Groups and Group-Instantiations in Mobile Communities – Detection, Modeling and Applications

Georg Groh
Technische Universitaet Muenchen
Faculty for Computer Science / Informatics
85748 Garching, Germany
grohg@in.tum.de

Abstract

The paper investigates techniques for the detection and modeling of groups of people in a mobile community also interacting in the real world. In this scenario, the concept of contextual manifestations or instantiations of existing social groups is of special significance. In view of detecting and modeling of groups and their instantiations, several examples for classes of data (location- and velocity-data, natural language interest phrases and text-based communication-data) in the information spaces of mobile communities are investigated. Similarity measures between people are constructed and verified by empiric means or through stochastic simulation. On this basis, special clustering approaches for the detection and modeling of groups are developed and tested. The paper concludes with a discussion on possible applications for the methods which have been developed.

1. Introduction

The advent of Web 2.0 and a blooming of user generated content in various forms (Wikipedia, the Blogosphere etc.) has also fostered an increased interest in detecting and modeling social structures of web-users. Either social relations are explicitly declared via two-way handshaking and represented as lists of "friends" or "buddies" in social-networking-, community-, or communication-platforms such as Friendster, MySpace or Skype. On the other hand, the large amount of user generated content in discussion boards, wikis and blogs (directed or undirected 1:n or n:m communication acts) also allows for the detection and modeling of social structures (1:1 relations, group-relations etc.) via AI techniques such as NLP-analysis, clustering etc.. Adding a further interesting aspect to that is the widespread use of mobile communication means (phones, PDAs etc.) and the increasing spread of built-in context-sensors such as GPS receivers also provides for an even broader basis of data which can be used to characterize social structures (like groups), especially on a shorter time-scale. Instantiations of longer-term social relations (e.g., family meetings (in case of group relations)) become analyzable. The scenario of mobile interaction with social media is especially interesting because it is often seamlessly integrated with "normal" real-life social interaction and provides much deeper insights into "real-life" social structures (as opposed to purely "virtual" relations) especially with respect to applications that support social interaction in a context-aware fashion. An obvious example for social structures are groups. Besides 1:1 social relations to other persons, groups strongly determine our perception of our social existence. We thus aim at modeling groups on the basis of data that is available in the information spaces associated with mobile communities that is communities of people that interact via face to face communication, electronically via mobile media and also via "conventional" community support and communication support platforms instantiating blogs, wikis, discussion boards etc..

In section 2 we review some basic notions such as the socio-psychological notion of a group and characterize the scenario we aim at in more detail. Section 3 introduces several types of data that are available in the collaborative information spaces of mobile communities and how similarity measures between people implementing socially valid heuristics can be constructed on the basis of these data. We then investigate in section 4 how we can construct multi-modal-models for groups and their contextual instantiations and finally section 5 presents ideas for the application of these models.

2. Basic scenario

Virtual Communities and support fo these have been extensively discussed in the late 1990s (see e.g., [13]). In [7] the following definition for virtual community was given:

> A set of people which have a high degree of community-awareness, communicate with other members via electronic media, and have a common pursuit which can be identified with the pursuit to collaboratively build up a thematically focused, information- or knowledge-space. This collaborative information- or knowledge-space (CIKS) predominantly contains semi-formal implicit "warm" information or knowledge with a strong emphasis on textual form.

Contributing to the CIKS is understood in this context as a communication act. A single user's blog is a special case of a very simple CIKS where the communication acts are 1:n. The community in this case could be informally bounded by the Blog's author and her friends. More explicitly bounded communities or platforms for several communities that contain richer CIKS-structures (e.g., implementing n:m communication) are also very common on the web.

In the COSMOS project [7, 5], we investigated how the concept of community support platforms containing support-means and services for the buildup of CIKS (social-network visualizations, discussion boards etc.) could be extended in the case of *mobile interactions*. Together with a German

ICWSM'2007 Boulder, CO, USA

TelCo-provider we implemented a concept for a mobile community support platform with services allowing e.g., for the mobile lookup and management of the whereabouts of one's friends or for attaching messages to locations (virtual graffiti) or moving persons (see [8, 5]).

Usually, a mobile community support system and it's CIKS is bound to a server-platform on the web and is accessed through mobile (and also stationary) clients. It is also possible to develop concepts for a completely distributed system based on user agents which we currently investigate.
Besides elements characterizing users (profiles (also containing a user's highly dynamic context)) and information-items, a CIKS of a mobile community also contains elements characterizing various relations e.g., user-user social relations in varying degrees of explicitness.

An interesting question that can be posed in such a scenario is: What can we infer from either contextual (highly dynamical) and/or other elements of a mobile community's CIKS about the group structures between community members? How can we amalgamate several such inferred group models and /or with models of explicitly stated group memberships? What are applications for such group models?

2.1 Groups and group instantiations

Groups play an important role in almost any branch of science that investigates structures which are either human generated or human related. Such sciences include computer science (e.g., teams in groupware) law science (e.g., groups as legal entities), economics (e.g., working teams), ethnology (ethnic groups), history (e.g., social and political groups of the past), art (e.g., artist groups), etc. While all these scientific disciplines investigate rather special aspects of groups, sociology and especially social psychology try to characterize groups from a more generic point of view.

The field of research in sociology and *social psychology* which deals with groups like we have them in mind is usually designated *small group research*. The term small group attempts to distinguish the scientific subject from sets of people of the size of political parties, ethnic groups and the like.

The *main characterizing features of a group* are described in the following points:

- The *minimal definition* of a group is a comparatively small number of people which interact with each other directly via face-to-face-interactions [11, 6]
- The *number of group members* is usually so small that direct *face-to-face-interactions* are possible between all members [11, 6]. An often stated number for an upper bound is 20 [6]. This can also be justified by considering results from cognitive sciences which suggest respective limitations of human cognition and perception.
- The interaction situations must be of a *certain duration* in order to allow for common structures like norms or goals [6].
- Group members share a *network of interpersonal attraction* (Hare in [1]).
- Often, the members of a group have interdependent characteristics: *common goals*, *common norms*, a *special communication structure*, a *role- and affect structure*, and a *group awareness* [6, 1].
- Groups are often characterized by *immediately perceivable features* (like names, uniforms etc.) which allow others to perceive the group as a whole and which define borders of the group. [4, 10].

A definition from Homans (1950) [11], which is an often cited common denominator in small group research sums up the notion of a group:

> "A group is a number of persons who communicate with one another often over a span of time, and who are few enough so that each person is able to communicate with all the others, not at second hand, through other people, but face-to-face."

Although the face to face communication is emphasized, the above definitions are to some extend transferable to groups that communicate using electronic means.

Considering social structures like groups with respect to Context and Context-Parameters such as location one can make further interesting observations. The first observation is that space and resulting propinquity have direct influence on groups. As reviewed in [7], several studies in social psychology indicate that distance has an inverse effect on social impact and thus on group formation and stabilization parameters. Furthermore, since our sensor apparatus restricts our ability to communicate and also restricts the number of people we can (more or less) simultaneously communicate with we see that space also plays an interesting role for group processes on scales as small as several meters.

What we further observe (and support by the aforementioned conclusions) that is that we have to distinguish between social groups that exist in an abstract way without having to consider highly contextual parameters such as the exact location and instantiations of these groups or even Ad-Hoc-groups that both only exist within the boundaries of a certain context. As an example for a *group with abstract existence* consider a family. An *instantiation of this group* in a certain context would be the family meeting on Christmas.

Means for detecting and modeling such group instantiations, amalgamating them with models of abstract groups and designing services for their support with respect to (mobile) social media have not jet been very extensively investigated.

3. Data in CIKS cuitable for group detection and modeling

We will now review data available in mobile community CIKS and suitable algorithms that allow for the detection and modeling of groups and group instantiations. We will look at three general classes of data: highly dynamic, context data, explicit self description data and communication content. These classes will be discussed by investigating examples from every class.

3.1 Location and velocity

In order to detect instantiations of groups in certain contexts we need to model the context of individual community members and then analyze which subsets of these are good candidates for group instantiations. The most obvious context parameter often discussed in literature is location. We assume that the locations and also the velocities of potential group members are accessible to the group instantiation detection algorithms at any time (which implies that we can also estimate the velocity). We furthermore require a certain minimum accuracy (within 5 m at least).

Our experiments (involving a lifestyle community with 100 prototype users with mobile devices (PDAs)) show that the resulting privacy problems can be dealt with in a satisfactory way (see [14]). Furthermore these field studies also clearly prove that GSM cell based location techniques with an accuracy of roughly 200 m in urban areas are not suitable for context aware support for socially interacting mobile communities. Thus a GPS/Galileo based accuracy is a minimum requirement for suitable group instantiation detection. Central (server-side) knowledge of whereabouts can be replaced by a distributed scenario where the personal information is completely controlled and maintained in the (mobile) user agent. Since these algorithms are still subject to ongoing research we will assume the simpler "centralized" approach.

Although social psychology predicts a positive influence of spatial proximity with respect to social impact and group formation (see [7]) common sense dictates that naive clustering of locations will not necessarily provide good candidates for group instantiations. There are several reasons for that:

- Spatial proximity is relative. Compared to the long distance between New York and Munich, the distance of several people being 2 km apart seems neglectable. Nevertheless it seems obvious that people 2 km apart cannot reasonably be assumed to be engaged in an instantiated group relation context.
- Spatial proximity without aligned velocities cannot be of contextual social significance either. People driving by each other at opposite speeds will most certainly not be engaged in an instantiated social group relation
- Standing in line at a fast food restaurant with several other people certainly accounts for spatial proximity and also for proximity with respect to velocities but is also probably not an indicator for these people to form an Ad-Hoc-Group or an instantiation of an abstract group.

While we have to account for the last point by combining and checking the group models on the basis of context data with group models derived from data with slower dynamics, the first two points and similar arguments can be taken care of by introducing suitable social heuristics into the clustering process.

Due to the lack of a broad basis of real people's movement data, we developed an extensive stochastic mobility simulation of people moving in urban areas. The difficulty was, that models for individual motion (such as Random Walk based models (see [3]) and models for group motion (such as the reference point group model [12]) are readily available but no models are available which combine and smoothly transfer individual motion into group motion and vice versa. Thus a new mobility model based on probability models and Markov chains and realistic estimation of the corresponding parameters was created that allows the simulated entities to come together, move or rest as a group and to dissolve again. The simulation also provides detailed data about the "real" group structures at any iteration thus allowing for a quality measure for the proposed group instantiation detection algorithms. (see [7] for an extensive discussion on the details).

In order to detect and model group instantiations and Ad-hoc-Groups and especially to distinguish them from pseudo groups with no social relevance, algorithms on the basis of

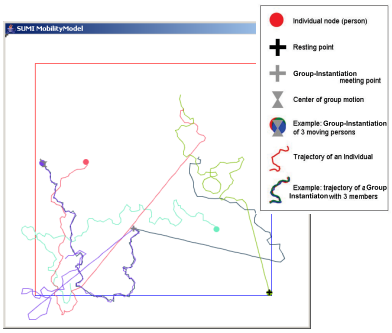

Fig. 1: *SUMI simulation with 5 nodes*

crisp clustering approaches were developed. In order to be able to compare the "real" groups occurring in the simulation with the groups found by our detection algorithm, we developed a precision / recall- based performance measure. This measure ensures that groups can not only be compared on a basis of identity but also with respect to group members, thus allowing for small errors in detection (lets say a group with one member missing) account only for small decrease in detection quality, although the found group indeed differs from the "true" group (e.g., by a member missing etc.). For the details see [7].

Accounting for the importance of velocity that has been stated above, standard crisp clustering approaches such as SAHN, K-Means and MST (see [7]) were used to cluster individuals represented as vectors from $\mathbb{R}^4$ treating the 2-D-location on an equal footing as the 2-D-velocity. Since these algorithms all require the desired number of clusters as an input parameter, the algorithms are run with all possible values for the number of clusters as input and a cluster validation measure is needed to distinguish the best clustering and thus the "optimum" number of clusters (groups). Using a fixed self developed cluster validation measure (which will be discussed below) we found that the actual clustering algorithm did not make much of a difference (see table1) with respect to cluster detection quality. What is more interesting is the

	SAHN	K-Means	MST
Precision P_{tot}	0.8283	0.7843	0.8283
Recall R_{tot}	0.7111	0.6663	0.7111
F-Measure $F_{0.5\text{tot}}$	0.7523	0.7073	0.7523
Computing Time [sec]	367.5	150.5	11910.7

Table 1: *Varying the clustering algorithm. SAHN was single link. A Euclidean distance measure and the* SCVS *were used for all three experiments. The same simulation data was used for all three experiments*

difference between several cluster validity measures. In general we can assume that (besides soldiers moving in combat), spherical clusters are more suitable for human group instantiations than elongated clusters. Thus we didn't take algorithms targeted especially for these into account. Furthermore, since we can assume that face to face contact is important for group-instantiations and also assuming that the limit for human communication distances which is also important for group instantiations is roughly $\tau_l = 30$ meters we have to take this into consideration when constructing an adopted social cluster validation measure. Assuming cut off values of

τ_l for location-variations and τ_v for velocity variations within the group instantiation we can construct such a social validation measure. Together with further heuristics with respect to consolidating smaller groups found by the algorithm into larger groups (see [7]) we developed the Social Cluster Validation and Selection procedure (SCVS) which performed on the simulation data (which also were aimed at maximum realism) much better than e.g., the well known Dunn-Index [9]. Varying the parameters τ_l and τ_v yields the result that either

	SCVS	Dunn
Precision P_{tot}	0.7843	0.2451
Recall R_{tot}	0.6663	0.2451
F-Measure $F_{0.5\text{tot}}$	0.7073	0.2451
Comp. Time [sec]	150.5	162.7

Table 2: *Varying the Distance Measure. K-Means and the same Simulation data were used for all experiments. SCVS parameters were* $\tau_l = 1.0$ *and* $\tau_v = 2.0$ *(Relative Units)*

too small values (restriction of cluster acceptance is too restricted) or too large (cluster acceptance is too loose) doesn't lead to optimal performance. [7] yields the details of this investigation. Summing up was has been said so far we can see that integrating heuristics from social psychology can be a great help in detecting group instantiations on the basis of context parameters and context models.

We will now investigate a second class of information that is interesting to detect groups: Explicit self information.

3.2 Explicit self-information

The most obvious example of explicit self information is the direct declaration of social information. Most community platforms (mobile or not) allow for buddylists or even social networks to be set up by the users themselves or at least via handshaking between two users. Obviously, this kind of self information allows for the direct derivation of group models which clearly have a high significance. In case of social networks we e.g., can rely on group models proposed by sociometry which describes groups as special sub-graphs (cliques, clans etc.) in the graph of dyadic social relations [16].

But there are less obvious examples of explicit self information which can be used to detect groups. Since in case of explicit self information we usually have a slow dynamic w.r.t. changes in that data, they are much more useful for detecting abstract groups. (we will see later how to amalgamate them with group instantiations found on the basis of highly dynamic context data. One type is explicit self information on personal interests which can be found in various profiles on most community- or communication-platforms on the Web.

These interest phrases are either free text or chosen from a fixed (often hierarchical) ontology. We collected samples for both types and developed similarity measures for both types that allow for comparing two people with respect to their interests. These measures could also be used to infer dyadic social relations. We used them as an input to a special relational clustering algorithm which delivers groups with respect to interests.

The simpler of both types of interest utterances consist of vectors of interest keywords chosen from a fixed ontology of interest keywords. Comparing them is quite easy, since we deal with vectors instead of sets. Based on a normalized inner product, we incorporated an idf-like measure to account for the fact that very common interests shared by most of the community members are no good discriminators with respect to similarity. Furthermore the taxonomic structure of the interest ontology was taken into consideration by weighing interest correspondences at the leaf level (low generality) higher than correspondences at an inner node level (higher generality). The precise description of the similarity measure is given in [7].

More demanding with respect to similarity computations are sets of free text interest phrases. They may often contain very special vocabulary often in unusual orthography or with abbreviations which makes the use of stemming or spelling correction algorithms for the unification of terms in view of semantic comparability difficult. Furthermore, since the free text phrases are given without a fixed order, we have to compare each with each if we want to give an overall similarity for two such sets. Syntactic and grammar analysis (e.g., POS tagging) is also difficult due to the proprietary language often used. Furthermore word sense disambiguation techniques that would also be helpful w.r.t. to semantic analysis is also difficult because of the lack of proper training data considering the special community vocabulary. Conventional other approaches from statistical NLP are also difficult to apply because the phrases are often very short. We thus opted for a Semantic Net based approach using a WordNet-based comparison algorithm allowing to state the semantic relatedness of terms in the phrases and a dictionary based approach where terms were semantically compared by comparing their Encarta-dictionary entries. The dictionary approach is based on a simple vector model based cosine similarity between word vectors of the dictionary texts. The Word-Net-based part of the measure compares the synsets of two terms by their path-lengths to the nearest common super-concept (synset) in a taxonomy of generalized abstraction relations (GAR) and is a further development of an approach suggested in [17]. Figure 2 depicts the paths. The complete approach is shown in figures 3 and 4. Due to space restrictions not all details could be explained here. The interested reader is referred to [7].

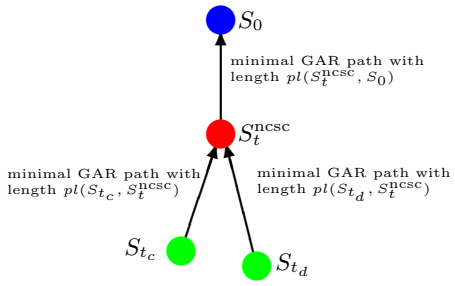

Fig. 2: *Path lengths between Synsets, their nearest common generalized super-concept and the root node*

In order to verify the algorithm, we collected 100 sets of free text interest phrases through a survey. We then conducted a second survey where people where asked to rate the similarities between a subset of 20 of these sets on a scale from 0 to 10. We furthermore conducted a third survey where people where asked to estimate the similarity between single interest phrases chosen from the second survey's sets. For the second and third survey we calculated the mean and standard devi-

$\mathrm{sim}(\mathcal{X}^{(j_1)}, \mathcal{X}^{(j_2)})$(Part(I)

- `Input:`
  - `Two sets of free text interest concepts` $\mathcal{X}^{(j_1)}, \mathcal{X}^{(j_2)}$:

$$\mathcal{X}^{(j_1)} = \{C_0^{(j_1)}, C_1^{(j_1)}, \ldots, C_{|\mathcal{X}^{(j_1)}|-1}^{(j_1)}\} \text{ and}$$
$$\mathcal{X}^{(j_2)} = \{C_0^{(j_2)}, C_1^{(j_2)}, \ldots, C_{|\mathcal{X}^{(j_2)}|-1}^{(j_2)}\} \quad ;$$

    `where each of the concepts` C `has a free textual phrase representation of length` $N(C)$:

$$C_i^{(j_1)} \triangleq \hat{C}_i^{(j_1)} = (w_{n_0}, w_{n_1}, \ldots, w_{n_{N(\hat{C}_i^{(j_1)})-1}})$$
$$C_k^{(j_2)} \triangleq \hat{C}_k^{(j_2)} = (w_{m_0}, w_{m_1}, \ldots, w_{m_{N(\hat{C}_k^{(j_2)})-1}})$$

  - `Required resources: WordNet 2.0; Microsoft Encarta Dictionary; Stop-list`
- `Preprocess:`
  - `Iterate over all free text interest phrases in both sets:`

    `for(`$i, k = 0; i < |\mathcal{X}^{(j_1)}|, k < |\mathcal{X}^{(j_2)}|; i$`++,` k`++)`

    `Remove Stop-words (SR) and multiple instances of words (MIR):`

$$\hat{C}_i^{(j_1)} \xrightarrow{\text{SR, MIR}} \tilde{C}_i^{(j_1)} = (w_{p_0}, w_{p_1}, \ldots, w_{p_{N(\tilde{C}_i^{(j_1)})-1}})$$
$$\hat{C}_k^{(j_2)} \xrightarrow{\text{SR, MIR}} \tilde{C}_k^{(j_2)} = (w_{q_0}, w_{q_1}, \ldots, w_{q_{N(\tilde{C}_k^{(j_2)})-1}})$$

    `endfor`
- `Prepare(I):`
  - `A Hashtable of all words` $\mathscr{V}(\mathcal{X}^{(j_1)}) = \{w_{r_0}, w_{r_1}, \ldots w_{r_{|\mathscr{V}(j_1)|-1}}\}$ `in` $\mathcal{X}^{(j_1)}$ `and a Hashtable of all words` $\mathscr{V}(\mathcal{X}^{(j_2)}) = \{w_{s_0}, w_{s_1}, \ldots w_{s_{|\mathscr{V}(j_2)|-1}}\}$ `in` $\mathcal{X}^{(j_2)}$
  - `Retrieve and preprocess Encarta Dictionary entries for all words:`

    `for(`$a, b = 0$; $a < |\mathscr{V}(\mathcal{X}^{(j_1)})|, b < |\mathscr{V}(\mathcal{X}^{(j_2)})|$; a`++,` b`++)`

    * `Get dictionary entries for all senses for` w_{p_a} `and` w_{q_b}`. Concatenate them into` $d(w_{p_a})$ `and` $d(w_{q_b})$`.`
    * `Do stop-word removal, and compute word-vectors by counting term frequencies`

$$d(w_{p_a}) \xrightarrow{SR} \tilde{d}(w_{p_a}) \xrightarrow{CTF} \check{d}(w_{p_a}) = (\nu_{u_0}, \nu_{u_1}, \ldots, \nu_{u_{|\mathscr{V}(\tilde{d}(w_{p_a}))|-1}})$$
$$d(w_{q_b}) \xrightarrow{SR} \tilde{d}(w_{q_b}) \xrightarrow{CTF} \check{d}(w_{q_b}) = (\nu_{l_0}, \nu_{l_1}, \ldots, \nu_{l_{|\mathscr{V}(\tilde{d}(w_{q_b}))|-1}})$$

    `endfor`

Fig. 3: *Similarity Measure for sets of free text interest phrases (part I)*

$\mathrm{sim}(\mathcal{X}^{(j_1)}, \mathcal{X}^{(j_2)})$ (Part(II)

- `Prepare(II):`
  - `Compute a similarity matrix` $\mathrm{Sim}^{\mathrm{dict}} = \mathscr{V}(\mathcal{X}^{(j_1)}) \times \mathscr{V}(\mathcal{X}^{(j_2)}) \to [0,1]$

$$\mathrm{Sim}_{ab}^{\mathrm{dict}} = \mathrm{sim}^{\mathrm{dict}}(w_{p_a}, w_{q_b}) =$$
$$= \begin{cases} 1 & \text{if} \quad w_{p_a} = w_{q_b} \\ \frac{1}{|\tilde{d}(w_{p_a})||\tilde{d}(w_{q_b})|} \sum_{\{(e,f)|w_{u_e} \in \tilde{d}(w_{p_a}) \wedge w_{l_f} \in \tilde{d}(w_{q_b})\}} \nu_{u_e} \nu_{l_f} \; \delta_{u_e l_f} & \text{if} \quad w_{p_a}, w_{q_b} \in \texttt{Dictionary} \\ 0 & \text{else} \quad . \end{cases}$$

  - `Compute a similarity matrix` $\mathrm{Sim}^{\mathtt{WordNet}} = \mathscr{V}(\mathcal{X}^{(j_1)}) \times \mathscr{V}(\mathcal{X}^{(j_2)}) \to [0,1]$

$$\mathrm{Sim}_{ab}^{\mathrm{WordNet}} = \mathrm{sim}^{\mathrm{WordNet}}(w_{p_a}, w_{q_b}) =$$
$$= \begin{cases} 1 & \text{if} \quad w_{p_a} = w_{q_b} \\ \left(1 + \min_{\{(c,d)|w_{p_a} \in S_{t_c} \wedge w_{q_b} \in S_{t_d}\}} \left(\frac{pl(S_{t_c}, S_t^{\mathrm{psec}}) + pl(S_{t_d}, S_t^{\mathrm{psec}})}{2\, pl(S_t^{\mathrm{psec}}, S_0)^{\frac{1}{2}}}\right)\right)^{-1} & \text{if} \quad w_{p_a}, w_{q_b} \in \texttt{WordNet} \\ 0 & \text{else} \quad . \end{cases}$$

- `Compute sim:`
  - `Compare unordered sets of free text interest phrases:`

$$\mathtt{sim}(\mathcal{X}^{(j_1)}, \mathcal{X}^{(j_2)}) = \frac{1}{2} \frac{1}{|\mathcal{X}^{(j_1)}|} \sum_{i=0}^{|\mathcal{X}^{(j_1)}|-1} \max_k \mathtt{sim}(C_i^{(j_1)}, C_k^{(j_2)}) + \frac{1}{2} \frac{1}{|\mathcal{X}^{(j_2)}|} \sum_{k=0}^{|\mathcal{X}^{(j_2)}|-1} \max_i \mathtt{sim}(C_i^{(j_1)}, C_k^{(j_2)})$$

  - `where pairs of free text interest phrases are compared by:`

$$\mathtt{sim}(C_i^{(j_1)}, C_k^{(j_2)}) = \mathtt{sim}(\tilde{C}_i^{(j_1)}, \tilde{C}_k^{(j_2)}) = \frac{1}{2} \frac{1}{N(\tilde{C}_i^{(j_1)})} \sum_{a=0}^{N(\tilde{C}_i^{(j_1)})-1} \max_b \mathtt{sim}(w_{p_a}, w_{q_b}) + \frac{1}{2} \frac{1}{N(\tilde{C}_k^{(j_2)})} \sum_{b=0}^{N(\tilde{C}_k^{(j_2)})-1} \max_a \mathtt{sim}(w_{p_a}, w_{q_b})$$

  - `where pairs of words (lexicalized) concepts are compared by:`

$$\mathtt{sim}(w_{p_a}, w_{q_b}) = \frac{1}{\alpha + \beta} \left(\alpha\, \mathtt{sim}^{\mathtt{WordNet}}(w_{p_a}, w_{q_b}) + \beta\, \mathtt{sim}^{\mathtt{dict}}(w_{p_a}, w_{q_b}) \right)$$

- `Return sim.`

Fig. 4: *Similarity Measure for sets of free text interest phrases (part II)*

ation of the answers yielding two 20x20 matrices.

We also conducted additional surveys to judge the quality of our similarity measure for the vectors of interests chosen from a fixed ontology. For this we used a sample of 100 such vectors from a dating community. The results are also very encouraging and showed that our measure is able to reflect human judgment to a high precision.

The other surveys also showed that our similarity measure for sets of free text interest phrases also performs very well. Since the number of survey participants (30 in total) was too small to reasonably use statistical tests to determine the "randomness" of the human judgments so we chose the square root of the squared mean error (SQMSE) as a means to compare the human judgments with our measure and to determine the result of purely random judgments.

For the sake of brevity we will pick survey one, the details of the other surveys can be found in [7]. Assuming a uniformly random distribution for either the human judgments on similarity and for our measure (scaled to the [0,10] scale of the survey) the calculations give a value of SQMSE=4.47. For the optimum setting of the two parameters α and β of our algorithm we get a correlation of SQMSE=1.158 which shows that our measure is significantly better than random. Since the average human judgment standard deviation was 1.62, so one could conclude that the correlation between our measure and the averge judgment of the survey participants is within the boundaries of human judgment variations.

Varying the parameters α and β which control the weight of dictionary based comparison vs. WordNet-based comparison yields the results in table 3, yielding the optimum setting of $\alpha = 0.7$, $\beta = 0.3$.

Parameters	α=1.0, β=0.0	α=0.7, β=0.3	α=0.5, β=0.5	α=0.3, β=0.7	α=0.0, β=1.0
SQMSE	1.73	1.15	1.36	1.65	2.14

Table 3: *Varying the settings for α and β*

3.3 Communication contents

The third class of data in CIKS are all kinds of communication content in the narrower sense. (Regard that all accessible data in a CIKS (even the self information) can be regarded content of communication acts). We have e.g., tree-like discussion boards (n:m communication), messaging (1:1 communication), blogs (1:n communication) or other forms.

Especially the directed forms of communication allow for various social heuristics with respect to developing a similarity- or social-relatedness-measure. The spectrum ranges from heuristics concerning the communication frequency, the number of messages exchanged, the length of messages or the degree of mutuality. Obviously the content itself can also be investigated. In [7] we proposed a simple measure comparing the contents of messages and the number and mutuality of messages into a simple measure for social connectedness.

A simple example to illustrate that: A very simple heuristic that is incorporated into the measure is based on two plausible assumptions:

(1) The larger the difference between the numbers of replies

is, the lower is the intensity of interaction.

(2) The larger the number of mutual replies is, the more intense is the interaction of the two persons.

Thus, a possible similarity measure between person k_1 and person k_2 could be constructed like this (denoting the number of replies from person k_1 to postings authored by person k_2 by $m_{k_1 \to k_2}$ and denoting the number of replies from person k_2 to postings authored by person k_1 by $m_{k_2 \to k_1}$ respectively).

$$\mathrm{sim}(k_1, k_2) = 1 - \frac{1}{(1 + m_{k_1 \to k_2} m_{k_2 \to k_1} \exp(-\frac{|m_{k_1 \to k_2} - m_{k_2 \to k_1}|^2}{\sigma^2}))^q} \quad (1)$$

The factor $m_{k_1 \to k_2} m_{k_2 \to k_1}$ in the denominator was introduced in accordance with observation (2) above and the factor $\exp(-\frac{|m_{k_1 \to k_2} - m_{k_2 \to k_1}|^2}{\sigma^2})$ was introduced in accordance with observation (1) above. The exponent q and the deviation σ are parameters which can be adapted.

3.4 Fuzzy clustering of relational data

Unfortunately, for the data classes with slower dynamics we did not have social network data at hand to compare the results of clustering with "real world" groups e.g., among the creators of our 100 sets of free text interest sets. Nevertheless a sound and promising clustering algorithm using the relational data that can be computed from the proposed similarity measures.

First of all it is clear that in contrast to the location case we need fuzzy clustering approaches since e.g., groups with interests cannot well be modeled with crisp sets. Surprisingly few of the numerous fuzzy clustering approaches were able to operate on relational data.

Usually, algorithms like Fuzzy C-Means optimize intra cluster compactness by minimizing a target function such as a generalized square error function

$$J_{SQE} = \sum_{k=1}^{n} \sum_{i=1}^{c} U_{ik}^m ||x_k - \pi_i||^2 \quad (2)$$

(where the n patterns are denoted by x_k, the c cluster prototypes by π_i, the fuzzy membership matrix by U_{ik}, and the fuzzyness control parameter by m) through iterative alternating optimization, where the membership-matrix and cluster prototypes which depend on each other are adapted alternatingly in every step.

What we have is similarities between the users (patterns) on the basis of our heuristic social measures which can also be interpreted as strengths of the associated social ties. A metric pattern space that is required for the usual approaches is missing. Thus the idea is that switching from continuous adaptation of the cluster prototypes to a discrete version, where always a single pattern (user) is used as a cluster (group) prototype. The result is called RACE (relational alternating cluster estimation) [15]. Unfortunately the original algorithm had some severe drawbacks that were corrected by introducing a simulated annealing technique into the clustering process (see [7]).

4. Group models

What we have gained so far are reasonable heuristics for social ties based on contextual (highly dynamic) and other data in mobile communities which have been checked by simulation, surveys or are supported by findings from social psychology. Furthermore, we developed suitable algorithms that can detect and model groups and group instantiations on the basis of these data. How can we now amalgamate these into overall models for groups that have a high degree of real social relevance that can be used in intelligent support services? Due to space restrictions we can only review parts of the considerations (see [7] for more aspects).

First one can consider that from the many groupings detected through the context data, only a subset can be regarded instantiations of abstract groups. How can these be distinguished from ad-hoc-groups (that have social significance in the context but do not correspond to an abstract group) and pseudo groups (which have no social significance)? One good criterion is regularity. It can be argued [7] that abstract social groupings that have a high impact on a user are instantiated on a *regular* basis. For example, the group of colleagues that people work with on a daily basis, the members of the football team that they play with weekly or the group of friends that they meet every Saturday etc.. Thus if we detect that a certain group shows a regular pattern of instantiations we can conclude that an underlying abstract group exists which is of social significance. In order to detect such regularities we have to compare found clusters of people with respect to their members and furthermore we have to analyze the time-pattern of their appearance. For both aspects we have developed metrics. Comparing the groups w.r.t. to members is achieved through a modified Hamming distance based approach which is normalized with respect to group size. Regularity analysis is achieved either by an adapted Fourier analysis or through a statistical measure (see [7]). The measures were tested against the mobility simulation and were able to find a large fraction of the simulated groups that "met" on a regular basis (based on abstract groups of the simulation).

Using this approach we cannot rule out that regular semi-pseudo-groups are found. For example we might be in close proximity to the same people on the bus we use every day to go to work without having a deep social relation to them. We could certainly use these semi-pseudo-groups (which have indeed a "real" contextual common denominator) for applications that proactively bring people together. But even then it is more reasonable to combine the findings on the basis of contextual data with the findings from less dynamic data to give group models that have an even higher social relevance.

In order to do that we have to fusion the crisp models from context with the fuzzy group models gained from the other data. Since the crisp models can easily be represented in the fuzzy framework by using $\{1, 0\}$ as values for U_{ik} we can combine them with fuzzy set operations: Let $\mathcal{A}_1, \mathcal{A}_2$ be fuzzy sets $\mathcal{A}_i = \{(x, \mu_{\mathcal{A}_i}) | x \in \mathcal{X}; \mu : \mathcal{X} \to [0, 1]\}$ in $\mathcal{X}$, then [18]:

$$\mathcal{B} = \mathcal{A}_1 \cap \mathcal{A}_2 \quad \Rightarrow \quad \mu_{\mathcal{B}}(x) = \min(\mu_{\mathcal{A}_1}(x), \mu_{\mathcal{A}_2}(x)) \quad (3)$$

$$\mathcal{B} = \mathcal{A}_1 \cup \mathcal{A}_2 \quad \Rightarrow \quad \mu_{\mathcal{B}}(x) = \max(\mu_{\mathcal{A}_1}(x), \mu_{\mathcal{A}_2}(x)) \quad (4)$$

$$\mathcal{B} = \mathcal{A}^c = \mathcal{X} \setminus \mathcal{A} \quad \Rightarrow \quad \mu_{\mathcal{B}}(x) = 1 - \mu_{\mathcal{A}}(x) \quad (5)$$

5. Applications

Having models for groups facilitates some applications in the context of social sensitive support for communities. We will look at two examples. More aspects and applications are described in [7].

5.1 Indication of groups

Visualizing (indicating) groups seems at first glance to be of minor interest, since the eye is a much better clusterer than any algorithm. Some arguments nevertheless strongly speak in favor of this application:

- If the number of mobile community members is very large, the eye can well identify groups but the *relevance of the groups* for the single user cannot be easily determined visually, because a large number of points on a map can only be visualized with the help of legends (e.g., the users are visualized with numbered dots and the association of the numbers with the user-ids must be looked up in the legend). If only members of *explicitly declared personal groups* (e.g., Buddylists) are displayed on the Map, this problem does not occur but the user will *miss groups* that are not formed by his buddies or that have only one buddy as a member. Thus such a restricted indication service does not facilitate the build up of new social relations but is only a potentially useful bookkeeping and awareness functionality for existing ones.

- Especially when (Ad-Hoc-)Groups with respect to context parameters and abstract groups with respect to less dynamical parameters (e.g., interests) are *combined* (see section 4), the aspect of relevance can be filtered much more effectively.

- In order to display a sufficiently large number of user-locations for an appropriate overview, a map with a sufficiently *large scale* needs to be used to visualize a sufficient number of user-locations. On this scale, *individual users* in socially relevant groups are usually *not distinguishable* any more because the symbols are printed on top of each other. If groups are modeled explicitly, expandable group symbols could be used instead of printing the individual users on top of each other.

- Furthermore, techniques like the Social Cluster Validation and Selection procedure *SCVS* allow for a degree of *expressiveness of the models* (with respect to general relevance and validity of clusters as groups) that is not accessible with the "naked" eye.

- Without detecting and modeling groups and group instantiations, no *abstract groups with respect to context parameters* can be found and thus no such abstract groups can be indicated. In order to give an overview of a social situation, the indication of abstract groups (e.g., their meeting locations and their meeting frequency) is of great value.

5.2 Social sensitive collaborative filtering

Collaborative Filtering is essentially about *predicting the degree of relevance of an information-item* for one person *on the basis of degrees of relevance* of that information item for *other persons*. The degrees of relevance of the item for other persons is either estimated *implicitly* or raised *explicitly* in form of *ratings*. The relevance degrees for the other persons are generally weighted with the strength of a relation from the person in question to these other persons.

Denoting the rating of user k for item j by v_{kj} and the set of items that user k has voted on by I_k and the average voting of user k by

$$\overline{v}_k = \sum_{j \in I_k} v_{kj} \tag{6}$$

we can *predict* or estimate the voting of user k for an item with index j_m that he has not seen or rated yet by [2]

$$v_{kj_m} = \overline{v}_k + \alpha \sum_{k_a=1}^{n} w_{kk_a}(v_{k_a j_m} - \overline{v}_{k_a}). \tag{7}$$

In this general formulation, the weights w_{kk_a} are a measure for the similarity or correlation between users k and k_a and α is a normalization parameter. and their calculation distinguishes the basic CF approaches. The most popular variants are the Pearson correlation or cosine measure.

All these ways to calculate the similarity or correlation between users are based solely on the ratings of the users. These ratings can be gathered explicitly (by interacting with the users) or implicitly (by inferring them via measurable parameters such as the time a user views item j or the frequency of interaction etc. In any case, the similarity is calculated on the basis of *user-item relations* ("ratings") alone and neglects other relations such as *user-user relations, group structures* etc. that are very useful for calculating the weights.

In the *conventional Collaborative Filtering* approach, we have the *assumption* that the rating for an item from a user k_1 is more important for guessing the usefulness of the item for a user k_2 the more similar the rating behavior of both users is. In that way, the Collaborative Filtering process is assumed to be "self-adjusting": No matter what the other relations between the users are like: If they like the same items, a new comparable or similar item liked by one of them may be of use for the other. But a key point is that the new *item should be similar* to the previously related ones. Thus the idea of Collaborative Filtering works best when the items to be filtered are similar to the items which have been used to train the Collaborative Filtering system Thus, if we have very heterogeneous information items the conventional approach does not work well.

The argument can also be turned around: if the *users are similar* to each other and the new item is not very similar to the previously seen ones, then the similarity between the users is a criterion to nevertheless positively recommend the item. However, the similarity aspect between the users should match the "topic" of the item.

So what can be generally assumed is that *Collaborative Filtering works best*, when

- *User-Item-Relations* (ratings) are similar between users.
- *Item-Item-Relations* indicate similarity between filtered items
- *User-User-Relations* indicate similarity between filtering users

While the first aspect is (as has been discussed above) directly respected in the conventional approach. the second and third aspects are usually only implicitly respected by restricting the topic focus of the platform that offers CF.

The most obvious idea to include user-user-relations is to *complement the weights* calculated on the basis of rating-similarity with weights calculated on the basis of *user-user*

similarity. We should look for suitable social structures which have a high probability of being reflected in algorithmically accessible data. *Ad-Hoc-Groups, group instantiations and abstract groups* are natural candidates for such structures on the sub-community level.

Where a community's common pursuit would be too general and where the mere organizational aspect of a community (as a provider of service bundles) is very prominent, abstract groups within the community are the "*self-adjusting*" component for CF from the user-user-relation perspective. Groups play a tremendously important role in structuring the personal knowledge and information sphere and it is reasonable to assume that within an abstract group

- users have *tight social relations* with one another. Example: Family
- a *group-CIKS* can be abstractly associated with the group which is likely to be *topically focused* on the group's main social *coherence aspect.*
- if the items in this group-CIKS are *not directly topically related* (trivial relation), they are likely to be *related indirectly* via the *social brace* that the group represents.

Especially if two or more aspects are combined in the course of the detection phase, the group is likely to have a real existence (e.g., abstract groups wrt. location and velocity and abstract groups wrt. interests are intersected) and thus is likely to be a natural basic population for knowledge management and information flow control services such as CF.

We denote (as usually) the membership degree of a user k in a group $\mathcal{G}_i$ (which can be a combined group (via the methods described in section 4) or a single aspect group) by U_{ik}, the group's prototype by π_i and the item that is to be filtered by I_j. There are various possibilities to include the group model into the augmentation / complementation equation for the weights

$$w'_{kk_a} = \frac{1}{1+\beta}\big(w_{kk_a} + \beta \;\; U_{ik}U_{ik_a}\mathrm{sim}(k,k_a)\big) \tag{8}$$

$$w'_{kk_a} = \frac{1}{1+\beta}\big(w_{kk_a} + \beta \sum_i U_{ik}U_{ik_a}\mathrm{sim}(k,k_a)\big) \tag{9}$$

$$w'_{kk_a}(I_j) = \frac{1}{1+\beta}\big(w_{kk_a} + \beta \sum_i \mathrm{sim}(\pi_i, I_j)\; U_{ik}U_{ik_a}\mathrm{sim}(k,k_a)\big) \tag{10}$$

Equation 8 uses only one group membership to filter the general user-user similarity $\mathrm{sim}(k,k_i)$ (for which we can set e.g., the interest similarity $\mathrm{sim}(\mathcal{X}_k, \mathcal{X}_{k_a})$). Equation 9 uses all existing groups to filter the general user-user similarity and equation 10 introduces a further similarity which aims at comparing the group's "topic" with the item to be filtered. This similarity measure would have to be defined e.g., on the basis of the vector model or more sophisticated means.

6. Conclusions

In this paper we have shown how groups in mobile communities can be detected modeled and applied. Transforming the proposed approaches to a completely distributed agent-oriented scenario where every user controls his / her own data and no central community platform is necessary are the next goals for future work.

References

[1] E. Bradner. *Computer Mediated Communication Among Teams: What are 'Teams' and how are they 'Virtual'?* in C. Lueg, D. Fisher (Ed.): From UseNet to CoWeb - Studying Social Information Spaces, Springer, 2003.

[2] J. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. *Technical Report MSP-TR-98-12, Microsoft Research*, 1998.

[3] T. Camp, J. Boleng, and V. Davies. A survey of mobility models for ad-hoc-network research. *Wireless Communications and Mobile Computing (WCMC): Special issue on Mobile Ad Hoc Networking: Research, Trends and Applications, nr. 5, vol. 2, p. 483-502, 2002*, 2002.

[4] G. Endruweit and G. T. (Eds.). *Woerterbuch der Soziologie.* F. Enke Verlag, Stuttgart, Germany, 1989.

[5] R. R. et al (Eds.). *Community Services: Lifestyle.* Eul Verlag, Germany, 2006.

[6] L. Fischer and G. Wiswede. *Foundations of Social Psychology.* Oldenbourg Verlag, Germany, 1997.

[7] G. Groh. *Ad-Hoc-Groups in Mobile Communities - Detection, Modeling and Applications.* PhD-Thesis, TU-Muenchen, 2005.

[8] G. Groh and C. Hillebrand. Lbcs – location based community services. *Tech.Report TU-Mnchen, Computer-Science*, TUM-I0408, 2004.

[9] M. Halkidi, Y. Batistakis, and M. Varzirgiannis. Cluster validity methods: Part i. *http://citeseer.nj.nec.com/534869.html (URL, Jan. 2003)*, 2003.

[10] G. Hartfiel and K. Hillmann. *Woerterbuch der Soziologie.* A. Krner Verlag, Stuttgart, Germany, 1994.

[11] G. Homans. *The Human Group.* Transaction Pub., New York, Reprint 2001 (orig. 1950).

[12] X. Hong, M. Gerla, G. Pei, and C. Chiang. A group mobility model for ad hoc wireless networks. *Proceedings of ACM/IEEE MSWiM'99, Seattle, WA, USA, Aug. 1999, pp.53-60.*, 1999.

[13] M. Koch, G. Groh, and C. Hillebrand. Mobile communities - extending online communities into the real world. *Proc. Americas Conf. on Information Systems (AMCIS2002), Dallas, TX, Aug. 2002*, 2002.

[14] C. H. O. Brakel, G. Groh and A. Tasch. Privacy-einstellungen bei lokalisierungsdiensten. *Proc. MKWI04, Essen, Germany*, March 2004.

[15] T. Runkler and J. Bezdek. Alternating cluster estimation: A new tool for clustering and function approximation. *IEEE Transactions on Fuzzy Systems, Vol. 7 Nr. 4 Aug 1999, pp. 377-393*, 1999.

[16] S. Wassermann and K. Faust. *Social Network Analysis. Methods and Applications.* Cambridge University Press, 1994.

[17] Z. Wu and M. Palmer. Verb semantics and lexical selection. *Proc. 32nd Ann. Meeting of the Association of Computational Linguistics, pp 133-138, Las Cruces, Jun. 1994*, 1994.

[18] H.-J. Zimmermann. *Fuzzy Set Theory - and Its Applications.* Kluwer Academic Publishers, 1991.