Mining Groups Stability in Ubiquitous and Social Environments: Communities, Classes, Clusters

Mark Kibanov
Knowledge & Data Engineering Group
ITeG Research Center
University of Kassel
Wilhelmshöher Allee 73
34121 Kassel, Germany
kibanov@cs.uni-kassel.de

ABSTRACT

Ubiquitous Computing is an emerging research area of computer science. Similarly, social network analysis and mining became very important in the last years. We aim to combine these two research areas to explore the nature of processes happening around users.

The presented research focuses on exploring and analyzing different groups of persons or entities (communities, clusters and classes), their stability and semantics. An example of ubiquitous social data are social networks captured during scientific conferences using face-to-face RFID proximity tags. Another example of ubiquitous data is crowd-generated environmental sensor data. In this paper we generalize various problems connected to these and further datasets and consider them as a task for measuring group stability. Group stability can be used to improve state-of-the-art methods to analyze data. We also aim to improve the performance of different data mining algorithms, e. g., by better handling of data with a skewed density distribution. We describe significant results some experiments that show how the presented approach can be applied and discuss the planned experiments.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Applications—*Data Mining*; G.3 [**Mathematics of Computing**]: Probability and Statistics—*Statistical Computing*; J.4 [**Computer Applications**]: Social and Behavioral Science—*Sociology*

General Terms

Algorithms

Keywords

Data Mining, Ubiquitous Systems, Physical Computing

WSDM 2015, February 2–6, 2015, Shanghai, China.

http://dx.doi.org/10.1145/2684822.2697034.

1. INTRODUCTION

Ubiquitous computing as it was described by by Weiser [21] is becoming more real: the technologies become "integral, invisible part of people's lives". Such seamless integration is enabled by different factors: small sensors, affordable smartphones and advanced algorithms. On the other side social networks gained popularity and became an important research topic for different disciplines. The combination of these two areas may create new applications and is an interesting and challenging research topic.

Ubicon (cf., Section 3.1) is an example of a platform which aims to combine both ubiquitous and social computing utilizing RFID-technology, developed by the Sociopatterns initiative[1], and environmental sensors, utilized by EveryAware project[2]. Such applications will need in the future improved data mining techniques and machine learning algorithms to provide more support for the users in their everyday-activities. Furthermore these technologies can be applied in different scenarios – e. g., RFID tags can be used to implement idea of the smart university [4] or in healthcare settings [5].

One of the main data mining principles is finding similar elements in order to generalize knowledge about them. In my ongoing PhD thesis I concentrate on stability of different groups of elements in ubiquitous environments and social networks. We investigate the possibility of improvement of existing algorithms and better predictive analysis. As the social and ubiquitous systems have very diverse data, a group can be defined as community, cluster, class or even a group of nearly placed elements depending on context of data and application type.

The contributions of the presented thesis can be summarized as following:

1. **Definition of groups in social and ubiquitous environments and modeling their stability as a general problem;**
2. **Analysis and evaluation of groups semantics and their stability;**
3. **Application of the groups stability for data mining algorithms and data analysis;**

To the best of the author's knowledge, such analysis has not been addressed before.

[1] `http://www.sociopatterns.org`
[2] `http://www.everyaware.eu`

The rest of the paper is structured as follows. We give basic descriptions and definitions of groups, describe the methodology and give short overview about related work in Section 2. Section 3 provides a description of the datasets which are used for the experiments. The evaluation of the first results and the description of the future experiments is described in Section 4. Finally, Section 5 concludes the paper with a short summary.

2. BACKGROUND AND RELATED WORK

I give a short overview of the different types of the groups considered in the presented PhD thesis and define the task of measuring their stability and finding semantics. Moreover, this section contains key references which are relevant for the presented thesis.

2.1 Community Detection in Graphs

Graph is the most common mathematical representation of the social network. An (undirected) *graph* $G = (V, E)$ is an ordered pair, consisting of a finite set V containing the *vertices/nodes*, and a set E of *edges/connections* between the vertices. A weighted graph is a graph $G = (V, E)$ together with a function $w : E \to \mathbb{R}^+$ that assigns a positive weight to each edge. The larger is the weight of the edge (u, v) the stronger is connection between nodes u and v. The concept of a community intuitively describes a group of vertices out of all nodes of graph $C \subseteq V$ such that members of C are strongly "related" among each other but sparsely "related" to nodes outside of C. We consider non-overlapping communities, where each node belongs to exactly one community, so $C_1 \cup \ldots \cup C_k = V$ and $C_1 \cap \ldots \cap C_k = \{\}$, where $C_1 \ldots C_k$ are the detected communities, are true.

As a community is a concept, without strict definition, there are different algorithms which compute communities in graphs, e. g., *InfoMap* [16]. Fortunato et al. made a comparative analysis of community detection algorithms [9, 10], Papadopoulos et al. considered particularly community detection for social networks [15]. Many authors considered evolution of communities [19, 20].

2.2 Classification Problem

Classification is an important problem for many applications based on the data generated by ubiquitous computing applications because the data collected and used by such applications is often sparse and the lacking information needs to be recovered (predicted with given accuracy or confidence). We define the problem of classification as follows. Each point $p \in P$, from a set of points P, can be assigned to a class $x \in X$, from a finite set of classes X, by a classification function $c : P \to X$. In general, the classification function c is unknown. Thus, given a subset of all classified points $C_{train} \subset C$, the goal is to approximate c by a classification function $\hat{c}$. The most common measures for quality of the classification function (or algorithm) are *Precision*, *Recall* and *F1-Measure* which can be computed for each class x. The *Accuracy* and AUC are the measures which can be applied to measure the quality of classifier for the whole dataset.

Naive Bayes, *decision trees* and *k Nearest Neighbor (kNN)* are prominent and well-known families of classification algorithms. *Support Vector Machine* is a newer classification algorithm [7] which gained popularity during the last years. In the proposed PhD thesis I will try to develop methods to improve classification, particularly by changing from kNN algorithm to adaptive kNN algorithm. In this context, e. g., Ougiaroglou et al. [14] and Sun and Huang [18] used variable adaptive k to improve the algorithm.

2.3 Clustering

Clustering is another prominent data mining task. Cluster is a group of similar objects but members of different clusters should be not similar. Given a dataset consisting of different items (objects) D in a metric space. Cluster is a set of objects $Cl \subseteq D$. Similarly to communities (cf. Section 2.1), we consider non-overlapping clusters, each object belongs to exactly one cluster. $Cl_1 \cup \ldots \cup Cl_k = D$ and $Cl_1 \cap \ldots \cap Cl_k = \{\}$, where $Cl_1 \ldots Cl_k$ are the detected clusters, are true.

To determine similarity of two elements, a distance function should be used. The choice of the distance function depends on dataset and task. The most common functions are Euclidean Distance, Manhattan Distance and Max Distance. Unlike in graph, it is usually possible to compute distance between any two elements in a metric space (some nodes in graph are not connected and thus distance between them is not known). In contrast to classification problem, clustering algorithms are not supervised.

Clustering is widely discussed topic in data mining. There are different types of clustering methods: density-based (e. g., DBSCAN [8]), hierarchical (e. g., SLINK [17]), centroid-based (e. g., k-means [13]) and other methods. Xu and II [22] made an overview of different clustering algorithms.

2.4 Groups Stability and Semantic

We define the problem of group stability as follows. Given elements $el \in I$, where I is a set of objects. Each element belongs exactly to one group G_i: $el \in G_i$. The group can be defined, for instance, as cluster, class or community.

We consider two time points t_1 and t_2. The dataset I has a particular group structure at t_1 and t_2: $G_1^{t_1} \ldots G_n^{t_1}$ and $G_1^{t_2} \ldots G_m^{t_2}$. Consider two elements from the dataset which belong to the same group at t_1: $u, v \in G_i^{t_1}$. The group is more stable the larger is the relative frequency that u and v belong to the same group at t_2.

To understand group semantic we consider any feature, set of features or dimension instead of time: e. g., consider multi-layer network. Each layer can be represented as a graph which has the same nodes but different edges. So instead of groups in different time points we consider groups in different graphs $Graph_1$ and $Graph_2$: $G_1^{Graph_1} \ldots G_n^{Graph_1}$ and $G_1^{Graph_2} \ldots G_m^{Graph_2}$. So we can find implications of connections in different layers and thus explore semantics of the networks.

3. DATASETS

As mentioned below, we used the datasets collected using Ubicon-based applications and plan experiments with further datasets. In this section, I shortly introduce the Ubicon platform and describe collected data.

3.1 Ubicon

Ubicon[3] is an open source platform[4] which provides an extensible framework for building and hosting applications targeting both ubiquitous and social environments [2, 3].

[3] `http://ubicon.eu`

[4] `http://code.ubicon.eu`

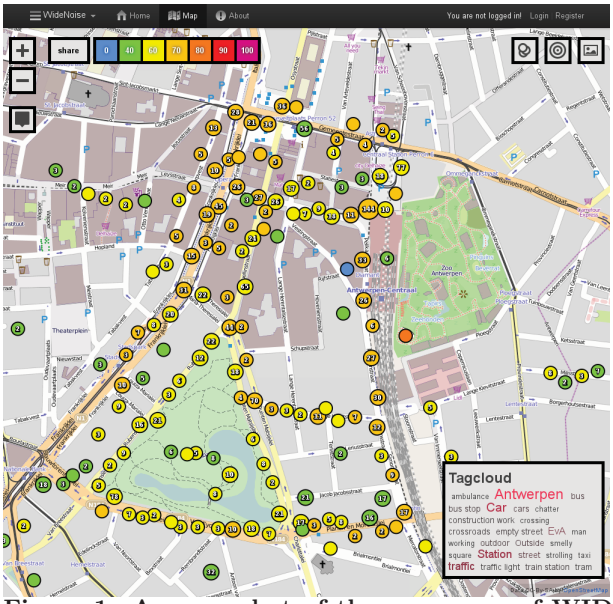

Figure 1: A screenshot of the map page of WIDENOISE [3]

From a data-centric view, Ubicon implements a data storage, processing, and serving pipeline. Ubicon hosts different applications, such as MyGroup, Conferator (which utilize RFID proximity tags to support workgroup members or conference participants), WideNoise and AirProbe (which utilize environmental sensors to enable participatory environmental monitoring). The Ubicon platform combines applications which use and collect both ubiquitous and social data. We present data collected using these applications.

3.2 Face-to-face Proximity Networks

We used RFID proximity tags to detect reliable face-to-face contacts. This technology was created by Sociopatterns initiative. The first experiments utilizing this technology were made during the *ESWC 2009* conference [1].

MyGroup and Conferator applications use the described technology to enhance users' experience during conferences and during their everyday workflows in the workgroups.

In our experiments we concentrate on the five datasets collected during the following conferences: LWA 2010, LWA 2011, LWA 2012, Hypertext 2011 and Informatik 2013. We asked participants of these conferences to wear RFID tags during the conferences. All these computer science conferences have different nature (e. g., in terms of size national\international context), so it is especially interesting to compare social interactions during these conferences. We also have been collecting face-to-face contacts data in the "Knowledge and Data Engineering" (*KDE*) research group of the University of Kassel for over two years.

Additional information about conference participants (institute affiliation, gender, country and position) and research group members (publication coauthorship, project cooperations) is available.

3.3 Environmental Sensor Data

Sensor data is another example of ubiquitous data. Using Ubicon-based applications WideNoise and AirProbe, users are able to collect environmental sensor data. The both applications were developed during the EveryAware project. The project created a set of tools for acquisition of sensor and subjective data so that different sensor data can be collected using smartphone embedded and pluggable sensors and subjective data can be input by users [6].

Smartphone application WideNoise Plus lets users to make noise level observations using embedded sensors, tag them, label with subjective perceptions and share these measurements in online social networks. A regular observation is made over 5 seconds with a sampling rate of 0.5 seconds. The location where the measurement was made can be obtained using smartphone GPS module or IP address (if GPS not available). The collected data is available for users online. Till now more than fifty thousand noise measurements were made by users all over the world.

The second EveryAware application – AirProbe – aims measuring air quality, and particularly black carbon concentration in the air. The AirProbe smartphone application and low-cost sensorbox were developed to enable such measurements. The sensorboxes can be connected to the smartphone via bluetooth. These sensorboxes may continuously measure the air quality while the users may tag interesting places or events using the AirProbe application. The majority of data was collected during AirProbe International Challenge case studies in Antwerp, Kassel, London and Turin where users used the tools for two weeks to collect as much data as possible.

The described datasets are examples of the ubiquitous and social networks data. They were collected by Ubicon-based applications. I am planning to use further datasets for the future experiments. The exact list of the datasets is not fix yet, but the group detection algorithms presented in Section 2 should be applicable to these datasets.

4. METHODOLOGY AND EVALUATION

General approach described in Section 2 should be justified for each dataset in order to reach the research goals (cf., Section 1), to improve specific applications and afterwards to generalize the results and make conclusions about groups stability and semantics in ubiquitous environments.

4.1 Community Stability in Face-to-Face Proximity Networks

Community detection is one of the most common approaches of finding groups of densely connected elements in graphs. We applied different community detection algorithms – InfoMap, Label Propagation, Leading Eigenvector, Walktrap, Edge Betweenness and Fast Greedy – to different conferences to determine if the stability of communities is dependent on chosen algorithms or conferences [11, 12].

To analyze the stability of the community structure we define a Community-pair (*c-pair*) as follows: if two nodes u and v belong to the same community, then $cp = (u, v)$ is a *c-pair*. To estimate and compare the stability of communities during different conferences, we compare the *c-pairs* that were formed during different days (e. g., at timepoints t_1 and t_2). Considered community detection algorithms show similar performance with respect to the community stability for these conferences. An interesting observation is that the active communication does not make communities stable – even vice versa. On hypothesis for explaining the negative correlation of community stability and communication is the

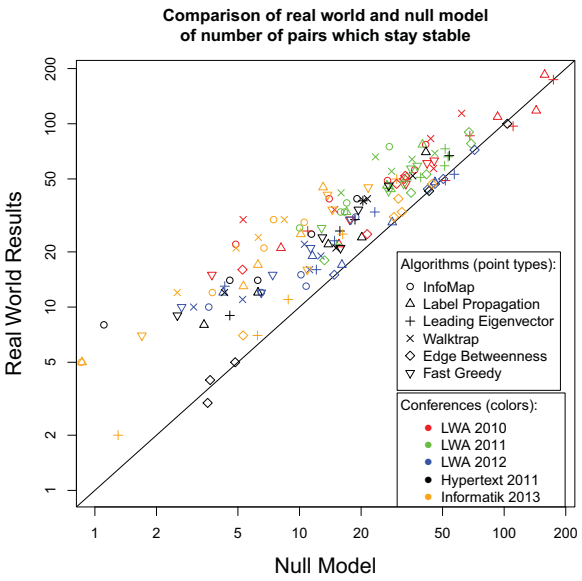

Figure 2: Comparison of the real community stability with the null model of the considered algorithms (marked by the different point types) and conferences (marked by different colors): the x-axis contains the respective null model values, the y-axis contains the respective real values. Both axes are scaled logarithmically.

following: the participants tend to stick to the known persons and tend to have less contacts with new persons that implies both lack of new contacts and stability of the existing communities over the whole conferences.

To clarify the significance of stability, we compared them to a null model and particularly the F1 score of the real data and the null model (cf., Figure 2). The F1 values computed from the real data are almost in every case larger than from the null model. On average, the real world F1 score is twice larger than the obtained null model F1 score.

4.2 Groups in Offline and Online Social Networks

We demonstrated some technical analysis in previous section. Such analysis can give insights about human interactions and improve recommender systems. However, for deep understanding of the nature of the human interactions it is necessary to perform further analysis. Here I give an overview of planned experiments.

Datasets collected using Conferator system show one of the layers of the scientific networks: face-to-face interactions during the conference. However these networks may have more different layers, such as:

- Coauthorship network
- Online professional networks (LinkedIn, ResearchGate)
- Online personal networks (Facebook)
- WWW networks with homepages of the scientists as nodes

An additional information about participants of the experiments is also available (for example, their institutions, conference status, academic status, research area). It is also possible to compute similarity between participants (e. g., based on papers similarity or homepage content similarity).

Comparison of listed layers and properties can be a key for understanding of the nature of human communication and cooperation. Such analysis corresponds with the research goals of the presented PhD thesis: we measure stability of groups, e. g., if the groups (communities) stay the same in different layers of the network; search for semantics of groups (e. g., by comparing which factors have more influence on emerging edges); improve data mining techniques (e. g., by adding additional information to recommender algorithms).

Further interesting topic is the possibility of finding one particular group from the offline world in different online networks (e. g., finding participants of one conference in DBLP coauthorship graph or ResearchGate Network).

4.3 Classification of Noise Measurements

We consider classification problem for sensor data and particularly the task is to improve the performance of data-mining algorithms based on group stability. We plan to use item-centered groups for this task: k nearest neighbors are the members of the same group. The task is finding k so that stability of the group in terms of class equality of its members is maximal. Basically, this is an idea of the standard k Nearest Neighbor (kNN) algorithm.

Data collected using ubiquitous applications is crowd- generated and thus has some problems: it is usually almost impossible to control how, where and when the data will be generated. Datasets may be sparse and have skewed density distribution of the data points, thus application of k Nearest Neighbor algorithm may not deliver constant performance. To solve this issue, we try to find an optimal k for the each item to maximize the accuracy of the algorithm.

In the first experiments, the algorithm showed a better performance in terms of accuracy. We are working on improvement of the runtime of the algorithm. We plan to generalize the approach in the future to be able to estimate the group stability based on similarity in general, not only for k Nearest Neighbors algorithm.

4.4 Clustering for origin detection

As described in Section 2.3, the elements of one cluster are elements located near each other in the n-dimensional metric space. Clustering may be useful if a group of similar items needs to be found: for instance, for origin detection of items. I am planning to setup the experiments to explore how the clustering can be used to identify the origin of the item. Considering Widenoise dataset this means finding the user who made current measurement using only features of the measurements (not knowing device ID).

A number of special research questions are connected to this task: *which clustering algorithm fits the best for this purpose? How the features for clustering should be chosen and weighted? What are the best and the worst accuracies of this approach?* These experiments are however not limited to research questions listed above but also lead to general research aims discussed in Section 1: stability of clusters may play a crucial role in accuracy estimation for the algorithm as more stable clusters may deliver more reliable detection of its items origin. Should the approach deliver reliable results, it will show that clusters have semantics: its origin.

The described problem may have a lot of practical applications: restoring of corrupt information, deep analysis of different data, verification of privacy policy. These experiments will be done with different datasets.

5. CONCLUSIONS AND FUTURE WORK

This paper gives a short overview of ongoing PhD thesis which considers mining stability of groups in social and ubiquitous computing. I gave a general motivation for this problem and defined different types of groups (communities, classes, clusters). I described the datasets I used for my experiments. These were collected using Ubicon platform which was shortly overviewed as well. The first experiments and their evaluation were presented in this paper as well.

It is essential for future work to generalize the results and build a unified model for the stability of groups in ubiquitous environments. Also, definition of group can be extended by considering subgroups discovery. Furthermore, I am planning to work with further datasets and to implement further experiments. Moreover I hope to be able to show the proposed general model can be applied to solve different problems.

Acknowledgement

This publication was supported by Kurt-Pauli-Foundation. I appreciate the help of Prof. Dr. Gerd Stumme and Dr. PD Martin Atzmueller who help me to work on my thesis. I appreciate also Martin Becker and Juergen Mueller who assisted me in writing papers and give me important feedback.

6. REFERENCES

[1] H. Alani, M. Szomszor, C. Cattuto, W. Van den Broeck, G. Correndo, and A. Barrat. Live social semantics. In *The Semantic Web - ISWC 2009*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2009.

[2] M. Atzmueller, M. Becker, S. Doerfel, M. Kibanov, A. Hotho, B.-E. Macek, F. Mitzlaff, J. Mueller, C. Scholz, and G. Stumme. Ubicon: Observing social and physical activities. In *Proc of CPSCom 2012*, pages 317–324. IEEE Computer Society, Nov. 2012.

[3] M. Atzmueller, M. Becker, M. Kibanov, C. Scholz, S. Doerfel, A. Hotho, B.-E. Macek, F. Mitzlaff, J. Mueller, and G. Stumme. Ubicon and its applications for ubiquitous social computing. *New Rev. of Hyper- and Multimedia*, 20(1), Mar. 2014.

[4] M. Atzmueller, S. Bobek, M. Kibanov, and G. J. Nalepa. Towards the ambient classroom: An environment for enhancing collaborative educational processes. In T. Roth-Berghofer, S. Oussena, and M. Atzmueller, editors, *Proceedings of the Smart University Workshop, SmartUni 2013 – International and Interdisciplinary Conference on Modeling and Using Context, Context 2013, Annecy, Haute-Savoie, France, 28 October, 2013*, 2013.

[5] A. Barrat, C. Cattuto, V. Colizza, L. Isella, C. Rizzo, A. E. Tozzi, and W. Van den Broeck. Wearable sensor networks for measuring face-to-face contact patterns in healthcare settings. In M. Szomszor and P. Kostkova, editors, *Proceedings of the 3rd International ICST Conference on Electronic Healthcare for the 21st century (eHealth 2010)*, 2011.

[6] M. Becker, A. Hotho, J. Mueller, M. Kibanov, M. Atzmueller, and G. Stumme. Subjective vs. objective data: Bridging the gap. CSSWS 2014, Poster, 2014.

[7] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[8] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc of KDD 1996*, pages 226–231. AAAI Press, 1996.

[9] S. Fortunato and C. Castellano. Community structure in graphs. In R. A. Meyers, editor, *Encyclopedia of Complexity and Systems Science*, pages 1141–1163. Springer, 2009.

[10] S. Fortunato and A. Lancichinetti. Community detection algorithms: a comparative analysis: invited presentation, extended abstract. In G. Stea, J. Mairesse, and J. Mendes, editors, *VALUETOOLS*, page 27. ACM, 2009.

[11] M. Kibanov, M. Atzmueller, C. Scholz, and G. Stumme. On the evolution of contacts and communities in networks of face-to-face proximity. In *Proc of CPSCom 2013*, pages 993–1000. IEEE Computer Society, 2013.

[12] M. Kibanov, M. Atzmueller, C. Scholz, and G. Stumme. Temporal evolution of contacts and communities in networks of face-to-face human interactions. *Science China Information Sciences*, 57(3), Feb. 2014.

[13] S. P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–136, 1982.

[14] S. Ougiaroglou, A. Nanopoulos, A. N. Papadopoulos, Y. Manolopoulos, and T. Welzer-Druzovec. Adaptive k-nearest-neighbor classification using a dynamic number of nearest neighbors. In *Proc of ADBIS 2007*, pages 66–82. Springer, 2007.

[15] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos. Community detection in social media - performance and application considerations. *Data Min. Knowl. Discov.*, 24(3):515–554, 2012.

[16] M. Rosvall, D. Axelsson, and C. T. Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23, 2009.

[17] R. Sibson. Slink: An optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1):30–34, 1973.

[18] S. Sun and R. Huang. An adaptive k-nearest neighbor algorithm. In *Proc of FSKD 2010*, pages 91–94. IEEE, 2010.

[19] H. Sundaram, Y.-R. Lin, M. De Choudhury, and A. Kelliher. Understanding community dynamics in online social networks: A multidisciplinary review. *Signal Processing Magazine, IEEE*, 29(2):33 –40, Mar. 2012.

[20] M. Toyoda and M. Kitsuregawa. Extracting evolution of web communities from a series of web archives. In *Hypertext 2003*, pages 28–37. ACM, 2003.

[21] M. Weiser. The computer for the 21st century. *Scientific American*, 265:94–104, 1991.

[22] R. Xu and D. C. W. II. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.