Analysis of User Roles and the Emergence of Themes in Discussion Forums

Tobias Hecking, Irene-Angelica Chounta, H. Ulrich Hoppe
University of Duisburg-Essen
Department of Computer Science
and Applied Cognitive Science
Duisburg, Germany
{hecking,chounta,hoppe}@collide.info

***Abstract*— This work explores network analysis methods for the analysis emergent themes as well as types of users in discussion forums. The paper provides both, a description of the analysis approach and its application as a case study. To that end, keywords are extracted from forum threads and then linked the users in the forum thread resulting in a bipartite network. Applying bipartite clustering on those networks, both, groups of users with common interest in themes can be identified as well as groups of related keywords based on their common relations to users. As a case study, the approach is applied to a discussion forum of a Coursera MOOC. The results reveal some interesting patterns and phenomena of thematic development that take place in such large scale learning courses.**

Keywords—Bipartite Networks;Community Detection;Thematic development;Discussion Forums;MOOCs

I. INTRODUCTION

Discussion forums are a common means for enabling asynchronous information exchange on the web. The content of those forums can grow very large and can be of different nature, ranging from question-answer discussions to coordination and socializing among users. Especially in the case of the recently emerged, massive open online courses (MOOCs), discussion forums play an important role. In such courses, the discussion forums are often the only possible channel for communication and knowledge exchange between participants as well as between participants and the course staff. The discussion topics range from technical issues support to peer help and social conversations [1].

In order to understand the processes and underlying mechanisms in such online discussion forums, information has to be extracted on different levels. On the level of individuals, the challenge is to model roles of contributors and to identify the important users who are indispensable for the cohesiveness of the community while, on a more abstract level, the content of the forum discussions and the patterns of user contribution is of particular interest. Since online discussions can be very dynamic, methods for analysing thematic development are needed. Probabilistic topic models like Latent Dirichilet Allocation [2] can be used to model topics in texts (in our case forum threads). However, further insight can be acquired if relations among keywords and interest user groups as well as their evolution over time are extracted from the data. To that end, we present an approach for clustering bipartite networks of forum users and keywords extracted from the forum threads. A bipartite cluster of users and keywords can be interpreted as a group of users with common interests and a group of keywords that are related since they are densely connected to a common set of users. The evolution of those clusters can then be tracked over time, and thus, enabling the identification of evolutionary events like merging, splitting, and continuing of interest groups and keyword groups. Furthermore, it is shown that the proposed framework should not be considered as a substitute of probabilistic topic models but rather as an additional analysis method for thematic development. All results are presented along a case study on an anonymized forum dataset of the Coursera course titled "Global Warming: The Science and Modeling of Climate Change" provided by the University of Chicago (See [1] for data description).

The paper is structured as follows: Section II gives an overview of the related work in discussion forum analysis in and methods for analysing emergent themes from textual data that are relevant for this work in general. Section III outlines the methodological foundation of this work. The case study on the mentioned Coursera MOOC forum is presented in Section IV. A more formal evaluation and comparison with existing methods follows in Section V. Finally, Section VI concludes the main findings of this work and gives an outlook of possible further work.

II. RELATED WORK

The analysis of thematic development in online communities has been an active research topic in the recent years. One of the first studies on online mass communication by Whittaker et al. [3] described dependencies between different properties of Usenet groups such as thread depth, message length, and demographics. Later the roles of users and knowledge diffusion processes were investigated in more detail [4, 5]. However, these kinds of studies did not yet incorporate text mining of the content of user contributions. This has been done later to solve tasks like post classification, and discussion disentanglement. Especially in the case of MOOC discussion forums it is of huge interest to identify content related threads in which exchange of knowledge between participants takes place [1, 6] as well as the estimation of discussion quality [7].

Probabilistic models of thematic development or topics in online discussions most methods rely on the principle of Latent Dirichilet Allocation (LDA) [2]. Dynamics is especially taken into account in dynamic topic models (DTM) [8] and authors' relations to topics in author-topic-models (ATM) [9]. The combination of both in one model is presented by Xu et al. [10] which comes closest to the idea of this paper. However, it still restricted to a fixed number of topics and interest groups and does not allow tracing the full history of thematically related keywords and interest groups of users simultaneously.

Apart from probabilistic topic analysis, network based methods have as well been applied for discourse analysis, for example, in the context of communication in organisations or scientometrics [11, 12]. Network analysis has found to be appropriate for modelling thematic dynamics in online discussions as well. Introne and Drescher [13] applied the clique percolation community detection method [14] to networks of words extracted from chat messages of users who collaboratively solve a fictive criminal case. The found sub-communities of words are interpreted as topics. By applying community tracking [15, 16] - the re-identification of sub-communities across time slices, they were able to trace the life time of a topic including evolutionary events like topic splits and merges of topics (word clusters). This approach has some similarities with ours presented in Section III. However, in instead of extracting word-to-word networks from the forum posts, our approach links thread keywords to users who are active posters in the thread. The clustering of those bipartite networks does not only allow finding groups of related keywords but also groups of users with a common interest as well as their evolution over time. This approach of modelling artefacts related to the users who use them rather than the direct induction of relations between keywords or users has successfully been applied in social media analytics [17] and scientometrics [18]. A previous step into this direction was presented by Lipizzi et al. [19]. In their work bipartite networks of users linked to keywords extracted from their Twitter tweets are modelled based on tweets on certain products. Different to the work presented in this paper, for further analysis they project the user-keyword relations into a unipartite network of keywords which loses the information about the users. These concept networks are then analysed in terms of cohesiveness and sentiment.

III. Methodology

A. Network Extraction

For the purpose of the study, we build time slices of the evolving network of users and keywords corresponding to the forum activity in each week of the course. In each time slice there is a set of active users *U* posting to a subset of threads *T*. The first step is to extract a set of keywords *K* from the active threads in the corresponding period. This is done by aggregating the posts within one thread to a single document and application of the keyword extraction algorithm provided by Alchemy API[1]. This algorithm computes a ranked list of keywords with relevance values ranging between 0 and 1. In this work we used keywords with relevance above 0.8. The result can be modelled as a bipartite network where threads are represented as nodes of one type linked to keywords as nodes of the other type. Its adjacency matrix A_{TK} is of the following form:

$$\begin{bmatrix} 0 & B_{TK} \\ B^T{}_{TK} & 0 \end{bmatrix} \quad B_{TK} \in \{0,1\}^{|T|\times|K|} \tag{1}$$

In the same manner, another bipartite network of users and forum threads can be build where users are linked to threads they posted in during the course week. The resulting adjacency matrix A_{UT} has the form:

$$\begin{bmatrix} 0 & B_{UT} \\ B^T{}_{UT} & 0 \end{bmatrix} \quad B_{UT} \in \{0,1\}^{|U|\times|T|} \tag{2}$$

If the rows of B_{TK} and the columns of B_{UT} are in the same order, the adjacency matrix A_{UK} of the final user – keyword network can be calculated as:

$$A_{UK} = \begin{bmatrix} 0 & B_{UK} \\ B^T{}_{UK} & 0 \end{bmatrix}, B_{UK} = B_{UT} \times B_{TK} \tag{3}$$

B. Network Clustering

The next step is to discover bipartite clusters of closely related users and keywords. A cluster of users and keywords can be seen as a group of users with a common thematic interest. In case of discussion forums users who post in the same thread automatically form such an interest group since they form a biclique with all the keywords of the thread, provided that the thread has enough keywords. However, defining interest groups only on the thread level might give an incomplete picture. Therefore, the biclique percolation method [20] is appropriate to discover interest groups of people in multiple threads based on common keywords. Figure 1 gives an example of a bipartite cluster which is not restricted to authors and keywords from only one forum thread. The biclique percolation method has two further advantages for the task of clustering users and keywords simultaneously. First, the method allows overlaps between the clusters. This property is essential since it is likely that a keyword can be used in different contexts. The same is true for users. A user can be part of more than one interest group. Second, the method only assigns users and keywords to clusters if they are part of a biclique with *a* nodes of the one mode and *b* nodes of the second type. This can be considered as an inherent filtering procedure that helps to focus on the important parts of the network. Keywords and users who are not well connected within the network are not part of the final clustering result.

The principle of biclique percolation adapts the well-known clique percolation method to bipartite networks where no cliques in the original sense are present. The method relies on the definition of a $K_{a,b}$ biclique. This is a maximal connected bipartite subgraph with a nodes of the first mode and b nodes of the second mode. Thus, if a set of a actors all are connected to each of b resources, they form a $K_{a,b}$ biclique. A bipartite subgroup also called biclique community is defined as the union of a series of adjacent $K_{a,b}$ bicliques. Two $K_{a,b}$ cliques are considered adjacent if they share at least *a-1* nodes of the one type and *b-1* nodes of the other (see Figure 1).

[1] http://www.alchemyapi.com/

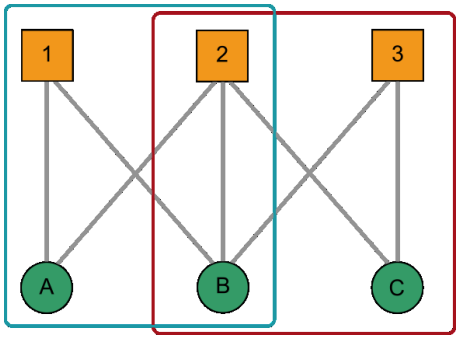

Fig. 1. Example of two adjacent $K_{2,2}$ bicliques. First clique: {A, B, 1, 2}, second clique: {B, C, 2, 3} (Source [24]).

C. Thematic dynamics

As described in [21], in a static snapshot of an evolving bipartite network both parts of a cluster can be considered at once. However, regarding the evolution of the network these two parts should be considered separately. It is possible to re-identify the keyword group or the user group of a bipartite cluster across time slices. An example is given in Figure 2.

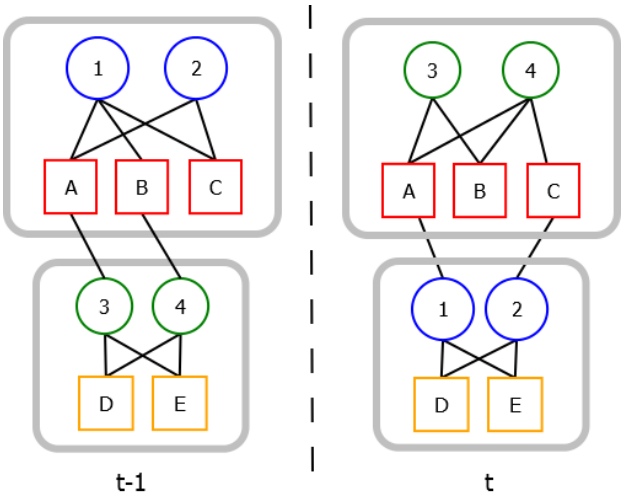

Fig. 2. Example of independent evolution of the two parts of bipartite clusters. The particular parts of the clusters in t-1 persist over time but in different clusters. (Source: [21])

This analytical approach can also be applied to the user-keyword networks that are extracted from successive time slices, as described before. Thus, it can provide insight regarding the analysis of thematic dynamics in discussion forums. The assumption is that keywords which occur in one bipartite cluster can be interpreted as semantically related, since they are densely related to a subset of the users in the network. Furthermore, the users of such a cluster can be considered as an interest group with common relations to a subset of concepts. Such a mixed cluster of users and concepts not necessarily correspond to a single thread where all users in a thread are linked to all keywords assigned to the thread, as explained in Section III.A, but may contain users and keywords of different threads. This allows finding keywords that belong to a certain thematic area and subgroups of users who are interested in this area, but also to trace the whole history of thematic areas and interest groups by applying community matching [22, 23]. A group of keywords in a bipartite cluster at a given time slice *t* can sometimes be matched to groups of keywords in a time slice *t-x* by different measures of similarity. The same is true for groups of users. Since there is not always a one-to-one matching between two groups in two different time slices of the network, it becomes also possible to identify splitting and merging topics (groups of keywords) or interest groups (groups of users) solely based on the changes in the user-keyword network. The advantage of the bipartite clustering approach to other dynamic topic modelling approaches is that the model completely maintains the relation of users and keywords in bipartite user-keyword clusters. This enables to track the shifts of interests of certain groups of users simultaneously.

In the case of bipartite clusters, the re-identification of parts of clusters is more complicated as in unipartite networks since the two groups of nodes in such clusters can evolve independently (Figure 2), and therefore, the matching procedure has to be applied to both parts of a bipartite cluster separately. There are different approaches to match groups of nodes in a dynamic network over time. The work of Bródka et al. [22] gives a good overview and evaluation. In this work we chose the inclusion measure. Its value indicates to what extend a smaller group is contained in a larger group and can be calculated as:

$$sim(g1, g2) = \frac{|g1 \cap g2|}{\min(|g1|, |g2|)} \quad (4)$$

The inclusion measure is most suitable to identify merges and splits of groups. According to the approach presented by Greene and Doyle [23] and its adaptation to bipartite clusters [24] the matching of groups across time slices two sets of not matched groups are maintained. A group (keyword group or user group) is considered as not matched if there is no group with similarity above a certain threshold. The method proceeds as follows:

The procedure subsequently examines the clusters found in each time slice. As stated before, a bipartite cluster contains two groups, one keyword group and one user group. For all groups in a particular time slice *t*, the similarity to all the not matched groups in previous time slices is computed. If the similarity to one or more not matched group is above the threshold, a relation between the groups is introduced. Consequently, the matched groups from previous time slices are deleted from the set of not matched groups and all the groups of the current time slice are added to this set. This has to be done for the keyword groups and the user groups of the bipartite clusters separately in order to identify evolutionary events of these groups separately. The matching threshold has been fixed to 0.75 in this study.

IV. Case Study

A. Dataset description

For this paper, we study a Coursera MOOC tiled “Global Warming: The Science and Modeling of Climate Change”. A discussion forum was used to support the communication between the users and to promote social interaction and

information exchange. The activity in the discussion forum was recorded in log files from October 21, 2013 to January 11, 2014 [1]. The discussion forum consisted of forums and sub-forums, further divided in threads consisting of posts and comments to posts. The structure of the discussion forum also reflects different thematic areas that are facilitated, such as general discussion, help on assignments, feedback on course organizational issues, etc.

Users were able to start their own threads, post in different threads and forums and also comment directly on existing posts. The MOOC participants could post or comment in the discussion forum either using a personalized user account or anonymously. Overall, we identified four different user types: students (1000 users), staff (4 users), instructors (1 user) and anonymous users. In the present study, the anonymous users were removed from the analysis since we cannot gain insight with respect to their user status (students or staff) or to personalize user activity.

The dataset consisted of 1005 participants, who created 5336 posts in 2027 threads distributed over 1874 sub-forums. In Figure 3, we present the distribution of user activity, with respect to posting, over the threads and forums of the discussion board. On average, each user posts on 3 threads (mean=2.94, σ=6.9) and 2 forums (mean=1.86, σ=2.08). As it is shown from the distribution, users do not get involved or spread over many threads and forums. This indicates limited activity that focuses on particular thematic areas without further expanding to others.

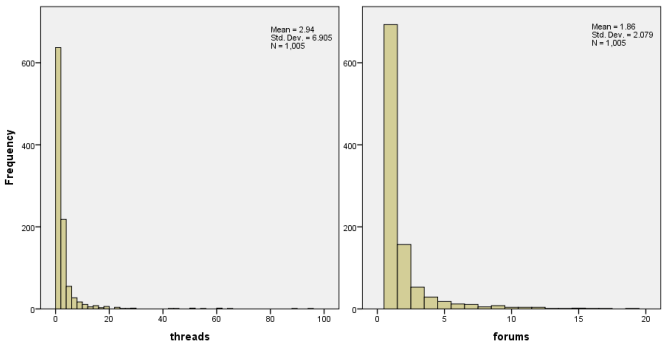

Fig. 3. Distribution of user posts over threads and forums of the discussion board for the whole duration of the MOOC.

In Figure 4, we present the distribution of posts per individual user for the whole duration of the course. On average, each user created 5 posts (mean=5.31, σ=18.86) while it was shown that the majority of users posted less than 50 posts in the discussion forum. The users of the discussion forum could vote for the posts that were created by other users by adding a +/-1. This way, the users themselves provided a rating of the quality and usefulness of posts. Overall, 3844 posts (72% of the total number of posts) received no votes, while the rest posts were rated with 0.47 votes on average. The number of posts per user, was found to correlate highly with the number of threads (ρ=0.876, $p<0.01$) and the number of forums (ρ=0.819, $p<0.01$) that the particular user had been posting. Additionally the number of posts correlated statistically significantly with the average number of votes per user (ρ=0.234, $p<0.01$). This suggests, that the users who have a high posting activity, also contribute to different thematic areas and are acknowledged as influential by the other users of the discussion forum.

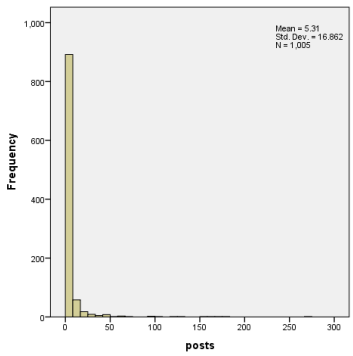

Fig. 4. Distribution of user posts over the whole duration of the MOOC course.

B. *Bipartite Clustering of Users and Keywords*

The users and keywords were clustered using the described biclique percolation method. The parameters were set to $a=b=3$. This means that a biclique should contain at least 3 keywords and 3 users. Consequently, the resulting clusters only comprise of users and keywords with at least 3 connections. A typical result is depicted in Figure 5.

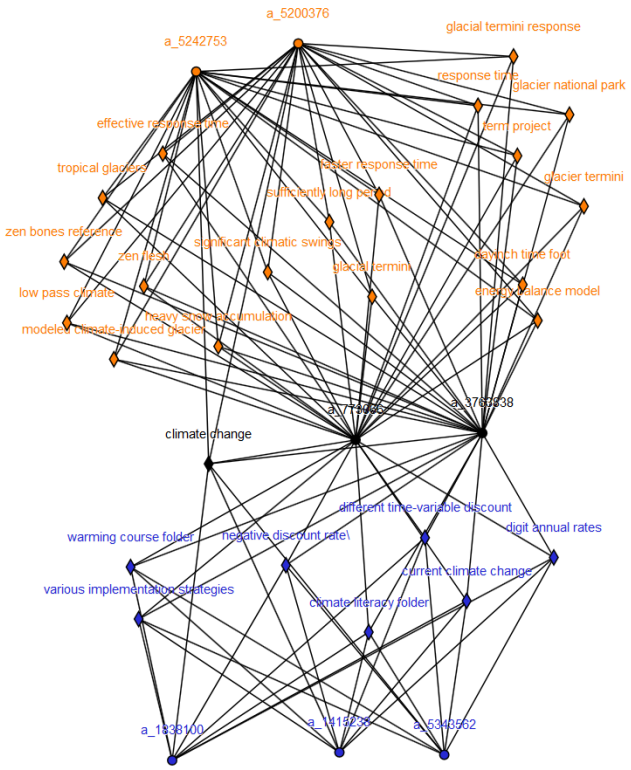

Fig. 5. Bipartite clusters of users and keywords

The depiction is an excerpt of the resulting clusters in the seventh week of the course and shows two user-keyword clusters that share a common user. The orange cluster mainly contains keywords related to glaciers and the blue cluster relates to technical terms.

C. *Bridging concepts and bridging users*

Since users can be active in different threads, and hence, can have broader thematic interests, it may occur that those users cannot be assigned to a cluster uniquely. As one would expect this is often the case for the course administration. However, there are also regular course participants who bridge between different thematic areas as the example in Figure 5 shows. On the other hand, there are also concepts that are of general importance. These concepts frequently appear in the overlap of bipartite clusters as well. Consequently, the proposed approach can be used to identify, both, concepts of general importance for the community and users who have the potential to spread information between different thematic areas. Table I depicts the concepts that occur in the overlaps of the user-keyword clusters over time most often. In addition the third column of Table I subsumes the time slices in which the corresponding node occurs. Concepts that often occur in more than one cluster at a time are of different types. For example, "Climate change", which is central to the course topic "Global warming" in general, occurs in overlap of clusters in 6 of 11 time slices. This is expected for these kinds of concepts. Furthermore, there are concepts like "Environmental Science Master" that are not directly related to the course content. Those can frequently be found in the overlap of bipartite clusters, as well. An explanation could be that topics like potential further activities after the course are discussed in many different threads and in different contexts.

TABLE I. CONCEPTS MOST FREQUENTLY IN OVERLAPS

Keyword	Present in time slices	Times in overlap
Climate change	1,2,3,5,6,7,8,9	6
Sustainabe ressources	3,4,6,7,8	3
Environmental Science Master	6,7,8	3

Since the retrieved cluster can overlap in both modes, namely concepts and users, Table II subsumes the users who occur in overlaps most often. These users can be considered to have a broader interest, and thus, are closely connected to more than one user-keyword cluster. Since MOOC discussion forums are partly moderated by course staff who answer important questions regarding different topics, it is reasonable to find course staff people most often in the overlap. However, it is interesting that we could find such people also among the regular users. These users have a broader interest and participate in more threads than others, which can as well result from superposting behaviour (c.f. Huang et al. [25]).

TABLE II. USERS MOST FREQUENTLY IN OVERLAPS

User	User type	Present in threads	Times in overlap
a_4977322	Course staff	1,2,3,4,5,6,9,10	6
a_5219940	Regular user	1,4,5,6,7	6
a_1947440	Regular user	1,2,3,6,7,8	5

For the purpose of this study, we identified the users who appear in overlaps (Group 1) and further studied their practice with respect to their posting activity during the course. Additionally we compared them to users who don't appear in overlaps (Group 2). The results for the two groups are presented in Table III.

TABLE III. USER ACTIVITY STATISTICS FOR GROUP 1 AND GROUP 2

Groups	#threads	#forums	#posts	#votes per user
Group 1	5.55	2.74	10.9	0.91
Group 2	1.7	1.45	2.65	0.26

Overall, 324 users were found to appear in the overlaps (Group 1), 321 students, 2 staff members and 1 instructor. These users posted on average 11 posts throughout the duration of the course, in 6 threads (mean=5.55) and 3 forums (mean=2.74). Moreover, the posts they created were rated on average with 0.91 votes. On the contrary, the users who did not appear in overlaps (Group 2 - 681 users) created fewer posts which were distributed in fewer threads and over fewer forums in comparison to the activity of Group 1. Furthermore, the users of Group 2 were voted lower for their posts. From this data, it can be concluded, using the Mann-Whitney U Test, that Group 1 had statistically significantly higher posting activity, more distributed over thematic areas and was rated higher than the activity of Group 2 ($p<0.01$). The distribution of user activity per Group with respect to posts and threads is presented in Figure 6.

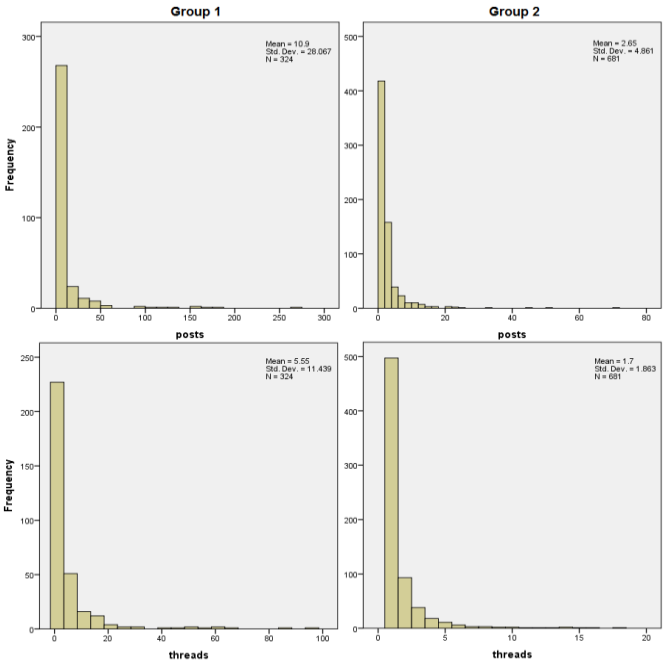

Fig. 6. Distribution of number of posts and threads for the users of Group 1 and Group 2.

D. Evlolution User-Keyword Clusters

Since user parts and keyword parts of the clusters can continue, merge, and split independently, there is a wide variety of possible evolutionary events. Some could be identified in our dataset. The most frequent events are listed in Figure 6. Event 1 in the depiction occurs most often. In these cases a topic is taken up by a different group of users. This happens often when forum threads remain active over several weeks or become active again after a period of inactivity. In this case study this is, for example, the case for threads on general discussions or keywords related to technical or organizational issues of the course.

On the contrary, interest groups of users could be identified across several time slices but with changing connections to thread keywords (event 2). This pattern is very interesting since it shows that groups of users sometimes change their interests simultaneously without necessarily having direct communication in the forum. As a prototypical example, a group of students with common connections to keywords of the area "carbon emissions" in week 4 of the course could be re-identified in week 7. However, in this week they talk about number crunching techniques for measuring climate change. An explanation could be that they first discuss more general about factors of climate change and afterwards show more interest into concrete measuring techniques.

Splitting groups of keywords (event 3 in Figure 7) could be observed occasionally. In the beginning of the course when a long thread for the general discussion of the topic "climate change" was taking place that evolve a large amount of the users. The resulting keyword group splits afterwards into three smaller keyword groups resulting from more focused discussions on special areas.

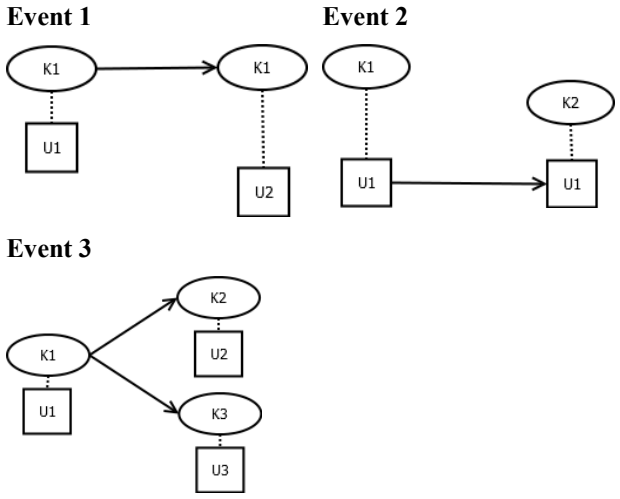

Fig. 7. Evolutionary events for discovered for the bipartite user-keyword clusters.

V. Comparison with Traditional Topic Modelling

The presented approach has some similarities with topic modelling. To some extent the keyword part of a user-keyword cluster can be interpreted as a topic since their common relations to a set of users introduce some semantic relatedness. It is very difficult to compare these two approaches quantitatively. However, for the sake of completeness Table IV depicts the similarity of topics found by LDA and keyword groups of the user-keyword clusters identified in each time slice. LDA requires the number of topics as a parameter. For comparison this parameter is set to the number of user-keyword clusters found in the corresponding time slice. The input of LDA is a document-term-matrix which elements are the number of occurrences of each term in the document. In our case a document is defined to be a single discussion thread in the forum. The term vocabulary corresponds to the keywords extracted with the Alchemy API as for the clustering approach. However, the keywords for each thread were not restricted to have a relevance value above a certain threshold as for the clustering approach. Since in LDA each topic is represented as a probability distribution of words, it is necessary to transform this distribution into a bag of words by specifying a minimum rank for words to be representative for a topic. For this evaluation this value was set to 10. The resulting bag of words and the keyword groups introduced by the clustering approach were compared in terms of inclusion similarity. It measures to what extent the smaller group of words is contained in the larger one similar to equation 4. For each keyword cluster the similarity with the best matching bag of words induced by LDA was calculated and average similarity of all pairs calculated. The problem is that a restriction to a fixed number of keywords is not possible for the clustering approach since there is no ranking of the keywords. Thus, the comparison with LDA is only limited.

The results in Table IV show clearly that the results of the user-keyword clusters are very different to the topics modelled by LDA. This huge differences result from the different interpretation of a topic. While LDA models topics as a probability distribution of words based on their occurrences in forum threads, the keywords in a bipartite cluster are the result of common relations to users.

TABLE IV. Comparison of LDA and bipartite clustering.

Time slice	*Number of topics*	*Avg. incl. similarity (LDA, clusters)*
1	30	0.19
2	21	0.175
3	20	0.135
4	9	0.01
5	12	0.08
6	11	0.2
7	10	0.15
8	11	0.29
9	11	0.29
10	7	0
11	0	1

VI. Conclusion

In this paper we explored bipartite clustering of users and keywords for the analysis of thematic development in online discussion forums. We demonstrated the potential of the

approach with the application of the method to a MOOC discussion forum. The interesting feature of co-clustering of users and keywords in a bipartite user-keyword network is that users and keywords extracted from forum threads are grouped simultaneously into mixed clusters. In such clusters, the group of users can be interpreted as a group with a common interest and introduce semantic relations among the keywords. Since the biclique percolation method [20] was applied, overlapping clusters are possible. Overlaps in the user dimension indicate users with broader interests and overlaps in the keyword dimension help to discover keywords that are important for more than one interest group.

In order to gain further insight, we applied the proposed methodology to a three-month MOOC course. In general, the participants of the course appeared to have low posting activity that focused on certain thematic areas. However, posting activity correlated to the number of threads and number of forums users were active as well as to the number of votes they received. After applying bipartite clustering of users and keywords, we were able to identify two groups of MOOC's users: those who appear in overlaps of the discovered clusters and those who do not. From the analysis of user activity of these groups, it was shown that users who appear in overlaps create more posts that also spread in more threads and forums, and thus over various thematic areas, acting as disseminators of knowledge. Additionally, these users received more votes, i.e. get higher ratings from other users of the discussion forum, indicating that they are identified as influential and important nodes. Regarding the emergence of themes in the discussion forums the found clusters were tracked over successive time slices. Different evolutionary events could be detected for, both, the keyword groups and the user groups of the bipartite clusters. One common event discovered was the persistence of a group of keywords across time slices. However, these groups of keywords were not necessarily connected to the same group of users over time. This indicates that themes are sometimes taken up by different groups of users. On the contrary the same pattern could be identified for persisting groups of users who build bipartite clusters with different groups of keywords over time. Splitting groups of keywords could be identified as well.

There are some commonalities to existing topic modelling methods. However, the presented approach is not directly comparable to these methods and should not be considered as a substitute of traditional topic modelling approaches. Most existing approaches like LDA and derivatives use probabilistic modelling. They do not allow for the tracking of topic evolution considering a topic as the result of evolutionary events like "merge", "split" of previous topics. Some methods model user relations to topics but not the dynamics of interest groups of users over time. Our approach allows for both, tracking of the history of emerging groups of related keywords as well as maintaining the relations to users. Simultaneously tracking of changing interests of user groups can also be identified as presented in section IV.D. The results in Section V have shown that the results by the bipartite clustering approach are very different compared to the results of LDA topic modelling. Hence, one should be careful with the interpretation of a cluster of users and keywords. While in probabilistic modelling a topic is a probability distribution of words based on the word's occurrences in documents, a group of words in a bipartite cluster is solely based on dense relations to a group of users with similar interests. This is a completely different view on thematic development in online discussions and the clustering approach yields different insights into the thematic dynamics.

In future work the presented approach should be applied to a wider range of online communication, like chats and Twitter. This would be a further step to understand thematic development in these areas and the identification of important users and their interests.

REFERENCES

[1] L. A. Rossi and O. Gnawali, "Language independent analysis and classification of discussion threads in coursera MOOC forums," in *Proceedings of the 15th {IEEE} International Conference on Information Reuse and Integration, {IRI} 2014, Redwood City, CA, USA, August 13-15, 2014,* 2014, pp. 654-661.

[2] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation," *J.Mach.Learn.Res.,* vol. 3, pp. 993-1022, mar, 2003.

[3] S. Whittaker, L. Terveen, W. Hill and L. Cherny, "The dynamics of mass interaction," in *From Usenet to CoWebs*Anonymous Springer, 2003, pp. 79-91.

[4] J. Zhang, M. S. Ackerman and L. Adamic, "Expertise networks in online communities: Structure and algorithms," in *Proceedings of the 16th International Conference on World Wide Web,* Banff, Alberta, Canada, 2007, pp. 221-230.

[5] L. A. Adamic, J. Zhang, E. Bakshy and M. S. Ackerman, "Knowledge sharing and yahoo answers: Everyone knows something," in *Proceedings of the 17th International Conference on World Wide Web,* 2008, pp. 665-674.

[6] Y. Cui and A. F. Wise, "Identifying content-related threads in MOOC discussion forums," in *Proceedings of the Second (2015) ACM Conference on Learning @ Scale,* Vancouver, BC, Canada, 2015, pp. 299-303.

[7] J. Kim, E. Shaw, D. Feng, C. Beal and E. Hovy, "Modeling and assessing student activities in on-line discussions," in *Proc. of the AAAI Workshop on Educational Data Mining,* 2006, .

[8] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd International Conference on Machine Learning,* 2006, pp. 113-120.

[9] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth and M. Steyvers, "Learning author-topic models from text corpora," *ACM Transactions on Information Systems (TOIS),* vol. 28, pp. 4, 2010.

[10] S. Xu, Q. Shi, X. Qiao, L. Zhu, H. Jung, S. Lee and S. Choi, "Author-Topic over Time (AToT): A Dynamic Users' Interest Model," vol. 274, pp. 239-245, 2014.

[11] J. Diesner and K. M. Carley, "Revealing social structure from texts: meta-matrix text analysis as a novel method for network text analysis," *Causal Mapping for Information Systems and Technology Research: Approaches, Advances, and Illustrations,* pp. 81-108, 2005.

[12] L. Leydesdorff and I. Hellsten, "Measuring the meaning of words in contexts: An automated analysis of controversies about'Monarch butterflies,"Frankenfoods,'and'stem cells'," *Scientometrics,* vol. 67, pp. 231-258, 2006.

[13] J. E. Introne and M. Drescher, "Analyzing the flow of knowledge in computer mediated teams," in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work,* San Antonio, Texas, USA, 2013, pp. 341-356.

[14] G. Palla, I. Derenyi, I. Farkas and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature,* vol. 435, pp. 814-818, 06/09, 2005.

[15] G. Palla, A. L. Barabasi and T. Vicsek, "Quantifying social group evolution," *Nature,* vol. 446, pp. 664-667, 2007.

[16] D. Greene, D. Doyle and P. Cunningham, "Tracking the evolution of communities in dynamic social networks," in *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference On,* 2010, pp. 176-183.

[17] K. El-Arini, M. Xu, E. B. Fox and C. Guestrin, "Representing documents through their readers," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* Chicago, Illinois, USA, 2013, pp. 14-22.

[18] E. Yan, Y. Ding and E. Jacob, "Overlaying communities and topics: an analysis on publication networks," *Scientometrics,* vol. 90, pp. 499-513, 2012.

[19] C. Lipizzi, L. Iandoli and J. E. R. Marquez, "Extracting and evaluating conversational patterns in social media: A socio-semantic analysis of customers' reactions to the launch of new products using Twitter streams," *Int. J. Inf. Manage.,* vol. 35, pp. 490, 2015.

[20] S. Lehmann, M. Schwartz and L. K. Hansen, "Biclique communities," *Phys Rev E.,* vol. 78, pp. 016108, Jul, 2008.

[21] T. Hecking, L. Steinert, T. Gohnert and H. U. Hoppe, "Incremental clustering of dynamic bipartite networks," in *Network Intelligence Conference (ENIC), 2014 European,* 2014, pp. 9-16.

[22] P. Bródka, S. Saganowski and P. Kazienko, "GED: the method for group evolution discovery in social networks," *Social Network Analysis and Mining,* vol. 3, pp. 1-14, 2013.

[23] D. Greene, D. Doyle and P. Cunningham, "Tracking the evolution of communities in dynamic social networks," in *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference On,* 2010, pp. 176-183.

[24] T. Hecking, S. Ziebarth and H. U. Hoop, "Analysis of Dynamic Resource Access Patterns in Online Courses," *Journal of Learning Analytics, JLA,* vol. 1, pp. 34-60, 2014.

[25] J. Huang, A. Dasgupta, A. Ghosh, J. Manning and M. Sanders, "Superposter behavior in MOOC forums," in *Proceedings of the First ACM Conference on Learning @ Scale Conference,* Atlanta, Georgia, USA, 2014, pp. 117-126.