cultural analysis of non-Western countries.

Wikipedia Network Analysis

Wikipedia provides a massive data set for a cross-cultural analysis as it collects contributors from all over the world collaborating with each other on Wikipedia instances differed by languages that define cultures. Wikipedia articles are created based on contributors' collaborations. Information about article revisions and their authors create a great opportunity to build author and article networks. These networks can be used in analyzing differences between Wikipedia communities hosted on various Wikipedia instances and thus originated from different cultures.

Voss (2005) was the first who analyzed Wikipedia networks. He mainly investigated the German Wikipedia and its network of articles. Articles as nodes are connected if they are linked to each other. Voss compared namespaces of the German, Japanese, Danish and Croatian Wikipedia. He found similar structures in the German and Japanese Wikipedia that had much more media talk pages comparing to the Danish and Croatian Wikipedia. But Voss focused on precise investigation of the German Wikipedia and omitted further explanations of differences or similarities that he observed in the namespaces.

Zlatic et al. (2006) examined precisely 11 Wikipedia networks of articles. The authors found that most of the Wikipedia networks are complex networks and results of their network measures are close to each other. While the results in some networks like Korean and Bulgarian are different, explanations of these differences were missing.

Nemoto and Gloor (2011) examined the English, Japanese, German, Korean, and Finish networks in Wikipedia user talks[1]. Their approach was based on 3-month sliding window networks. The number of edges and nodes in the networks were stable for the English and German Wikipedia and were fluctuating for the Japanese and Korean. The authors detected similarities in clustering coefficients in networks of different Wikipedia while the group degree centrality was the highest for the Japanese Wikipedia. They explained it through the hierarchical culture of the Japanese.

Klamma and Haasler (2008a) used Wikiwatcher (described in Section 3.3.1.3) to visualize different Wiki projects (Berlin Wiki, Google Wiki, Aachen Wiki) and to observe their changes in time. They found that registered users often serve as connectors in networks of anonymous users. Moreover, they showed that a tiny number of Wikipedia contributors had created or edited the majority of articles.

[1] a user talk is a discussion on a user page

All conducted works detected differences between user activities of Wikipedia instances. Some of the works even tried to explain the differences using Hostede dimensions though none of them considered recent studies from (Schwartz, 2012). Furthermore, the works indicated the need for the investigation of differences between broad number of Wikipedia instances. Some works validated the use of network analysis for detecting differences in Wikipedia instances. Therefore, relying on existing works, we monitor and analyzed 13 Wikipedia instances with help of network analysis.

7.1 Monitoring

In the following I describe how Wikipedia instances are monitored with our assumptions and limitations. After that i introduce the findings based on the monitoring according to users, user edits in articles, user locations, and participation in many Wikipedia instances.

7.1.1 Data Set

Using WikiWatcher (Klamma and Haasler, 2008a) we extract author networks from Wikipedia data dumps. We analyze the Wikipedia data starting from June, 30, 2001 till January, 1, 2009 and divide it into 16 time windows, half a year each. We choose both European and Asian Wikipedia. The instances are selected according to their size: large European Wikipedia (Spanish and Russian), large Asian Wikipedia (Japanese and Turkish), small European Wikipedia (Bulgarian, Catalan, Danish, Greek, Macedonian, and Ukrainian) and small Asian Wikipedia (Arabic, Hindi, and Korean). The set of small European Wikipedia instances includes Wikipedia of different Slavic languages (Bulgarian, Macedonian, and Ukrainian) and the Catalan Wikipedia, the Wikipedia instance of a one of minority language groups in Spain.

Selecting different Wikipedia instances we consider *power distance*[2] (Hofstede, 1991). In our data set we have the Danish Wikipedia with high *individualism* and the Korean with high *collectivism* scores. Moreover we operate with Wikipedia instances that belong to cultures with high *embeddedness* like in Slavic and Eastern countries or high *egalitarism* like in countries of Western Europe (Schwartz, 2008).

7.1.2 Assumptions and Limitations

The conducted study operates with activities of representative samples that are limited only by users of Wikipedia and not by their occupation or gender as it was in other studies such as (Uzuner, 2009). We consider Wikipedians as learners and their communities as learning communities since they acquire and share knowledge (check Section 3.3.1.3 for further details). Wikipedians or our samples decide to be registered

[2] from high respect for the hierarchy in Russia to relations based on equality in society in Denmark

or anonymous. Registered samples use names for their identification while anonymous samples are identified by Internet Protocol (IP) addresses. The samples are comparable as 1) they are using the same Wikimedia technology; 2) the number of registered Wikipedians in the investigated Wikipedia instances varies from 0,09% to 0,3%[3] of Internet users of a corresponding to Wikipedia country. Therefore, this study is more careful about used data while in many previous studies devoted to cultural differences surveys included the small amount of samples. Furthermore, just a few of the studies estimated culture differences not only based on surveys but as well based on interactions (Uzuner, 2009). Further assumptions and limitations are listed below.

- We operate with geographical location of anonymous users while geographical location of registered users is unknown as their IP addresses are hidden.

- It is possible that one person has several accounts (registered and anonymous). The Internet Provider address (IP) for an anonymous user can not serve as an identification of an anonymous user as most IPs change by Internet Service Providers continuously and using an unknown patterns. Therefore, we assume that 1) if one person has a registered account, it is the only account she has; and 2) we are talking not about anonymous contributors but about anonymous contributions that have been done by not registered contributors.

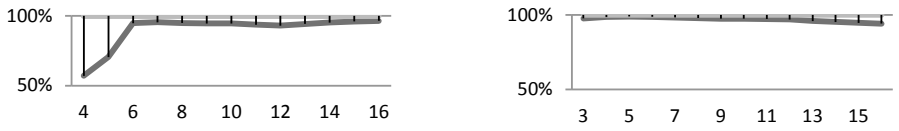
- Each of the Wikipedia instances has a language that connects the instance to one or more countries where the language is spoken (Hindi, Spanish and Arabic are exceptions). We assume that the majority of contributors are native speakers even if they contribute from countries where an official language is different to a Wikipedia instance language.

- Arabic and Spanish languages are native in many countries of the world. Cultural values of Arabic countries are close (Schwartz, 2008) while cultural values of Spanish-speaking countries in Latin America, in Central America and Spain are different. The Spanish and Arabic Wikipedia are edited a lot from many different parts of the world (Yasseri et al., 2012). Investigating both Wikipedia, we have to consider the difference of contributors' cultures.

7.1.3 Users and Edits

The ratio of registered users to all users in many Wikipedia instances is very low. From the beginning of existence the number of anonymous users of many instances is low, e.g., the Turkish Wikipedia in Figure 7.1. While in some of the instances the number of anonymous users is high from the beginning, for example, in the Danish Wikipedia (Figure 7.1).

[3]except of the Hindi Wikipedia, where at most 1/10000 of Internet users contribute to the Wikipedia since the beginning till 2009

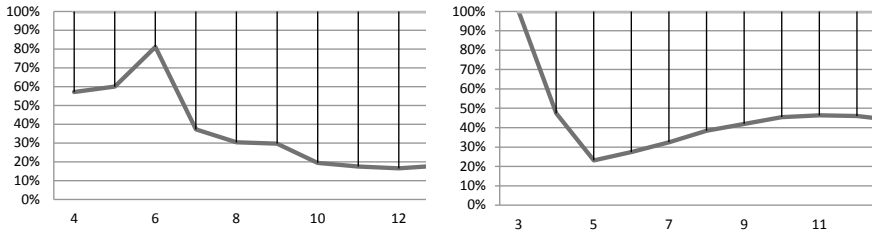


(a) Turkish (b) Danish

Figure 7.1: The development of the ratio of registered (above the line) to anonymous (under the line) contributors in Turkish (left) and Danish (right) Wikipedia.

In general, anonymous users make less than 20% edits in Wikipedia articles though the contributors from Spanish, Turkish and Japanese Wikipedia instances have a different behavior (Figure 7.2).

The reason for the low number of contributions from anonymous users as well as for the high number of anonymous users is their identification through IPs. Many contributions of one unregistered user are viewed as contributions from many anonymous users.



(a) Turkish (b) Japanese

Figure 7.2: The ratio of edits done by registered (above the line) and anonymous (under the line) users

We compare the ratio of involvement between anonymous and registered users by capturing ratios of pages that registered and anonymous users contribute to. In most of the cases anonymous users contribute to the most of articles in the beginning. But later the ratio of pages edited by registered users increases tremendously up to 100%.

Users in the Japanese and Ukrainian Wikipedia show different behavior. Many active Japanese users stay anonymous. Therefore about 50% of Wikipedia pages are edited by anonymous users. In the Ukrainian Wikipedia most of articles are edited by registered users while only 10% of all articles have been edited by anonymous users.

Moreover, we find following peculiarities of Wikipedia instances in Table 7.1.

Wikipedia	Anonymous contributors	Registered contributors	Revisions till 2009	Average number of edits	Number of pages in 2009
Arabic	320K	16K	2.4M	3.5	384K
Ukrainian	110K	5K	2.1M	3.89	338K
Macedonian	26K	1K	0.5M	4.25	65K
Catalan	260K	8K	3M	4.53	361K
Hindi	22K	1K	0.3M	4.52	50K
Turkish	1,054K	30K	4.2M	5.18	492K
Russian	1,520K	42K	10.9M	6.15	1.239K
Danish	305K	15K	2.7M	6.54	253K
Korean	248K	9K	2.6M	6.58	227K
Greek	190K	6K	1.2M	7.07	94K
Bulgarian	265K	7K	2M	7.33	155K
Spanish	5,178K	158K	19M	8.46	1.439K
Japanese	8,869K	106K	21M	10.04	1.398K

Table 7.1: Statistics of Wikipedia instances. The number of anonymous and registered contributors, revisions, and pages are rounded.

- The Turkish Wikipedia has twice as much edits per article on average than the Arabic Wikipedia though Arabic countries' representatives are close in cultural values to Turkish representatives (Hofstede, 1991; Schwartz, 2008). The reason can be that the number of registered users in the Turkish Wikipedia is twice as much as in the Arabic Wikipedia. Furthermore, Rask (2007) stated that countries with higher human development index (Turkish has a higher index than Arabic people) contribute to and benefit more from Wikipedia than others.

- The Bulgarian Wikipedia has the highest average number of page edits and the highest number of registered users between small and middle Slavic countries, though Ukrainian population is nine times as large as than Bulgarian population while the number of Bulgarian Internet users is only half as high as the number of Ukrainian Internet users in 2009[4].

- The Ukrainian and Greek Wikipedia have similar numbers of registered users

[4]The World Factbook, Internet usershttps://www.cia.gov/library/publications/the-world-factbook/rankorder/2153rank.html, Last access on 19.05.2015

but different number of average edits per article. The Greek Wikipedia articles have a higher number of revisions per article on average thus Greek Wikipedians are more active than Ukrainian ones.

- The Korean Wikipedia with 12K users and the Danish Wikipedia with 17K users have approximately the same average number of edits and revisions. While the Bulgarian Wikipedia with 8K registered users has even higher average number of edits per article than mentioned Wikipedia instances.

- The number of Wikipedia users is not influenced by population size or Internet population of a country. For instance, Russia has nearly twice as much citizens and about twice as much Internet users as Turkey[5] but the number of Turkish and Russian Wikipedians are similar. Nearly 64,8M of Hindi are Internet users in India, the Hindi Wikipedia has at least 1K contributors[6] while the Danish Wikipedia has 15K contributors with 86M Internet users.

7.1.4 Geographical Location of Anonymous Users

The location of anonymous Wikipedia users can be identified by Internet Provider addresses (IPs) using country codes from a freely available database [7]. These codes define locations of anonymous users that perform in most of investigated Wikipedia about 20% of edits. Nevertheless, we suppose that registered users, that contribute more to Wikipedia, have the same locations as anonymous users. Therefore, here we hypothesize that monitoring of geographical locations of anonymous users points out geographical distribution of registered users in Wikipedia instances since each Wikipedia has its common working hours (Yasseri et al., 2012).

We investigate anonymous contributors in case they are located in Germany and the U.S. as these countries usually include a high number of immigrants. We find that 40% of anonymous contributions in the Japanese, 27 % in the Danish and 15% in the Spanish Wikipedia have been done by users located in Germany. The number of immigrants in Germany for Japanese is 30K, for Danish is 18K and for Spanish is 140K though in case of Spanish contributors we should consider as well immigrants from other Spanish-speaking countries to make some conclusions. We suppose that representatives of Japanese and Danish immigrants in Germany are very active as Wikipedia contributors though we can not say if contributors located in Germany are immigrants or just visitors.

Users located in the USA perform 11% of all anonymous contributions in the Greek

[5]The World Bank, Data about internet usershttp://data.worldbank.org/indicator/IT.NET.USER.P2, Last access on 22.10.2014

[6]we are not considering anonymous contributors

[7]IP2Country mapping database http://www.ip2country.net/ip2country/ip_country.html, Last access on 24.10.2014

Wikipedia, 8% in the Arabic Wikipedia and 6% in the Hindi Wikipedia. But the numbers of immigrants in the USA is tremendously different from the numbers in Germany: 1.4M for Greek, 1.5M for Arabic and 1.5M for Hindi immigrants. Investigation of unusual locations of Wikipedians or users of any other media can suggest how national minorities preserve their language being abroad.

7.1.5 Cross-Wikipedia Users

A few Wikipedia users contribute to more than one Wikipedia, so called cross-Wikipedia users. We count contributions of the same authors (registered or anonymous[8]) in Table 7.2. A majority of the cross-Wikipedia anonymous contributors are from Germany (15 %) and the United States (11 %) (Figure 7.3).

Considering the set of our Wikipedia instances we count 1,7K contributors that did 618K edits to more than 3 Wikipedia (Table 7.3). We investigate instances cross-Wikipedia users contribute to in case of at least 3 Wikipedia. In many cases they edit articles in the Russian, Japanese and any other Wikipedia from our list of investigated instances (Figure 7.3).

Count of manipulated Wikipedia	Anonymous contributors	Registered contributors	All contributors
5	10	2	12
4	1,493	213	1,706
3	9,028	1,634	10,662
2	44,067	9,596	53,663
1	4,374,043	478,940	4,852,983

Table 7.2: Wikipedia contributors that edited one or more Wikipedia instances

7.2 Analysis

The previous section includes findings from the set of Wikipedia instances. Following the example of studies of Pfeil et al. (2006); Hara et al. (2010) I analyze Wikipedia author networks and author activities involving cultural theories from Hofstede (1991) and Schwartz (2008). Moreover, we visualize author networks and watch their growth using dynamic network analysis.

7.2.1 Cultural Differences

Previous work argued that due to openness in the digital world cultures with high power distance can profit since power distance influence is much less in the digital

[8] anonymous cross-Wikipedia users are identified by the same IPs

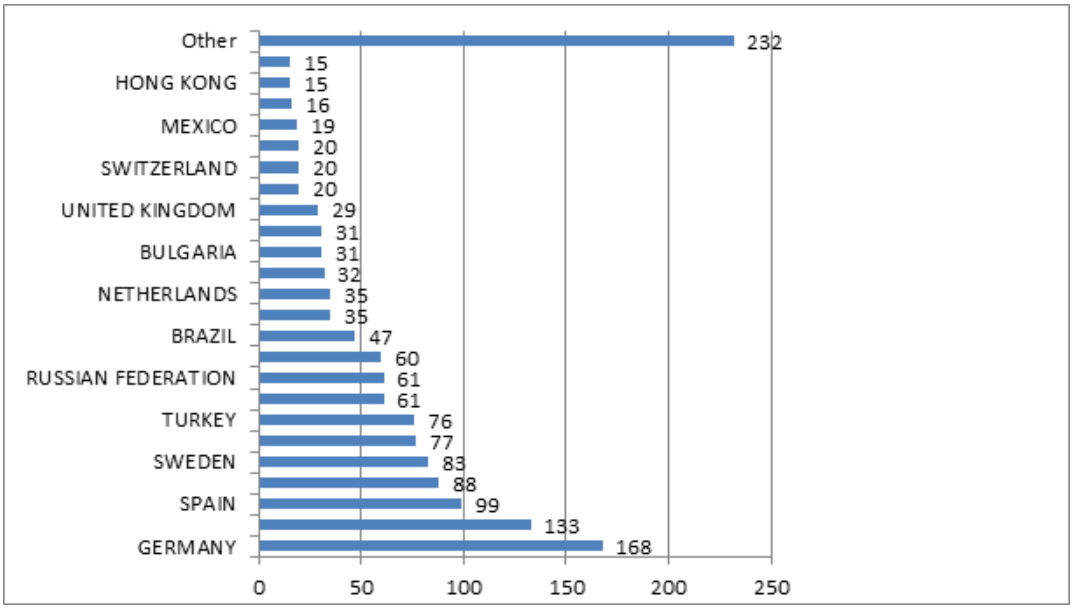


Figure 7.3: The geographical location of anonymous contributors that manipulated with articles in more than 3 Wikipedia instances

Cross-Wikipedia users	Wikipedia	Number of edits
505	Arabic	69,513
550	Bulgarian	24,060
291	Catalan	51,560
152	Danish	7,701
185	Greek	213
283	Spanish	5,387
91	Hindi	117
1,653	Japanese	75,642
119	Korean	231
52	Macedonian	14,389
1,698	Russian	314,323
1,159	Turkish	49,922
594	Ukrainian	5,525

Table 7.3: Statistics about edits of cross-Wikipedia users that manipulated more than 3 Wikipedia instances.

world (Gunawardena et al., 2003). Comparing to European countries Arabic and Turkish cultures have high power distance, e.g. respect for family. The Turkish Wikipedia has the higher average number of revisions per article than the Arabic as well as has a higher number of registered users (Table 7.1). Schwartz (2008) captured the difference between those cultures by the *embeddedness* value while Hofstede (1991) mentioned

difference in *power distance* (PD) dimension. Turkish people have less *embeddedness* and higher affective and intellectual *autonomy* than Arabic people while Arabic people have higher PD value (Hofstede, 1991). The affective and intellectual *autonomy* can result that Turkish Wikipedians are more engaged to contribute to Wikipedia than Arabic Wikipedians.

The Russian and Turkish Wikipedia are different in size (Table 7.1) and the average number of edits per article is higher in the Russian Wikipedia. *Embeddedness* values of Russian and Turkish people are similar while PD is much higher in the Russian culture. Differences in *embeddedness* or PD are not influencing differences in the average number of revisions per article. This finding is supporting conclusions from (Gunawardena et al., 2003) about *embeddedness* and PD.

Representatives of Slavic countries such as Ukraine and Macedonia have high *embeddedness* and *hierarchy* values (Schwartz, 2008). The average number of edits per article in their Wikipedia instances are much lower than those in the Greek and Danish Wikipedia that belong to cultures with much more respect to opposite values like *egalitarism* and *autonomy*. The other Wikipedia from a Slavic country, Russia, has a higher value for the average number of edits per article than in Ukrainian and Macedonian but it has a higher number of contributions/contributors as well. The Bulgarian Wikipedia is exceptional in our case: it has the highest average number of edits per article between Slavic cultures. Schwartz (2008) states about similarities of Slavic countries though Bulgarian Wikipedians are much more active than other Wikipedians from Slavic countries. Comparing Ukrainian with Greek and Arabic with Danish Wikipedians (since these Wikipedia instances have a similar number of registered users) we find the correlation between *autonomy* and the average number of edits: the higher *autonomy*, the higher is the average number of edits.

Most Wikipedia contributions from our dataset are anonymous. Even so, articles are created and edited mostly by registered users (more than 80 % of content). The Japanese Wikipedia is exceptional as their anonymous users create or edit 45% of articles. Ishii and Ogasahara (2007) found as well that Japanese prefer to stay anonymous. Japanese anonymous users are much more active then anonymous users in other Wikipedia.

7.2.2 Dynamic Analysis of Wikipedia Author Networks

With the help of Wikiwatcher (Section 3.3.1.3) we analyze networks of Wikipedia instances that emerge in previously defined 16 time windows. Firstly, we visualize networks of registered contributors. We define authors as nodes and their connections show if the authors revise the same articles. Nearly all networks' visualizations from our Wikipedia set follow the same pattern as visualized in Figure 7.4. Most registered users belong to a strongly connected component of a network that is the biggest group of nodes in the network where a node can reach any other node from the group. For example, authors that are working on popular articles are usually connected.

Secondly, we visualize a network of anonymous contributors (Figure 7.5). The

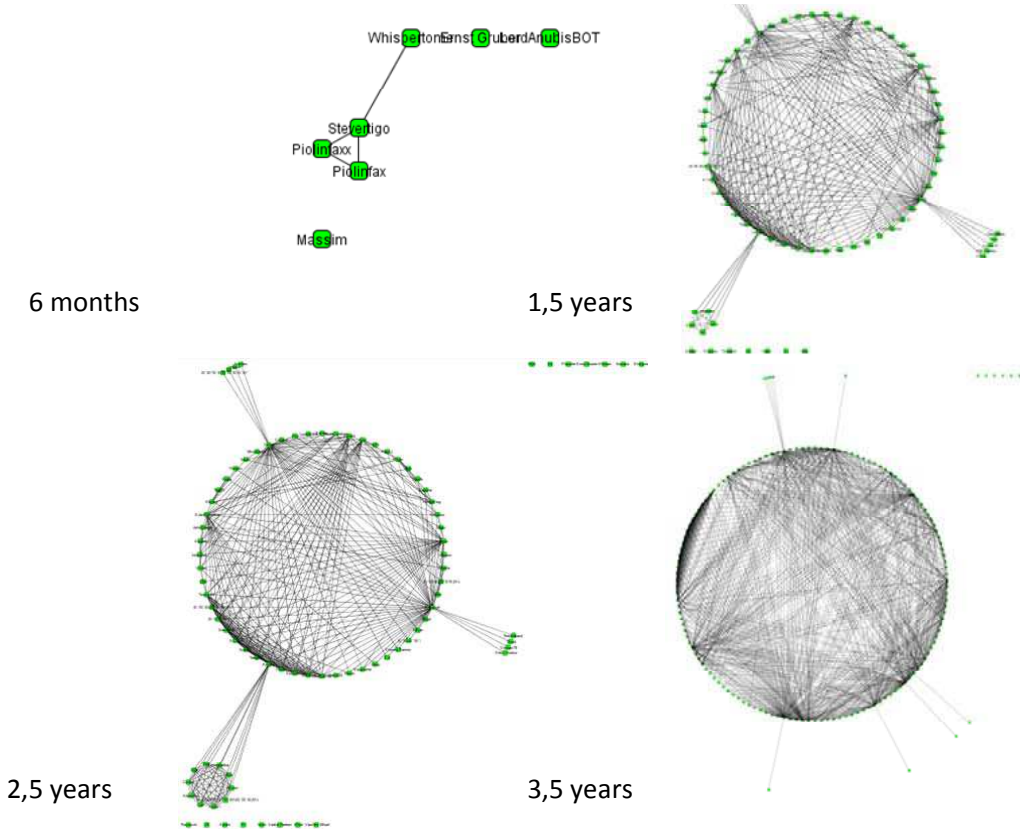


Figure 7.4: Evolution of a Wikipedia author network of registered contributors

contributors form isolated groups since the very beginning of the observation. The groups are growing but most of them stay isolated from each other. Considering limitations of the study, isolated groups of nodes can represent one user that logged in under different Internet Protocol (IP) addresses and therefore the user is presented by many nodes that edit the same articles. Such an investigation can be used for identification of unique anonymous users. We can not regretfully make any conclusions from the network of anonymous users: even if isolated groups consist of nodes that represent only one unregistered user we can not assume that the unregistered user had not representative nodes in other isolated groups. Therefore, this network is not dense since many sets of nodes are representatives of unregistered users respectively.

Later we construct a network with both anonymous and registered users (Figure 7.6). Most nodes belong to a strongly connected component and a minor amount of nodes are isolated or appear in isolated groups. Registered users become bridges that connect a network of anonymous users as Klamma and Haasler (2008b) defined for small Wiki projects. Though registered users in the Greek and Catalan Wikipedia behave differently.

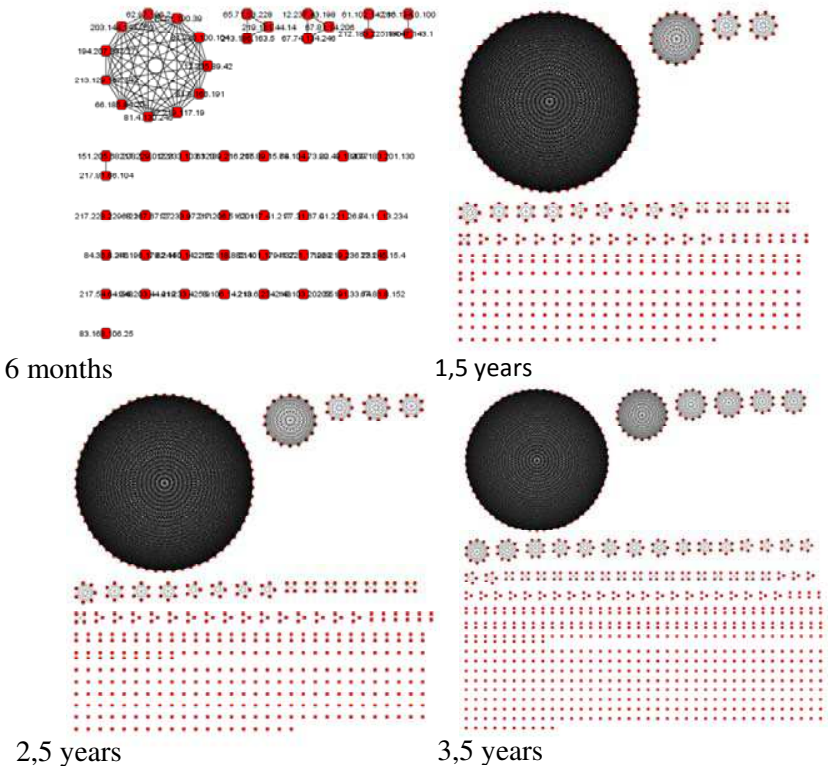


Figure 7.5: Evolution of a Wikipedia network of anonymous contributors

The registered authors of the Catalan Wikipedia have been forming a network with one large and small groups unrelated to each other. It seems that isolated groups consist of authors that are interested in particular topics. Ribé and Rodríguez (2011) defined that many Catalan Wikipedians miss to refer to any other article from the Catalan Wikipedia while Catalan Wikipedians operate only with a closed set of articles. The references to other articles could have evoked an interest of contributors to other kind of articles while the absence of the references can be a reason for a low density of the Catalan Wikipedia author network comparing with other networks.

The network of registered authors in the Greek Wikipedia (Figure 7.7) includes many groups isolated from each other. Thus, only some of registered users in the Greek Wikipedia serve as bridges between isolated groups of anonymous users though other Wikipedia author networks are connected (Figure 7.6).

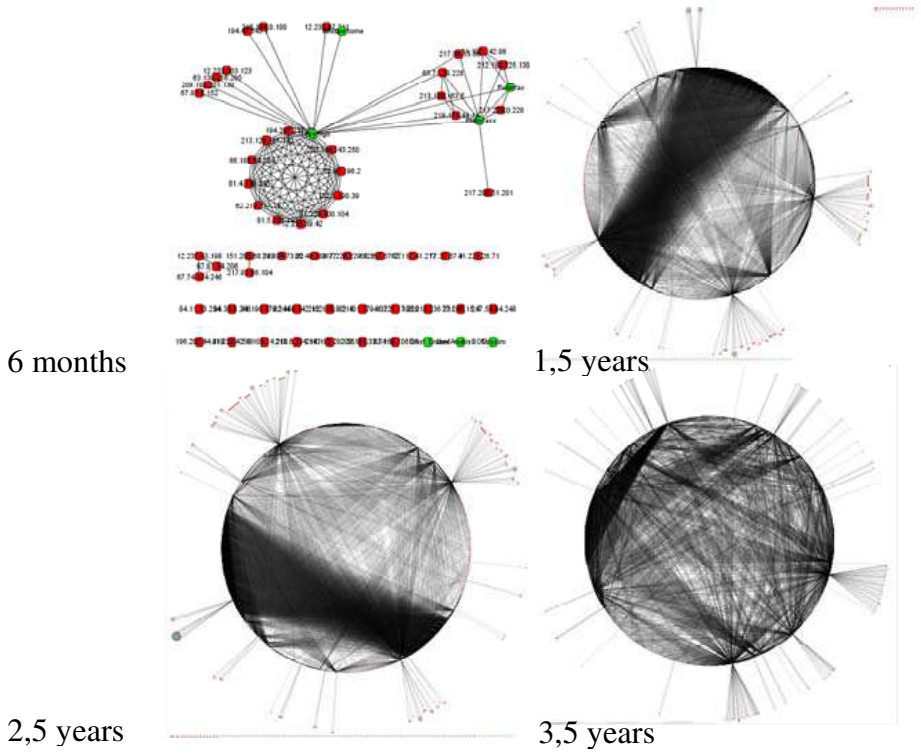


Figure 7.6: Evolution of a Wikipedia network of all contributors

7.2.3 Cultural Perspectives on Wikipedia Author Networks

The Korean and Danish Wikipedia have similar values of the average number of edits although their *power distance* (PD) values are totally different and their cultures support opposite values like *egalitarism* (Danish) and *hierarchy* (Korean). Zlatic et al. (2006) showed that the Korean Wikipedia has one of the highest clustering coefficient in article networks comparing to other 29 Wikipedia instances examined in their study. Korean Wikipedians editing many articles make them connected to other articles so that an article network consists of a number of tightly connected groups of articles. Therefore, we assume that an author network in the Korean Wikipedia have a high clustering coefficient as well as the authors are connected through articles they collaboratively edited[9]. It can be a reason why the Korean Wikipedia with smaller number of registered contributors has similar in the average number of edits per article with the

[9]the authors refer to other articles that were edited by them as well thus tightly connected groups in the article network cause tightly connected groups in the author network or vice versa

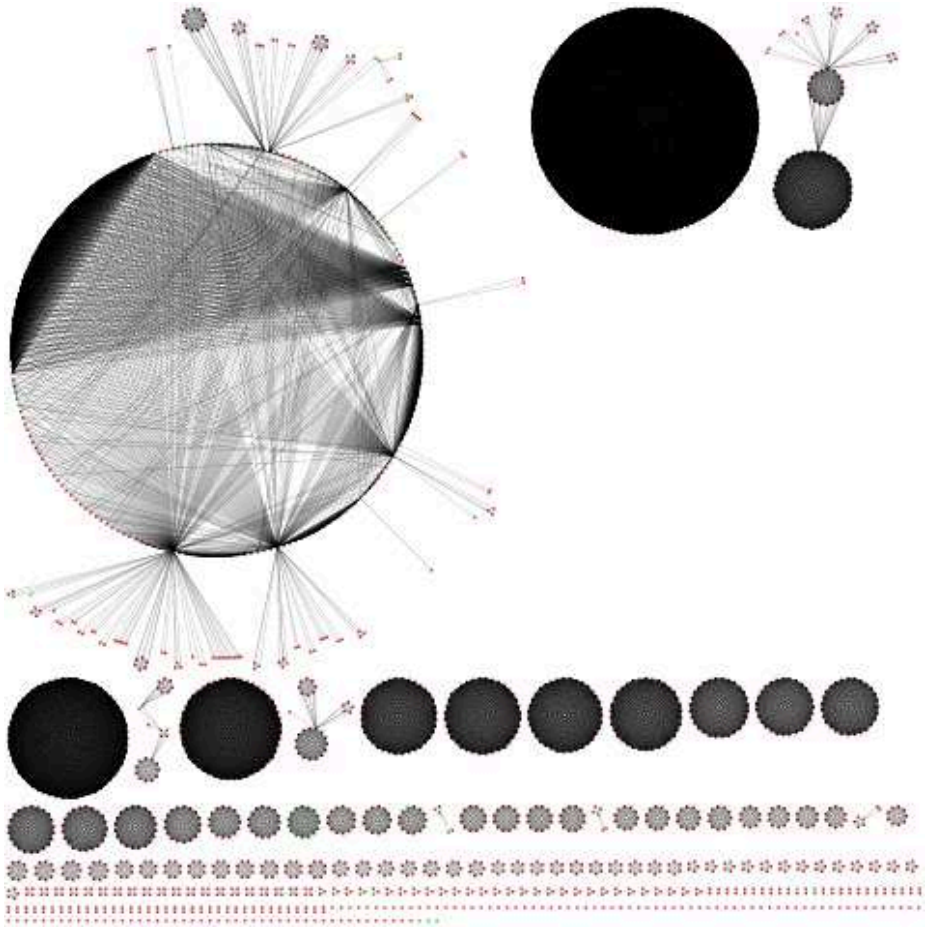


Figure 7.7: The network of all authors in the Greek Wikipedia after nearly 3,5 years

Danish Wikipedia.

We compare the Bulgarian and Korean Wikipedia as they have a similar size of registered users (Table 7.1). The Bulgarian Wikipedia has the highest average number of edits per article. According to Schwartz (2008) both cultures has a high respect for *hierarchy* and *embeddedness* though Zlatic et al. (2006) detected peculiarities of Bulgarian Wikipedia that make the Wikipedia instance exceptional, e.g. the directed article network in the Wikipedia is highly disassortiative, i.e., nodes of different degrees are connected with a high probability.

7.3 Implications for Culturally Sensitive Collaborative Technologies

The first attempts from McLoughlin and Oliver (2000) and Gunawardena et al. (2003) provided rules for constructing distance learning courses depending on the culture of learners. In this study we detect differences of Wikipedians that reveal peculiarities about behavior of culture representatives in social media. These peculiarities provide insights to changes community stakeholders can realize in community environments to make them more culture-sensitive.

Danish and Korean Wikipedia instances are similar in the number of users, number of revisions and average number of edits per article though their cultural values and dimensions are polar (Hofstede, 1991; Schwartz, 2012). Due to Wikipedia policies, the Koreans who usually have a respectful attitude to authorities, feel more comfortable in Wikipedia environment where everybody can contribute and most of contributors have same roles. Although the Korean (students) appreciate working in groups (Uzuner, 2009), they communicate between groups so that a dense web of connections between Wikipedia authors and articles is created(Zlatic et al., 2006). Therefore, we assume that countries with collectivistic cultures like in Korea benefit from Wikipedia since this medium weaken the meaning of authorities to Korean.

Similarities for the values of the Danish and Korean Wikipedia may show that the Danish Wikipedians who value intellectual autonomy (Schwartz, 2008) can be more active in Wikipedia since Western countries representatives are usually more critical in articles than Eastern countries representatives (Hara et al., 2010). Danish Wikipedians can benefit from the system of awards and roles based on the quality of contributions important for the Danish and not on the number of contributions – criticized in many crowdsourcing websites, e.g. stackoverflow [10].

The same change can be applied for Greek Wikipedians to enlarge the amount of brokers, users that connect isolated groups of authors and therefore possess powerful roles (Burt, 2005). Users that connect different topics and groups should be honored by a special badge visible for others.

Many Wikipedia instances corresponding to Slavic countries do not support the Neutral Point of View[11] that can be achieved if several editors contribute qualitative information. The Ukrainian and Macedonian Wikipedians edit rarely existing articles while more often create new ones. The Russian Wikipedia has a higher number of edits per article in average but anyway the number is still low comparing to other Wikipedia instances with similar number of pages such as the Spanish or the Japanese. Tendency of secrecy (Borker, 2012) or a need for a direct instruction (Schwartz, 2008) can cause

[10]Developer forum stackoverflow `http://stackoverflow.com/`, Last access on 10.02.2015

[11]All encyclopedic content on Wikipedia must be written from a neutral point of view (NPOV) `http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view`, Last access on 27.01.2015

such a behavior of many Slavic Wikipedians. Direct instructions may help Arabic Wikipedians that have even higher respect for the authority than Slavic representatives (Schwartz, 2008) to increase activities in Wikipedia.

Since the Japanese prefer to stay anonymous, Japanese anonymous users are extremely active and perform contributions in 50% of pages in the Wikipedia. The success of the Japanese Wikipedia lies in possibility for contributors to stay anonymous and this fact can be replicated by other media environments to adopt them for Japanese users.

7.4 Summary

In this case study we applied the technologies proposed by the framework technologies for data monitoring and analysis. We hypothesized that differences in cultures influence ways how learners collaborate in social media. For example, collaborators from collectivism-oriented cultures behave differently to collaborators of individualism-oriented cultures. Therefore, we detected learning community needs in social media that include representatives of different cultures.

To validate our methodology in estimating community states and needs with the focus on cultural differences (Hofstede, 1991; Schwartz, 2008) of community peers we took 13 Wikipedia instances, where the peers are coming from different countries and cultures. We monitored Wikipedia instances of large and small size, with representatives from Asia and Europe. We conducted a statistical analysis of data about Wikipedians and their edits and estimated correlations of the measures with cultural values and dimensions of Hofstede (1991) and Schwartz (2008). Additionally, we detected the geographical locations of anonymous users that let us infer the geographical locations of registered users. Furthermore, we followed activities of cross-Wikipedia users that contribute to two or more Wikipedia instances from our set. Using social network visualization we followed the evolution of author networks where Wikipedia authors are connected if they maintained the same articles.

According to our observations some results of monitoring and network analysis of Wikipedia networks can not be explained by existing values and dimensions of cultural theories. For example, similarities between Korean and Danish Wikipedia can not be predicted according to Hofstede (1991) and Schwartz (2008). Also, the fact that Bulgarian Wikipedians' activities differ from other Slavic Wikipedians' activities. Based on our investigations together with other experiments we made a number of assumptions for collaborative technology design that can be applied to create culturally-sensitive learning and collaborative technologies. Network analysis helps to explain the similarity between the Danish and Korean behaviors since Koreans manage to create a very tightly connected web between articles and authors (Zlatic et al., 2006). We assumed that the Koreans are more active in Wikipedia than others since they pay less or no attention to authorities in Wikipedia. While Danish Wikipedians need to be approved or motivated to get satisfied according to their cultural values of *intellectual*

autonomy. With this chapter we showed that our methodology can be used to investigate learning activities of different nationalities in social media with the purpose to find differences and similarities of culture representatives in creating and sharing their knowledge.

Chapter 8

Conclusion and Outlook

We have seen It has been recognized that community modeling is a prerequisite for creating and maintaining a successful community information system. Community modeling in contrast to user modeling provides a more systematic vision of communities, their states, their triggers to evolution, their agents and roles and many other things.

The main contributions of this work are the design, realization and validation of the process of community model creation. Its design is precisely described in Chapter 3. We introduced four phases of the process: *modeling*, *refinement*, *monitoring*, and *analysis*. Each phase is described precisely with methods and technologies that are applied for phase realization. The general and specific models of learning communities define in the chapter can be used to avoid cold start problems while modeling learning communities. The formal representation of a community in the *refinement* phase is one of rare agent-based models of learning communities. The multidimensional model of the Mediabase can serve for future realizations of the Mediabase cube. Furthermore, we described crawlers we used for monitoring that can be examples of other social media crawlers. Methods devoted to the analysis phase respect Community of Practice dimensions thus can be used for the analysis such communities.

In chapter 4 we have presented a data management solution that utilizes the data warehouse technology implementing the Mediabase Cube model that provides application-independent and cross-media views on data. The proposed solution can be applied for maintaining data from any social media community. Our solution is the only cube designed for learning communities in social media that allows to operate with the data independently from its source.

In chapter 5 we modeled learning communities in language learning forums. For this purpose we described services that represent all phases of the process of community model creation. We realized and adapted community detection and evolution algorithms to define community agents, that are learners and communities. Network analysis of communities helped to define roles of the learners while analysis of community texts mined for community intents, topics and emotions. The results of these community data analyses were stored according to the Mediabase model and exploited

to create community models. The models were used as inputs for multi-agent simulations that let us validate our formal representation of communities and their usage in supporting community evolution.

Usage of community data analysis is much broader than only creating community models. Since many self-regulated learners require support in directing their learning processes we utilized results of the analysis for competence management in Chapter 6. Implementing CAfe, the competence analyst for eTwinning, we provided an application that supports eTwinning community stakeholders in their life long learning. We informed teachers, the users of eTwinning, about their activities, activities of their peers and development of their own and their peers' competences. Managers of eTwinning can investigate eTwinning networks to define success or failure factors of competence development in eTwinning communities.

Finally, the third application of this work deals with the investigation of community needs that we discovered using community data analysis in Chapter 7. Based on activities of Wikipedia contributors and using our monitoring and analysis services we built networks of Wikipedians registered in different Wikipedia instances. Our data set consisted of instances that belong to different cultures. Therefore, investigating activities of the Wikipedia contributors and their community evolution we specified a number of community requirements for community information system designs that is sensitive to cultural differences.

8.1 Conclusions

Our process of community model creation has proven useful in various ways. The data management solution that was developed for *monitoring* allows to maintain heterogeneous data sources without adapting queries and applications.

Using an information modeling approach we provided a modeling solution that proposes information about early requirements of communities that can be useful for CIS developers. Such a solution fills the gap between customers of CIS and their developers that can benefit from investigating community models before they maintain CIS.

This work has an interdisciplinary character since it integrates views of learning theories on learning communities and analysis of learning community data captured in social media. The methods we applied for analyzing communities consider dimensions of learning theories. Using outcomes of our services learners can analyze their activities and reflect according to their achievements as well as compare their achievements with achievements of others. Other stakeholders such as community managers and administrators can find successful and failure-prone communities based on community data analysis and community models. Moreover, by simulation of community models, the stakeholders can identify steps that trigger successful community development.

In summary, this work has shown that community modeling provides objective information about communities that is relevant for all community stakeholders. The

process of community model creation discovers community states and needs that are relevant for community stakeholders' tasks. Focusing on learning communities, this work proposed a solution that supports communities of self-regulated learners by providing information that helps to enhance learning processes and lifelong learning competencies. Moreover, this solution can assist community managers with the required information for supporting community stability and community developers with early requirements of communities.

8.2 Outlook

Although this dissertation answers the research questions as given in the Chapter 1 it causes a set of new questions that motivate further investigations.

8.2.1 Extension of a General Community Model and Specific Models

The general model of learning communities relies on Community of Practice dimensions (Wenger, 1998) and transcriptivity theory (Jäger et al., 2008). However, the model can be extended with other important aspects of network learning theories such as cognitive processes and cognitive outcomes from (Dillenbourg, 1999) or self-regulated learning phases (Nussbaumer et al., 2011). These aspects will indicate cognitive work in communities and phases of learners. Furthermore, the aspects demand additional analysis of community content that detects cognitive processes and outcomes in community texts. And they require *i** models to extend these to strategic rationale models. Such models focus on decisions and strategies of actors that are taken internally and they can improve decisions about user and community strategies required for simulations.

In this work three community models were presented. Nevertheless, other communities exist that can not be mapped with provided models. These are communities of interest or hobby communities, teacher-student communities, and many others. Therefore, the set of community models can be extended with new types. To make classification of communities easy and faultless multiclass classifiers can be trained to identify types of communities.

8.2.2 Community Simulation Using Additional Community Information

Simulations of models are performed considering reciprocity and preferential attachment strategies that influence the process of community evolution (Schnegg, 2006). However, other factors such as learners' knowledge and their learning histories also contribute to strategies and payoffs of learners. Furthermore, changes in the environment like new learning resources or new learning partners have to be considered as

well in simulations. Results of such refined simulations should provide more accurate predictions about community evolution.

Simulations can also advise further directions of community development. Group formation, usually solved by simulations, follows the purpose of finding the best combination of community members when their payoffs have the highest values possible. It means that all community members are in a win-win situation or close to it. The problem of the win-win situation (Nash equilibrium) has a complex solution (Daskalakis et al., 2006) therefore it is a challenge to find the possible combination of community members that satisfies all community members.

8.2.3 Expansion of Data Sources

The data management solution we presented here can be used for many other media such as Facebook, Twitter, Moodle, Massive Open Online Courses. The collected data, cleaned and anonymous, can be shared as an open repository of social media that facilitates further research on social media. For example, the collected data could be integrated by entity resolution to detect learners and their profiles in different media.

8.2.4 Near-real Time Realization

Social media communities have short lives. Therefore, fast community modeling can help communities to sustain. It means, that the framework presented in this work needs to be refined with technologies relevant for this purpose. Also some algorithms applied in the work have to be adopted. First of all, data can be collected using data stream frameworks such as Apache Storm[1] or Apache S4[2]. All analysis techniques can be executed using distributed environments such as clouds or GPUs. We have already started to refine algorithms for community detection and evolution which are designed for and executed in the GPU environment. However, efficient execution of information retrieval algorithms can be done on frameworks for a distributed processing of large data sets such as Hadoop[3].

8.2.5 Extension of Methods for Community Analysis

We analyzed the collected data using methods that respect learning theories' dimensions. To respond to mutual engagement requirements (Wenger, 1998) we investigate graphs of user collaborations and calculate Social Network Analysis measures. However, the *mutual engagement* also emphasizes trust between community members that

[1] A free and open source distributed realtime computation system `https://storm.apache.org/`, Last access on 05.03.15

[2] A distributed stream computing platform `http://incubator.apache.org/s4/`, Last access on 05.03.2015

[3] The Apache Hadoop software library `http://hadoop.apache.org/`, Last access on 05.03.2015

can be considered by constructing trust networks from user collaborations. Furthermore, we detected patterns of learners based on clustering of Social Network Analysis measures. These can be refined by further machine learning algorithms and by adding other learner characteristics. To implement requirements of the *joint enterprises* dimension (Wenger, 1998) we analyzed intents based on syntactical language patterns. However, extraction of goals can be refined by other approaches coming from linguistics and information retrieval. Furthermore, it is important to extract not just a goal but a learning goal. To reply to shared repertoire's requirements (Wenger, 1998) we defined topics and concepts of communities and performed emotional analysis of community texts. Though further analyses can be executed such as classification of discussions types (Ferguson et al., 2013), categorization of texts and entities based on Linked Open Data repositories, topic mining using Latent Dirichlet Allocation and dynamic language models. All these extensions can improve the accuracy of community models.

In this work we detected communities and their evolution using algorithms discovering connected groups of nodes in graphs. The graphs are built according to interactions between users. However, Wenger (1998) emphasized not only interactions but as well other dimensions that can be taken into consideration for new community detection and evolution algorithms.

Bibliography

Creating effective teaching and learning environments: First results from TALIS. Organization for Economic Co-operation and Development, Paris, 2009. ISBN 9789264056053.

Fabian Abel, Dominikus Heckmann, Eelco Herder, Jan Hidders, Geert-Jan Houben, Erwin Leonardi, and van der Sluijs, Kees. Definition of an appropriate User profile format: GRAPPLE Deliverable 2.1 Version: 1.0, 2009. URL http://www.kbs.uni-hannover.de/Lehre/pers12/_source/04_user_modeling_frameworks/papers/D2.1-WP2-UserProfileFormat-v1.0.doc.

Fabian Abel, Ilknur Celik, Claudia Hauff, Laura Hollink, and Geert-Jan Houben. U-Sem: Semantic Enrichment, User Modeling and Mining of Usage Data on the Social Web. In *Proceedings of 1st International Workshop on Usage Analysis and the Web of Data at the 20th WWW Conference*, Hyderabad, India, March 28th, 2011. URL http://arxiv.org/abs/1104.0126.

Brad Adelberg. NoDoSE - a Tool for Semi-automatically Extracting Structured and Semistructured Data from Text Documents. *SIGMOD Rec*, 27(2):283–294, 1998. URL http://doi.acm.org/10.1145/276305.276330.

Toni Ahlqvist, Asta Bäck, Minna Halonen, and Sirkka Heinonen. *Social Media Roadmaps Exploring the futures triggered by social media*. VTT Research Notes 2454. Espoo, Finland, 2008.

Kirsti Ala-Mutka, Yves Punie, and Anusca Ferrari. Review of Learning in Online Networks and Communities. In *Proceedings of 4th European Conference on Technology Enhanced Learning*, Nice, France, September 29-October 2, 2009, pages 350–364. URL http://dx.doi.org/10.1007/978-3-642-04636-0.

Rëka Albert, Hawoong Jeong, and Albert-László Barabási. Diameter of the World-Wide Web. *Nature*, 401(9):130–131, 1999. URL http://www.nature.com/nature/journal/v401/n6749/full/401130a0.html.

Christopher Alexander. *A Pattern Language: Towns, Buildings, Construction (Center for Environmental Structure Series)*. Oxford University Press, New York, 1978. URL http://downlode.org/etext/patterns/.

Enrique Alfonseca, Rosa M. Carro, Estefanía Martín, Alvaro Ortigosa, and Pedro Paredes. The impact of learning styles on student grouping for collaborative learning: a case study. *User Modeling and User-Adapted Interaction*, 16(3-4):377–401, 2006. URL http://dx.doi.org/10.1007/s11257-006-9012-7.

Catarina Almeida, Miguel Goulão, and Joao Araujo. A Systematic Comparison of i* Modelling Tools Based on Syntactic and Well-formedness Rules. In *Proceedings of the 6th International i* Workshop 2013*, Valencia, Spain, June 17-18, 2013, pages 43–48. URL http://ceur-ws.org/Vol-978/paper_8.pdf.

Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Engaging with Massive Online Courses. In *Proceedings of the 23rd International Conference on World Wide Web*, Seoul, Korea, April 7-11, 2014, pages 687–698.

Chris Anderson. *The Long Tail: Why the Future of Business Is Selling Less of More*. Hyperion, 2006.

Chee Siang Ang and Panayiotis Zaphiris. Simulating Social Networks of Online Communities: Simulation as a Method for Sociability Design. In *Proceedings of the International Conference on Human-Computer Interaction*, San Diego, California, USA, 19-24 July, 2009, pages 443–456.

Panagiotis Antoniou and Apostolos Siskos. The Use of Online Journals in a Distance Education Course. In *European Distance and E-Learning Network Annual Conference*, Naples, Italy, 13-16 June, 2007.

Kimberly E. Arnold and Matthew D. Pistilli. Course Signals at Purdue: Using Learning Analytics to Increase Student Success. In *Proceedings of the Second International Conference on Learning Analytics and Knowledge*, Vancouver, British Columbia, Canada, 29 April – 2 May, 2012, pages 267–270. URL http://doi.acm.org/10.1145/2330601.2330666,AddtoCitaviprojectbyDOI.

Sitaram Asur, Srinivasan Parthasarathy, and Duygu Ucar. An Event-based Framework for Characterizing the Evolutionary Behavior of Interaction Graphs. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(4):16:1–16:36, 2009. URL http://doi.acm.org/10.1145/1631162.1631164.

Thomas Aynaud and Jean-Loup Guillaume. Static community detection algorithms for evolving networks. In *Proceedings of International Workshop on Dynamic Networks in conjuction with Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, Avignon, France, 2010, pages 508–514. URL http://merkur.informatik.rwth-aachen.de/bscw/bscw.cgi/d3504129/%5bAyGu10%5dStatic%20Community%20Detection%20Algorithms%20For%20Evolving%20Networks.pdf.

Albert Bandura. *Social learning theory*. General Learning Press, New York, 1971.

Albert Bandura. *Social foundations of thought and action*. Englewood Cliffs, NJ Prentice Hall, 1986.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *J. Stat. Mech.*, (P10008), 2008. URL `http://arxiv.org/pdf/0803.0476`.

Alex. Borgida and John Mylopoulos. A Sophisticate's Guide to Information Modeling. In *Metamodeling for Method Engineering*, pages 1–42. 2009.

David Borker. Accounting, Culture And Emerging Economies: IFRS in Central and Eastern Europe. *International Business and Economics Research Journal*, 11(9): 1003–1018, 2012.

David Boud, Rosemary Keogh, and David Walker. Promoting Reflection in Learning: a Model. In *Reflection: Turning Experience into Learning*, pages 18–40. 1985.

Ruth Breuer, Ralf Klamma, Yiwei Cao, and Riina Vuorikari. Social Network Analysis of 45,000 Schools: A Case Study of Technology Enhanced Learning in Europe. In *In Proceedings of 4th European Conference on Technology Enhanced Learning*, Nice, France, 29 September - 2 October, 2009, pages 166–180.

Reva Berman Brown and Sean McCartney. Competence Is Not Enough: Meta-Competence and Accounting Education. *Accounting Education*, 4(1):43–53, 1995.

Peter Brusilovsky. Adaptive Hypermedia. *User Modeling and User-Adapted Interaction*, 11(1):87–110, 2001.

Ronald S. Burt. *Structural Holes: The Social Structure of Competition*. Havard Business Press, Cambridge, MA, 1992.

Ronald S. Burt. Structural Holes and Good Ideas. *The American Journal of Sociology*, 110(2):349–399, 2004. URL `http://dx.doi.org/10.2307/3568221`.

Ronald S. Burt. *Brokerage and Closure: An Introduction to Social Capital*. Oxford University Press, 2005.

Rafael A. Calvo and Sidney D'Mello. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *Affective Computing, IEEE Transactions on*, 1(1):18–37, 2010.

Charles Carceller, Shane Dawson, and Lori Lockyer. Improving academic outcomes: does participating in online discussion forums payoff? *International Journal of Technology Enhanced Learning*, 5(2):117–132, 2013.

Carlos Cares, Xavier Franch, Anna Perini, and Angelo Susi. Towards Interoperability of i* Models Using iStarML. *Comput. Stand. Interfaces*, 33(1):69–79, 2011. URL http://dx.doi.org/10.1016/j.csi.2010.03.005.

Tanmoy Chakraborty, Sriram Srinivasan, Niloy Ganguly, Sanjukta Bhowmick, and Animesh Mukherjee. Constant communities in complex networks. *Scientific reports*, 3:1825, 2013.

Yu-Chuan Joni Chao and Hao-Chang Lo. Students' perceptions of Wiki-based collaborative writing for learners of English as a foreign language. *Interactive Learning Environments*, 19(4):395–411, 2011.

Surajit Chaudhuri, Umeshwar Dayal, and Venkatesh Ganti. Database technology for decision support systems. *Computer*, 34(12):48–55, 2001.

Graham Cheetham and Geoff Chivers. *Professions, Competence and Informal Learning*. Edward Elgar Publishing, Northampton, 2005.

Elayne Coakes and Peter Smith. Developing communities of innovation by identifying innovation champions. *The Learning Organization*, 14(1):74–85, 2007. URL http://dx.doi.org/10.1108/09696470710718366.

Jay Cross. *Informal Learning*. Pfeiffer, 2007.

Fabiano Dalpiaz and Jennifer Horkoff, editors. *Proceedings of the Seventh International i* Workshop co-located with the 26th International Conference on Advanced Information Systems Engineering (CAiSE 2014)*, CEUR 1157, 2014. Springer.

Mathieu d'Aquin. Linked Data for Open and Distance Learning, 2012. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.386.1720&rep=rep1&type=pdf.

Mathieu d'Aquin and Nicolas Jay. Interpreting data mining results with linked data for learning analytics. In *the Third International Conference*, Leuven, Belgium, 8-12 April, 2013, pages 155–164.

Mihai Dascalu, Traian Rebedea, and Stefan Trausan-Matu. A Deep Insight in Chat Analysis: Collaboration, Evolution and Evaluation, Summarization and Search. In *Artificial Intelligence: Methodology, Systems, and Applications*, pages 191–200. 2010. URL http://dx.doi.org/10.1007/978-3-642-15431-7.

Constantinos Daskalakis, Paul W. Goldberg, and Christos H. Papadimitriou. The Complexity of Computing a Nash Equilibrium. In *Proceedings of the Thirty-eighth Annual ACM Symposium on Theory of Computing*, Seattle, USA, May 21–23, 2006, pages 71–78. URL http://doi.acm.org/10.1145/1132516.1132527.

Nuno David. Validation and Verification in Social Simulation: Patterns and Clarification of Terminology. In *Epistemological Aspects of Computer Simulation in the Social Sciences*, volume 5466, pages 117–129. Springer Berlin Heidelberg, 2009. ISBN 978-3-642-01108-5.

Jo Davies and Martin Graff. Performance in e-learning: online participation and student grades. *British Journal of Educational Technology*, 36(4):657–663, 2005. URL `http://dx.doi.org/10.1111/j.1467-8535.2005.00542.x`.

Shane Dawson. 'Seeing' the learning community: An exploration of the development of a resource for monitoring online student networking. *British Journal of Educational Technology*, 41(5):736–752, 2010.

Michael Derntl and Ralf Klamma. The European TEL Projects Community from a Social Network Analysis Perspective. In *21st Century Learning for 21st Century Skills*, pages 51–64. 2012.

Michael Derntl, Susanne Neumann, and Petra Oberhuemer. Lost in Interaction in IMS Learning Design Runtime Environments. *Educational Technology & Society*, 17(3): 332–342, 2014.

Louis Deslauriers, Ellen Schelew, and Carl Wieman. Improved Learning in a Large-Enrollment Physics Class. *Science*, 332(6031):862–864, 2011.

Pierre Dillenbourg. What do you mean by collaborative learning? In *Collaborative-learning: Cognitive and Computational Approaches*, pages 1–19. 1999.

Robert J. Elliot, Lakhdar Aggoun, and John B. Moore. *Hidden Markov Models*. Springer-Verlag, New York, 1995.

Yrjö Engeström. *Learning by Expanding*. Orienta-Konsultit Oy, 1987.

European Parlament and the Council. Recommendation od the European Parliament nad the Council on key competences for lifelong learning., 2006.

Kayvon Fatahalian, Jeremy Sugerman, and Pat Hanrahan. Understanding the efficiency of GPU algorithms for matrix-matrix multiplication. In *the ACM SIGGRAPH/EUROGRAPHICS conference*, Sarajevo, Bosnia-Herzegovina, June 20-21, 2004, pages 133–137.

Rebecca Ferguson, Zhongyu Wei, Yulan He, and Simon Buckingham Shum. An evaluation of learning analytics to identify exploratory dialogue in online discussions. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, Leuven, Belgium, April 8-12, 2013, 2013, pages 85–93. URL `http://doi.acm.org/10.1145/2460296.2460313`.

Tiago Lopes Ferreira and Alberto Rodrigues Silva. Foster an Implicit Community Based on a Newsletter Tracking System. In *Proceedings On the Move to Meaningful Internet Systems*, Rome, Italy, 2012, pages 398–415. URL http://dx.doi.org/10.1007/978-3-642-33606-5.

Roy Thomas Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, 2000.

Danyel Fisher, Marc Smith, and Howard T. Welser. You Are Who You Talk To: Detecting Roles in Usenet Newsgroups. In *Proceedings of the 39th Hawaii International Conference on System Sciences*, Kauai, HI, 4-7 January, 2006, page 59b. URL http://www.hicss.hawaii.edu/Hicss39/bestpapers.htm.

Beatriz Florian, Christian Glahn, Hendrik Drachsler, Marcus Specht, and Ramón Fabregat Gesa. Activity-Based Learner-Models for Learner Monitoring and Recommendations in Moodle. In *Proceedings of 6th European Conference on Technology Enhanced Learning*, Palermo, Italy, September 20-23, 2011, pages 111–124. URL http://dx.doi.org/10.1007/978-3-642-23985-4_10.

Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40:35–41, 1977.

Linton C. Freeman. Centrality in Social Networks. Conceptual Clarification. *Social Networks*, 1:215–239, 1978/79.

Paul Fugelstad, Patrick Dwyer, Jennifer Filson Moses, John Kim, Cleila Anna Mannino, Loren Terveen, and Mark Snyder. What makes users rate (share, tag, edit...)?: predicting patterns of participation in online communities. In *Proceedings of the 2012 ACM conference on Computer Supported Cooperative Work*, Seattle. Washington, February 11-15, 2012, pages 969–978.

Günter Gans, Dominik Schmitz, Thomas Arzdorf, Matthias Jarke, and Gerhard Lakemeyer. SNet Reloaded: Roles, Monitoring and Agent Evolution. In *Proceedings of 6th Workshop on Agent-Oriented Information Systems*, New York, July, 2004, pages 2–16.

John Gantz and David Reinsel. The Digital Universe Decade – Are You Ready?, EMC Corporation, 2010. Retrieved from http://www.emc.com/collateral/analyst-reports/idc-digital-universe-are-you-ready.pdf.

James J. Gibson. The Theory of Affordances. In *Perceiving, acting, and knowing*, pages 127–141. 1977.

Nigel Gilbert and Klaus G. Troitzsch. *Simulation for the Social Scientist*. Open University Press, second edition, 2005.

Carlo Giovanella, Mihai Dascalu, and Federico Scaccia. Smart City Analytics: state of the art and future perspectives. *Interaction Design and Architecture(s)*, 20:72–87, 2014.

Christian Glahn, Marcus Specht, and Rob Koper. Using tag-clouds for supporting reflection in self-organised learning. *International Journal of Technology Enhanced Learning*, 3(1):61–79, 2011.

Bogdan Gliwa, Stanisław. Saganowski, Anna. Zygmunt, Piotr Bródka, Przemysław Kazienko, and Jarosław Kozlak. Identification of Group Changes in Blogosphere. In *Proceedings of International Conference on Advances in Social Networks Analysis and Mining*, Istanbul, Turkey, 26-29 August, 2012, pages 1201–1206.

Sten Govaerts, Katrien Verbert, Daniel Dahrendorf, Carsten Ullrich, Manuel Schmidt, Michael Werkle, Arunangsu Chatterjee, Alexander Nussbaumer, Dominik Renzel, Maren Scheffel, Martin Friedrich, Jose Luis Santos, Erik Duval, and Effie L.-C Law. Towards responsive open learning environments: the ROLE interoperability framework. In *Proceedings of the 6th European conference on Technology enhanced learning: towards ubiquitous learning*, Palermo, Italy, September 20-23, 2011, pages 125–138. URL `http://dl.acm.org/citation.cfm?id=2045445.2045458`.

Mark S. Granovetter. The Strength of Weak Ties. *American Journal of Sociology*, 78: 1360 – 1380, 1973.

Gemma Grau, Xavier Franch, and Neil A.M. Maiden. A Goal-Based Round-Trip Method for System Development. In *Proceedings of the 11th International Conference on Requirements Engineering: Foundations for Software Quality*, Porto, Portugal, June 13-14, 2005, pages 71–86.

Ralph Grishman and Beth Sundheim. Message Understanding Conference-6: A Brief History. In *Proceedings of the 16th Conference on Computational Linguistics*, 1996, pages 466–471. URL `http://dx.doi.org/10.3115/992628.992709`.

Charlotte. N. Gunawardena, Penne L. Wilson, and Ana C. Nolla. Culture and online education. In *Handbook of distance learning*, pages 753–775. 2003.

Edward T. Hall. *Beyond culture.* Doubleday, Garden City, New York, 1976.

Edward T. Hall. *The dance of life.* Doubleday, Garden City, New York, 1983.

Michael Hammer and Dennis Mc Leod. Database Description with SDM: A Semantic Database Model. *ACM Trans. Database Syst.*, 6(3):351–386, 1981. URL `http://doi.acm.org/10.1145/319587.319588`.

Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann Publishers, second edition, 2006.

Anna Hannemann and Ralf Klamma. Community Dynamics in Open Source Software Projects: Aging and Social Reshaping. In *Proceedings of 9th IFIP WG 2.13 International Conference on Open Source Software: Quality Verification*, Koper-Capodistria, Slovenia, June 25-28, 2013, pages 80–96. URL http://dx.doi.org/10.1007/978-3-642-38928-3.

Noriko Hara, Pnina Shachaf, and Khe Foon Hew. Cross–cultural analysis of the Wikipedia community. *Journal of the American Society of Information Science and Technology*, 61(10):2097–2108, 2010.

Andrew Hilts and Eric Yu. Intentional Modeling of Social Media Design Knowledge for Government-Citizen Communication. In *Analysis of Social Media and Ubiquitous Data*, pages 20–36. 2011. URL http://dx.doi.org/10.1007/978-3-642-23599-3_2.

Geert Hofstede. *Cultures and organizations: Software of the mind.* McGraw Hill, London, 1991.

Luca Iandoli and Giuseppe Zollo. *Organizational cognition and learning: Building systems for the learning organization*. Information Science Pub, Hershey, PA, 2008. ISBN 978-1-59904-313-5.

David Insa, Josep Silva, and Salvador Tamarit. Using the words/leafs ratio in the DOM tree for content extraction. *The Journal of Logic and Algebraic Programming*, 82 (8):311–325, 2013.

Kenichi Ishii and Morihiro Ogasahara. Links between Real and Virtual Networks: A Comparative Study of Online Communities in Japan and Korea. *Cyberpsy., Behavior, and Soc. Networking*, 10(2):252–257, 2007. URL http://dblp.uni-trier.de/db/journals/cbsn/cbsn10.html#IshiiO07.

Ludwig Jäger. Transkriptivität - Zur medialen Logik der kulturellen Semantik. In *Transkribieren - Medien/Lektüre*, pages 19 – 41. 2002.

Ludwig Jäger, Matthias Jarke, Ralf Klamma, and Marc Spaniol. Transkriptivität: Operative Medientheorien als Grundlage von Informationssystemen für die Kulturwissenschaften. *Informatik Spektrum*, 31(1):21–29, 2008.

Matthias Jarke, John Mylopoulos, Joachim W. Schmidt, and Yannis Vassiliou. DAIDA: An environment for evolving information systems. *ACM Transactions on Information Systems (TOIS)*, 10(1):1–50, 1992.

Matthias Jarke, Maurizio Lenzerini, Yannis Vassiliou, and Panos Vassiliadis, editors. *Fundamentals of Data Warehouses*. Springer, 1999.

Matthias Jarke, Ralf Klamma, Gerhard Lakemeyer, and Dominik Schmitz. Continuous, Requirements-Driven Support for Organizations, Networks, and Communities. In *Proceedings of the 3rd International i* Workshop*, Recife, Brazil, February 11-12, 2008, pages 47–50. URL `http://ceur-ws.org/Vol-322/paper12.pdf`.

Manfred A. Jeusfeld, Matthias Jarke, and John Mylopoulos, editors. *Metamodeling for Method Engineering*. MIT Press, 2009.

Larry Johnson, Adams Samantha, and Malcolm Cummins. The NMC Horizon Report: 2012 Higher Education Edition, 2012.

Roula Karam, Piero Fraternali, Alessandro Bozzon, and Luca Galli. Modeling End-Users as Contributors in Human Computation Applications. In *Proceedings of Model and Data Engineering*, Poitiers, France, 3-5 October, 2012, pages 3–15. URL `http://dx.doi.org/10.1007/978-3-642-33609-6_3`.

Ron S. Kenett, Xavier Franch, Angelo Susi, and Nikolas Galanis. Adoption of Free Libre Open Source Software (FLOSS): A Risk Management Perspective. In *Proceedings of IEEE International Computers, Software, and Applications Conference*, Västerås, Sweden, July 21-25, 2014, pages 171–180.

Greg Kessler and Dawn Bikowski. Developing collaborative autonomous learning abilities in computer mediated language learning: attention to meaning among students in wiki space. *Computer Assisted Language Learning*, 23(1):41–58, 2010.

Joachim Kimmerle, Johannes Moskaliuk, and Ulrike Cress. Understanding learning: the Wiki way. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, Orlando, Florida, October 25-27, 2009, pages 3:1–3:8.

Anastasia Kitsantas. Test Preparation and Test Performance: A Self-Regulatory Analysis. *Journal of Experimental Education*, 70(2):101–113, 2002.

Anastasia Kitsantas and Nada Dabbagh. Promoting Self-Regulation in Distributed Learning Environments with Web-Based Pedagogical Tools: An Exploratory Study. *Journal of Excellence in College Teaching*, 15:119–142, 2004.

René F. Kizilcec, Chris Piech, and Emily Schneider. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, Leuven, Belgium, April 8-12, 2013, pages 170–179. URL `http://doi.acm.org/10.1145/2460296.2460330`.

Ralf Klamma. Werkzeuge und Modelle für die übergreifende Untersuchung von Social Software. *i-com*, 9(3):12–20, 2010.

Ralf Klamma. Community Learning Analytics – Challenges and Opportunities. In *Proceedings of International Conference of Web-based Learning*, Kenting, Taiwan, October 6-9, 2013, pages 284–293.

Ralf Klamma and Christian Haasler. Dynamic Network Analysis of Wikis. In *Proceedings of I-KNOW '08 and I-MEDIA '08*, Graz, Austria, October 21-22, 2008a, pages 276–279.

Ralf Klamma and Christian Haasler. Wikis as Social Networks: Evolution and Dynamics. In *Proceedings of 2nd SNA-KDD Workshop Social Network Mining and Analysis in conjuction with International Conference on Knowledge Discovery and Data Mining*, Las Vegas, NV, USA, August 24-27, 2008b.

Ralf Klamma and Zinayida Petrushyna. The Troll Under the Bridge: Data Management for Huge Web Science Mediabases. In *Proceedings of the 38. Jahrestagung der Gesellschaft für Informatik e.V. (GI), die INFORMATIK*, Munich, Germany, September 8-13, 2008, pages 923–928.

Ralf Klamma and Zinayida Petrushyna. Pattern-based competence management: On the gap between intentions and reality. In *Proceeding of 11th IFIP WG 5.5 on Virtual Enterprises*, Saint-Etienne, France, 11-13 October, 2010, pages 364–371.

Ralf Klamma, Marc Spaniol, and Matthias Jarke. "Do you know a similar project I can learn from?" Self-monitoring of Communities of Practice in the Cultural Sciences. In *Proceedings of the 3rd International Conference on Information Technology and Applications ICITA'05, Sydney, Australia, July 4-7, 2005, Volume II*, 2005, pages 608–613. URL `http://www-i5.informatik.rwth-aachen.de/lehrstuhl/staff/klamma/download/klammar_self-monitoring336.pdf`.

Ralf Klamma, Marc Spaniol, and Yiwei Cao. MPEG-7 Compliant Community Hosting. *MPEG and Multimedia Metadata Community Workshop Results 2005, J.UKM Special Issue (Journal of Universal Knowledge Management)*, 1(1):36–44, 2006a.

Ralf Klamma, Marc Spaniol, Yiwei Cao, and Matthias Jarke. Pattern-Based Cross Media Social Network Analysis for Technology Enhanced Learning in Europe. In *Proceedings of the First European Conference on Technology Enhanced Learning*, Crete, Greece, October 3-5, 2006b, pages 242–256. URL `http://www.springerlink.com/content/d110425q358k4750/`.

Ralf Klamma, Marc Spaniol, and Dimitar Denev. PALADIN: A Pattern Based Approach to Knowledge Discovery in Digital Social Networks. In *Proceedings of I-KNOW '06, 6th International Conference on Knowledge Management*, Graz, Austria, September 6 - 8, 2006c, pages 457–464. URL `http://www-i5.informatik.rwth-aachen.de/lehrstuhl/staff/klamma/download/KSDe06.pdf`.

Ralf Klamma, Yiwei Cao, and Marc Spaniol. Watching the Blogosphere: Knowledge Sharing in the Web 2.0. In *Proceedings of the 1st International Conference on Weblogs and Social Media*, Boulder, Colorado, USA, March 26-28, 2007, pages 105 – 112.

Styliani Kleanthous and Vania Dimitrova. Semantic enhanced approach for modelling cognitive relationships in virtual communities. In *Proceedings of Workshop on Adaptation and Personalisation in Social Systems: Groups, Teams, Communities: Proceedings of SociUM Workshop held at the 11th International Conference on UM*, Corfu, Greece, July 25-29, 2007.

Styliani Kleanthous and Vania Dimitrova. Analyzing Community Knowledge Sharing Behavior. In *Proceedings of User Modeling, Adaptation, and Personalization: Proceedings of UMAP*, Big Island, HI, USA, June 20-24, 2010, pages 231–242. URL `http://dx.doi.org/10.1007/978-3-642-13470-8_22`.

Florence Kluckhohn and Fred L. Strodbeck. *Variation in Value Orientation*. Row and Peterson, Evanson, IL, 1961.

Kenneth R. Koedinger, Kyle Cunningham, Alida Skogsholm, and Brett Leber. An open repository and analysis tools for fine-grained, longitudinal learner data. In *Proceedings of 1st International Conference on Educational Data Mining*, Montréal, Québec, Canada, June 20-21, 2008, pages 157–166.

Julian Alexander Krenge, Zinayida Petrushyna, Milos Kravcik, and Ralf Klamma. Identification of Learning Goals in Forum-based Communities. In *Proceedings of the IEEE 11th International Conference on Advanced Learning Technologies*, Athens, Georgia, USA, 6-8 July, 2011, pages 307–309.

Maria Kyprianidou, Stavros Demetriadis, Thrasyvoulos Tsiatsos, and Andreas Pombortsis. Group formation based on learning styles: can it improve students' teamwork? *Educational Technology Research and Development*, 60(1):83–110, 2012. URL `http://dx.doi.org/10.1007/s11423-011-9215-4`.

Maarten De Laat, Vic Lally, Lasse Lipponen, and Robert-Jan Simons. Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for Social Network Analysis. *International Journal of Computer-Supported Collaborative Learning*, 2(1):87–103, 2007. URL `http://dx.doi.org/10.1007/s11412-007-9006-4`.

Richard J. Larsen and Morris L. Marx. *An Introduction to Mathematical Statistics and Its Applications*. Prentice Hall, 2000.

Bruno Latour. On Recalling ANT. In *Actor Network Theory and after*, pages 15–25. 1999.

Bruno Latour. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press, Oxford, 2005.

Tobias Ley, Barbara Kump, and Cornelia Gerdenitsch. Scaffolding Self-directed Learning with Personalized Learning Goal Recommendations. In *User Modeling, Adaptation, and Personalization*, pages 75–86. 2010. URL http://dx.doi.org/10.1007/978-3-642-13470-8_9.

Xiaochen Li, Wenji Mao, Daniel Zeng, and Fei-Yue Wang. Agent-Based Social Simulation and Modeling in Social Computing. In *Proceedings of the IEEE ISI 2008 PAISI, PACCF, and SOCO International Workshops on Intelligence and Security Informatics*, Taipei, Taiwan, June 17, 2008, pages 401–412.

Cindy X. Lin, Bolin Ding, Jiawei Han, Feida Zhu, and Bo Zhao. Text Cube: Computing IR Measures for Multidimensional Text Database Analysis. In *Proceedings of the Eighth IEEE International Conference on Data Mining*, Pisa, Italy, December 15-19, 2008, pages 905–910.

Pei-Chun Lin, Berlin Wu, and Junzo Watada. Kolmogorov-Smirnov Two Sample Test with Continuous Fuzzy Data. In Van-Nam Huynh, Yoshiteru Nakamori, Jonathan Lawry, and Masahiro Inuiguchi, editors, *Integrated Uncertainty Management and Applications*, volume 68 of *Advances in Intelligent and Soft Computing*, pages 175–186. Springer Berlin Heidelberg, 2010.

Shian-Hua Lin and Jan-Ming Ho. Discovering Informative Content Blocks from Web Documents. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, AB, Canada, July 23 - 25, 2002, pages 588–593. URL http://doi.acm.org/10.1145/775047.775134.

Lasse Lipponen, Marjaana Rahikainen, Jiri Lallimo, and Kai Hakkarainen. Patterns of participation and discourse in elementary students' computer-supported collaborative learning. *Learning and Instruction*, 13(5):487–509, 2003.

Shuangyan Liu, Mike Joy, and Nathan Griffiths. iGLS: Intelligent Grouping for Online Collaborative Learning. In *Proceedings on the Ninth IEEE International Conference on Advanced Learning Technologies*, Riga, Latvia, July 15-17, 2009, pages 364–368.

Charles M. Macal and Michael J. North. Agent-based Modeling and Simulation. In *Proceedings of the Winter Simulation Conference*, Austin, TX, USA, 13-16 December, 2009, pages 86–98.

Leah P. Macfadyen and Shane Dawson. Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, 54(2):588–599, 2010.

Ioannis Magnisalis, Stavros Demetriadis, and Anastasios Karakostas. Adaptive and Intelligent Systems for Collaborative Learning Support: A Review of the Field. *IEEE Transactions on Learning Technologies*, 4(1):5–20, 2011.

Nikos Manouselis, Hendrik Drachsler, Riina Vuorikari, Hans G. K. Hummel, and Rob Koper. Recommender Systems in Technology Enhanced Learning. In *Recommender Systems Handbook*, pages 387–415. 2011.

JoséAntonio Marcos-García, Alejandra Martínez-Monés, Yannis Dimitriadis, Rocío Anguita-Martínez, Inés Ruiz-Requies, and Bartolomé Rubia-Avi. Detecting and Solving Negative Situations in Real CSCL Experiences with a Role-Based Interaction Analysis Approach. In *Intelligent Collaborative e-Learning Systems and Applications*, pages 129–146. 2009. URL http://dx.doi.org/10.1007/978-3-642-04001-6_9.

Roberto Martinez, James R. Wallace, Judy Kay, and Kalina Yacef. Modelling and Identifying Collaborative Situations in a Collocated Multi-display Groupware Setting. In *Artificial Intelligence in Education*, pages 196–204. 2011. URL http://dx.doi.org/10.1007/978-3-642-21869-9_27.

Michael Mayo and Antonija Mitrovic. Optimising ITS behaviour with Bayesian networks and decision theory. *International Journal of Artificial Intelligence in Education*, 12:124–153, 2001.

David C. McClelland. Testing for Competence Rather than for "Intelligence". *American Psychologist*, 20:321–333, 1973.

C. McLoughlin and R. Oliver. Designing learning environments for cultural inclusivity: A case study of indigenous online learning at tertiary level. *Australian Journal of Educational Technology*, 16(1):58–72, 2000.

Fabian Menges, Bud Mishra, and Giuseppe Narzisi. Modeling and simulation of e-mail social networks: A new stochastic agent-based approach. In *Proceedings on the Winter Simulation Conference*, Miami, Florida, USA, December 7-10, 2008, pages 2792–2800. URL http://dblp.uni-trier.de/db/conf/wsc/wsc2008.html#MengesMN08.

Stuart E. Middleton, Nigel R. Shadbolt, and De Roure, David C. Ontological User Profiling in Recommender Systems. *ACM Trans. Inf. Syst.*, 22(1):54–88, 2004. URL http://doi.acm.org/10.1145/963770.963773.

Libby V. Morris, Catherine Finnegan, and Sz-Shyan Wu. Tracking student behavior, persistence, and achievement in online courses. *The Internet and Higher Education*, 8(3):221–231, 2005. URL http://www.sciencedirect.com/science/article/pii/S1096751605000412.

Donn Morrison, Ian McLoughlin, Alice Hogan, and Conor Hayes. Evolutionary clustering and analysis of user behaviour in online forums. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, Dublin, Ireland, June 4–7, 2012. URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4636.

John Mylopoulos. Information modeling in the time of the revolution. *Information Systems*, 23(3-4):127–155, 1998.

Jad Najjar, Michael Derntl, Tomaž Klobucar, Bernd Simon, Michael Totschnig, Simon Grant, and Jan Pawlowski. A Data Model for Describing and Exchanging Personal Achieved Learning Outcomes PALO. *Int. J. IT Stand. Stand. Res.*, 8(2):87–104, 2010. URL http://dx.doi.org/10.4018/jitsr.2010070107.

Keiichi Nemoto and Peter A. Gloor. Analyzing Cultural Differences in Collaborative Innovation Networks by Analyzing Editing Behavior in Different-Language Wikipedias. *Procedia - Social and Behavioral Sciences*, 26(0):180–190, 2011. URL http://www.sciencedirect.com/science/article/pii/S1877042811024013.

Mark E. J. Newman. Who Is the Best Connected Scientist?A Study of Scientific Coauthorship Networks. In *Complex Networks*, pages 337–370. 2004. URL http://dx.doi.org/10.1007/978-3-540-44485-5_16.

Mark E. J. Newman and Michelle Girvan. Finding and evaluating community structure in networks. *PHYSICAL REVIEW E*, 69, 2004.

Thierry Nodenot and Pierre Laforcade. CPM: A UML Profile to Design Cooperative PBL Situations at Didactical Level. In *Proceeding of the Sixth International Conference on Advanced Learning Technologies*, Kerkrade, The Netherlands, 2006, pages 1113–1114.

Alexander Nussbaumer, Dietrich Albert, and Uwe Kirschenmann. Technology-mediated Support for Self-regulated Learning in Open Responsive Learning Environments. In *Proceedings of the 2011 IEEE Global Engineering Education Conference (EDUCON)*, Amman, Jordan, April 4-6, 2011, pages 421–427.

Ifeyinwa Okoye, Tamara Sumner, and Steven Bethard. Automatic extraction of core learning goals and generation of pedagogical sequences through a collection of digital library resources. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, Indiana, USA, July 22-26, 2013, pages 67–76. URL http://doi.acm.org/10.1145/2467696.2467708,AddtoCitaviprojectbyDOI.

SUllivan A. Palinscar. Social Constructivist Perspectives On Teaching and Learning. *Annual Review Psychology*, (49):345–375, 1998.

Stuart Palmer, Dale Holt, and Sharyn Bray. Does the discussion help? The impact of a formally assessed online discussion on final student results. *British Journal of Educational Technology*, 39(5):847–858, 2008. URL http://dx.doi.org/10.1111/j.1467-8535.2007.00780.x.

Seymour Papert. *Mindstorms: Children, computers, and powerful ideas*. Basic Books, Inc, 1980.

Nikolaos Pappas, Georgios Katsimpras, and Efstathios Stamatatos. Extracting Informative Textual Parts from Web Pages Containing User-generated Content. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, Graz, Austria, September 5-7, 2012, pages 4:1–4:8. URL http://doi.acm.org/10.1145/2362456.2362462.

Alexandros Paramythis. Adaptive support for collaborative learning with ims learning design: Are we there yet. In *Proceedings of the Workshop on Adaptive Collaboration Support, held in conjunction with the 5th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, Hannover, Germany, 2008, pages 17–29.

G. Pask and Bernard Scott. Learning strategies and individual competence. *International Journal of Man-Machine Studies*, 4:217–253, 1972.

James W. Pennebaker, Cindy K. Chung, Molly. Ireland, Amy. Gonzales, and Roger J. Booth. The Development and Psychometric Properties of LIWC2007, 2007. URL http://www.liwc.net/LIWC2007LanguageManual.pdf.

Dilhan Perera, Judy Kay, Irena Koprinska, Kalina Yacef, and Osmar R. Zaïane. Clustering and Sequential Pattern Mining of Online Collaborative Learning Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(6):759–772, 2009.

Zinayda Petrushyna, Milos Kravcik, and Ralf Klamma. Learning Analytics for Communities of Lifelong Learners: A Forum Case. In *Proceedings of the IEEE 11th International Conference on Advanced Learning Technologies*, Athens, Georgia, USA, 6-8 July, 2011, pages 609–610. URL http://dx.doi.org/10.1109/ICALT.2011.185.

Zinayda Petrushyna, Ralf Klamma, and Matthias Jarke. The Impact of Culture On Smart Community Technology: The Case of 13 Wikipedia Instances. *Interaction Design & Architecture(s)*, 22:34–47, 2014a.

Zinayida Petrushyna and Ralf Klamma. No Guru, No Method, No Teacher: Self-classification and Self-modelling of E-Learning Communities. In *Proceedings of the Third European Conference on Technology Enhanced Learning*, Maastricht, Netherlands, September 16-19, 2008, pages 354–365.

Zinayida Petrushyna, Ralf Klamma, and Milos Kravcik. Designing During Use: Modeling of Communities of Practice. In *Proceedings of 4th IEEE International Conference on Digital Ecosystems and Technologies*, Dubai, UAE, 13-16 April, 2010, pages 612–617.

Zinayida Petrushyna, Alexander Ruppert, Ralf Klamma, Dominik Renzel, and Matthias Jarke. i*-REST: Light-Weight i* Modeling with RESTful Web Services. In *Proceedings of the Seventh International i* Workshop co-located with the 26th International Conference on Advanced Information Systems Engineering (CAiSE 2014)*, Thessaloniki, Greece, June 16-17, 2014b.

Zinayida Petrushyna, Ralf Klamma, and Milos Kravcik. On Modeling Learning Communities. In *Proceedings of the 10th European Conference on Technology Enhanced Learning*, Toledo, Spain, 15-18 September, 2015, pages 254–267.

Ulrike Pfeil, Panayiotis Zaphiris, and Chee Siang Ang. Cultural Differences in Collaborative Authoring of Wikipedia. *Journal of Computer-Mediated Communication*, 12 (1):88–113, 2006. URL `http://dx.doi.org/10.1111/j.1083-6101.2006.00316.x`.

Manh Cuong Pham and Ralf Klamma. Data Warehousing for Lifelong Learning Analytics. *Bulletin of the Technical Committee on Learning Technology*, 15(2):6–9, 2013. URL `http://lttf.ieee.org/issues/april2013/Pham.pdf`.

Manh Cuong Pham, Yiwei Cao, Zinayida Petrushyna, and Ralf Klamma. Learning Analytics in a Teachers' Social Network. In *Proceedings of the Eighth International Conference on Networked Learning (NLC 2012)*, Maastricht, the Netherlands, April 2-4, 2012, pages 414–421.

Jean Piaget. *The child's conception of the world*. Paladin, St. Albans, Great Britain, 1973.

Jenny Preece. *Online Communities: Designing Usability and Supporting Socialbility*. John Wiley & Sons, Inc, New York, NY and USA, 2000. ISBN 0471805998.

Reihaneh Rabbany k., Mansoureh Takaffoli, and Osmar R. Zaïane. Social Network Analysis and Mining to Support the Assessment of On-line Student Participation. *SIGKDD Explor. Newsl.*, 13(2):20–29, 2012. URL `http://doi.acm.org/10.1145/2207243.2207247`.

Sheizaf Rafaeli, Tsahi Hayat, and Yaron Ariel. Wikipedia Participants and "Ba": Knowledge Building and Motivations. In *Proceedings of the Cyberculture 3rd Global Conference*, Prague, Czech Republic, 2009.

Morten Rask. The Reach and Richness of Wikinomics: Is the free web-based encyclopedia Wikipedia only for rich countries? In *Proceedings of the Joint Conference of*

The International Society of Marketing Development and the Macromarketing Society, Washington, DC, June 2-5, 2007. URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=996158.

Andrew Ravenscroft. From conditioning to learning communities: Implications of fifty years of research in e–learning interaction design. *ALT-J*, 11(3):4–18, 2003.

Christophe Reffay and Marie-Laure Betbeder. Sharing Corpora and Tools to Improve Interaction Analysis. In *Learning in the Synergy of Multiple Disciplines*, pages 196–210. 2009. URL http://dx.doi.org/10.1007/978-3-642-04636-0_20.

Dominik Renzel and Ralf Klamma. From Micro to Macro: Analyzing Activity in the ROLE Sandbox. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, Leuven, Belgium, April 8-12, 2013, pages 250–254. URL http://doi.acm.org/10.1145/2460296.2460347, AddtoCitaviprojectbyDOI.

Dominik Renzel, Ralf Klamma, and Marc Spaniol. MobSOS - A Testbed for Mobile Multimedia Community Services. In *Proceedings of the Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, Klagenfurt, Austria, May 7-9, 2008, pages 139–142.

Günter Daniel Rey. *E-Learning: Theorien, Gestaltungsempfehlungen und Forschung*. Psychologie Lehrbuch. Huber, Bern, 1. aufl edition, 2009. ISBN 978-3-456-84743-6.

Marc M. Ribé and Horacio Rodríguez. Cultural Configuration of Wikipedia: Measuring Autoreferentiality in Different Languages. In *Proceedings of Recent Advances in Natural Language Processing*, Hissar, Bulgaria, September 12-14, 2011, pages 316–322.

Giuseppe Rizzo and Raphaël Troncy. NERD: a framework for unifying named entity recognition and disambiguation extraction tools. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, April 23 - 27, 2012, pages 73–76.

Aristama Roesli, Dominik Schmitz, Gerhard Lakemeyer, and Matthias Jarke. Modelling Actor Evolution in Agent-Based Simulations. In *Organized Adaption in Multi-Agent Systems*, pages 126–144. 2009. URL http://dx.doi.org/10.1007/978-3-642-02377-4_8.

Colette Rolland. Capturing System Intentionality with Maps. In *Conceptual Modelling in Information Systems Engineering*, pages 141–158. 2007. URL http://dx.doi.org/10.1007/978-3-540-72677-7_9.

Gert Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4):581–603, 1966. URL http://dx.doi.org/10.1007/BF02289527.

Maren Scheffel, Katja Niemann, Abelardo Pardo, Derick Leony, Martin Friedrich, Kerstin Schmidt, Martin Wolpers, and Carlos Delgado Kloos. Usage pattern recognition in student activities. In *Proceedings of the 6th European Conference of Technology Enhanced Learning*, Palermo, Italy, September 20-23, 2011, pages 341–355.

Oliver Scheuer and Bruce M. McLaren. Helping Teachers Handle the Flood of Data in Online Student Discussions. In *In Proceedings of Intelligent Tutoring Systems*, Montreal, Canada, June 23-27, 2008, pages 323–332. URL http://dx.doi.org/10.1007/978-3-540-69132-7_36.

Michael Schnegg. Reciprocity and the Emergence of Power Laws in Social Networks. *International Journal of Modern Physics*, 17(8), 2006.

Judith Schoonenboom, Henk Sligte, Ayman Moghnieh, Davinia Hernández-Leo, Krassen Stefanov, Christian Glahn, Marcus Specht, and Ruud Lemmers. Supporting life-long competence development using the TENCompetence infrastructure: a first experiment. *International Journal of Emerging Technologies in Learning (iJET)*, 3, 2008.

Shalom H. Schwartz. *Cultural value orientations: Nature & implications of national differences.* Publ. House of SU HSE, Moscow, 2008.

Shalom H. Schwartz. An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture*, 2(1), 2012.

Lemya Settouti, Nathalie Guin, Vanda Luengo, and Alain Mille. Adaptable and Reusable Query Patterns for Trace-Based Learner Modelling. In *Proceedings of 6th European Conference on Technology Enhanced Learning*, Palermo, Italy, September 20-23, 2011, pages 384–397.

Patrick Siehndel, Ricardo Kawase, Asmelash Teka Hadgu, and Eelco Herder. Finding Relevant Missing References in Learning Courses. In *Proceedings of the 22nd International Conference on World Wide Web Companion*, Seoul, Korea, April 7-11, 2013, pages 425–430. URL http://dl.acm.org/citation.cfm?id=2487788.2487957.

B.F Skinner. The science of learning and the art of teaching. *Harvard Educational Review*, pages 88–97, 1954.

Peter Sloep and Adriana Berlanga. Learning Networks, Networked Learning. *Comunicar*, 19(37):55–64, 2011.

Amy Soller. Supporting social interaction in an intelligent collaborative learning system. *IJAIED*, 12:40–62, 2001.

Amy Soller, Alejandra Martínez, Patrick Jermann, and Martin Muehlenbrock. From mirroring to guiding: A review of state of the art technology for supporting collaborative learning. *International Journal of Artificial Intelligence in Education*, 15(4): 261–290, 2005.

Dandan Song, Fei Sun, and Lejian Liao. A hybrid approach for content extraction with text density and visual importance of DOM nodes. *Knowledge and Information Systems*, 42(1):75–96, 2013.

Ergang Song, Zinayida Petrushyna, Yiwei Cao, and Ralf Klamma. Learning Analytics at Large: The Lifelong Learning Network of 160,000 European Teachers. In *Proceedings of 6th European Conference on Technology Enhanced Learning*, Palermo, Italy, September 20-23, 2011, pages 398–411.

Marc Spaniol and Ralf Klamma. Mediating Ontologies for Communities of Practice. In *Proceedings of Practical Aspects of Knowledge Management*, Vienna, Austria, December 2–3, 2004, pages 330–342.

Marc Spaniol, Ralf Klamma, and Yiwei Cao. Learning as a Service: A Web-based Learning Framework for Communities of Professionals on the Web 2.0. In *Proceedings of the International Conference on Web-based Learning*, Edinburgh, UK, August 15-17, 2007, pages 160–173.

Gerry Stahl. *Group cognition: computer support for building collaborative knowledge*. Acting with technology. MIT Press, 2006. ISBN 9780262195393. URL `http://books.google.de/books?id=6W7uAAAAMAAJ`.

Markus Strohmaier, Peter Prettenhofer, and Mathias Lux. Different Degrees of Explicitness in Intentional Artifacts: Studying User Goals. In *Proceedings on the CSKGOI'08 International Workshop on Commonsense Knowledge and Goal Oriented Interfaces, in conjunction with IUI'08*, Canary Islands, Spain, January 13, January 13. URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.142.8528&rep=rep1&type=pdf`.

Hee-Joen Suh and Seung-Wook Lee. Collaborative Learning Agent for Promoting Group Interaction. *ETRI*, 28(4):461–474, 2006.

Bernardo Tabuenca, Marco Kalz, Stefaan Ternier, and Marcus Specht. Stop and Think: Exploring Mobile Notifications to Foster Reflective Practice on Meta-Learning. *IEEE Transactions on Learning Technologies*, 8(1):124–135, 2015.

Chayant Tantipathananandh, Tanya Berger-Wolf, and David Kempe. A framework for community identification in dynamic social networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, San Jose, CA, USA, August 12 - 15, 2007, pages 717–726. URL `http://doi.acm.org/10.1145/1281192.1281269`.

Marta Tatu. *Discovering Intentions in Text and Semantic Calculus: Intention Overview, Classification, Representation, Discovery and Interactions with Other Semantic Relations*. VDM Verlag, Saarbrücken, Germany, Germany, 2008. ISBN 3639078683.

Fons Trompenaars and Charles Hampden-Turner. *Riding the waves of culture: Understanding cultural diversity in global business*. McGraw-Hill, New York, 1998.

Maksim Tsvetovat and Kathleen M. Carley. Modeling Complex Socio-technical Systems using Multi-Agent Simulation Methods. *Künstliche Intelligenz (Artificial Intelligence Journal)*, 18:23–28, 2004.

Kimberley Upton and Judy Kay. Narcissus: Group and Individual Models to Support Small Group Work. In *User Modeling, Adaptation, and Personalization*, pages 54–65. 2009. URL http://dx.doi.org/10.1007/978-3-642-02247-0_8.

US Department of Education. *Teacher quality. A report on the preparation and qualification of public school teachers*. Washington, DC, 1999.

Erdinç Uzun, Hayri Volkan Agun, and Tarık Yerlikaya. A hybrid approach for extracting informative content from web pages. *Information Processing & Management*, 49(4):928–944, 2013.

Sedef Uzuner. Questions of Culture in Distance Learning: A Research Review. *The International Review of Research in Open and Distance Learning*, 10(3), 2009.

Axel van Lamsweerde. Goal-oriented requirements engineering: a guided tour. In *Proceedings on Fifth IEEE International Symposium on Requirements Engineering*, Toronto, Ontario, August 27-31, 2001, pages 249–262. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=948567.

Kurt VanLehn. Student modeling. *Foundations of intelligent tutoring systems*, pages 55–78, 1988.

Julita Vassileva, Gordon McCalla, and Jim Greer. Multi-agent multi-user modeling in I-Help. *User Modeling and User-Adapted Interaction*, 13(1-2):179–210, 2003.

Katrien Verbert, Nikos Manouselis, Hendrik Drachsler, and Erik Duval. Dataset-driven research to support learning and knowledge analytics. *Educational Technology & Society*, 15(3):133–148, 2012.

Karane Vieira, da Silva, Altigran S., Nick Pinto, de Moura, Edleno S., Cavalcanti, João M. B., and Juliana Freire. A Fast and Robust Method for Web Page Template Detection and Removal. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, Arlington, VA, USA, November 5 - 11, 2006, pages 258–267. URL http://doi.acm.org/10.1145/1183614.1183654,AddtoCitaviprojectbyDOI.

Hubert Vogten, Rob Koper, Harrie Martens, and Jan van Bruggen. Using the Personal Competence Manager as a Complementary Approach to IMS Learning Design Authoring. *Interactive Learning Environments*, 16(1):83–100, 2008. URL `http://www.editlib.org/p/64776`.

Jakob Voss. Measuring Wikipedia. In *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*, 2005.

Riina Vuorikari and Santi Scimeca. Social Learning Analytics to Study Teachers' Large-Scale Professional Networks. In *Open and Social Technologies for Networked Learning*, pages 25–34. 2013. URL `http://dx.doi.org/10.1007/978-3-642-37285-8_3`.

Lev Vygotsky. *Thought and Language*. MIT Press, Cambridge, MA, 1934/1986.

Lev Vygotsky. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA, 1978.

Claudia Wagner, Matthew Rowe, Markus Strohmaier, and Harith Alani. What Catches Your Attention? An Empirical Study of Attention Patterns in Community Forums. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, Dublin, Ireland, June 4–7, 2012.

Priscilla Walmsley. *XQuery*. O'Reilly, Farnham, Calif, 2007. ISBN 9780596006341.

Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, 1994.

Duncan J. Watts and Steven H. Strogatz. Collective Dynamics of Small-World Networks. *Nature*, 393(6684):440–442, 1998.

E. Wenger. Communities of Practice and Social Learning Systems. *Organization*, 7 (2):225–246, 2000.

Etienne Wenger. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press, Cambridge, UK, 1998.

Martin Wolpers, Jad Najjar, Katrien Verbert, and Erik Duval. Tracking Actual Usage: the Attention Metadata Approach. *Educational Technology & Society*, 10(3):106–121, 2007. URL `http://www.ifets.info/journals/10_3/8.pdf`.

Peter. Woolf, Christopher Burge, Amy Keating, and Michael Yaffe. *Statistics and Probability Primer for Computational Biologists*. Spring, 2004.

Michael Wunder, Siddharth Suri, and Duncan J. Watts. Empirical Agent Based Models of Cooperation in Public Goods Games. In *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, Philadelphia, PA, USA, June 16 - 20, 2013, pages 891–908.

Ying Xie, Fengfeng Ke, and Priya Sharma. The effect of peer feedback for blogging on college students' reflective learning processes. *The Internet and Higher Education*, 11(1):18–25, 2008.

Taha Yasseri, Robert Sumi, and János Kertész. Circadian patterns of Wikipedia editorial activity: a demographic analysis. *PLOS ONE*, 7(1):e30091, 2012.

Eric Siu-Kwong Yu. *Modelling strategic relationships for process reengineering*. PhD thesis, University of Toronto, Toronto, Canada, 1995. URL `ftp://learning.cs.toronto.edu/dist/eric/DKBS-TR-94-6.pdf`.

Eric Siu-Kwong Yu. Towards Modelling and Reasoning Support for Early-Phase Requirements Engineering. In *Proceedings of the 3rd IEEE Int. Symp. on Requirements Engineering (RE'97): Proceedings of the 3rd RE*, Washington D.C., USA, January 6-8, 1997, pages 226 – 235. URL `http://www.cs.toronto.edu/pub/eric/RE97.pdf`.

Eric Siu-Kwong Yu. Social Modeling and i*. In *Conceptual Modeling: Foundations and Applications*, pages 99–121. 2009. URL `http://portal.acm.org/citation.cfm?id=1577331.1577340`.

Danuta Zakrzewska. Student Groups Modeling by Integrating Cluster Representation and Association Rules Mining. In *Proceedings of the 36th Conference on Current Trends in Theory and Practice of Computer Science*, Špindleruv Mlýn, Czech Republic, January 23-29, 2010, pages 743–754. URL `http://dx.doi.org/10.1007/978-3-642-11266-9_62`.

Yiwen Zhang and Mohan Tanniru. An Agent-Based Approach to Study Virtual Learning Communities. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, Big Island, HI, USA, 3-6 January, 2005.

Yuzhou Zhang, Jianyong Wang, Yi Wang, and Lizhu Zhou. Parallel Community Detection on Large Networks with Propinquity Dynamics. In *Proceedings of the 15th ACM SIGKDD Conference on Knowledge and Discovery and Data Mining*, Paris, France, June 28 - July 01, 2009, pages 997–1005.

Barry J. Zimmerman. Self-regulated learning and academic achievement: An overview. In *Educational Psychologist*, pages 3–17. 1990.

Vinko Zlatic, Miran Božicevic, Hrvoje Štefancic, and Mladen Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks. *PHYSICAL REVIEW E*, 74(1), 2006. URL `http://link.aps.org/doi/10.1103/PhysRevE.74.016115`.

Henri Zukier. The paradigmatic and narrative modes in goal-guided inference. In *Handbook of motivation and cognition: Foundations of social behavior*, pages 465–502. 1986.

Appendix A

An Example of SPARQL Query

```
PREFIX calaispred:<http://s.opencalais.com/1/pred/>
PREFIX calaistype:<http://s.opencalais.com/1/type/em/e/>
PREFIX calaiscategory:<http://s.opencalais.com/1/type/cat/>
SELECT distinct ?graphs ?name WHERE {
GRAPH ?graphs {{?x rdf:type calaistype:City}
UNION
{?x rdf:type calaistype:Company}
UNION
{?x rdf:type calaistype:Country}
UNION
{?x rdf:type calaistype:Movie}
UNION
{?x rdf:type calaistype:Organization}
UNION
{?x rdf:type calaistype:Person}
UNION
{?x rdf:type calaistype:ProvinceOrState}
UNION
{?x rdf:type calaistype:PublishedMedium}
UNION
{?x rdf:type calaistype:Technology}
.
?x calaispred:name ?name .
}
FILTER regex(str(?graphs), "^files://URCH_POST")
}
```

Listing A.1: SPARQL query for values of different entities like Person or Technology