Overlapping communities generation for online support forums

Fang Wang, Hakan Duman, Duong Nguyen, Simon Thompson
BT Innovate & Design
British Telecom
Adastral Park, Martlesham Heath, Ipswich, UK
fang.wang@bt.com

***Abstract*—This paper presents a novel approach of overlapping communities generation for online support forums. Different from traditional online forums that provide the most similar or relevant information to respond to a user query, the approach proposed in this paper manages online forums and provides user support based on overlapping communities. Inspired by natural societies, a forum is deemed as a complex network in which all entities (keywords, posts and user) of an online forum are grouped into a series of communities that can share members with each other. To enable this, a kind of keyword association graph is constructed based on the co-occurrences of keywords in user posts. CPM (Clique Percolation Method) is then applied to discover closely connected cliques (core clusters) in the graph. The core keyword clusters absorb related posts and users to form communities and the communities are naturally overlapping. The communities are also extended to include other un-clustered but relevant posts and users so all entities in the forum belong to at least one community. Overlapping communities in online forums provide a useful means to support various services including recommendation, alerting and profiling customer support.**

Keywords-Online Communities, Clique Percolation Method, Visualization

I. INTRODUCTION

Web-based forums provide a space for online users to share information and seek help from each other on various topics. To accomplish this goal, conventional approaches tend to find the most similar or relevant information (often posts in online forums) to respond to a user query or post [1], [2]. User queries in reality, however, are usually very complex, involving multiple aspects of concepts. The information that has the highest overall similarity therefore may not be the best or the only solution to answer a user's query, nor to satisfy all of the potential angles of interests of the user. This means that conventional approaches may fail to deliver some important useful information in different categories to satisfy user requirements from every related aspect. Therefore, this paper proposes a novel approach to managing and organizing online forums based on overlapping communities, a property increasingly recognized in several types of natural networks [3],[4]. Inspired by these natural systems, a forum is deemed as a complex network in which all entities (keywords, posts and user) may belong to multiple categories underlined by the interactive relationships of the entities.

Many approaches exist in the literature, attempting to form web communities (e.g. for e-learning systems [5],[6]) or in distributed peer-to-peer systems [7],[8],[9]. Most of them utilize hard-clustering algorithms (also termed as exclusive clustering in data mining) where each entity is assigned to a single cluster or community. However there are many situations (e.g. multiple interests of users [10]) in which an entity could reasonably be placed in more than just one community, and these situations are better addressed by *non-exclusive or overlapping clustering*.

In our previous work, we presented multi-interest communities [10] to cluster movie data in order to recommend more personalized movies for on-demand movie users. This approach worked well on movies that have clear and well defined genres and descriptive keywords, but may not be able to deal with unstructured, complex and noisy data such as those in online forums.

To address this limitation, we propose a novel data-driven approach capable of automatically generating overlapping communities for online support forums. Here, the system first extracts and processes essential semantic features of resources (keywords of posts) by carefully analyzing the posts submitted to an online forum. These keywords are then used to construct a complex graph according to keyword correlations and accordingly overlapping clusters are identified using the Clique Percolation Method. The core keyword clusters formed then absorb relevant posts and users to constitute overlapping communities. Furthermore, the communities are extended to incorporate other pertinent entities (such as keywords, posts and users) so as to widen the coverage of communities in the forum. The formed overlapping communities provide also the foundation for Cyclone [11], which is a visual environment allowing the user to explore, analyze, fine-tune the extracted communities simply by altering (e.g. adding new keywords, removing or amending) the set of keywords associated with the communities. The extracted keywords describing the communities are expected to be highly noisy, inconsistent as well as personal. Thus enabling the user to manually interact and manipulate correlations among keywords is an essential feature in reducing the amount of keyword (language) inconsistencies in the model.

In comparison to many clustering and community generation methods [12],[13], the proposed approach doesn't require any prior knowledge on explicit or implicit constrains at the number, size, shape or disjoint characteristics of target clusters.

The remaining of this paper is structured as follows. The next section describes the online forum, Hubbub which is used to explain the principles of our proposed approach and

forms the main data source for our experimental work. This section also briefly describes the Cyclone, which is used for online and real-time extracted communities' visualization and modification. Section 3 introduces and explains the proposed overlapping communities generation mechanism based on Clique Percolation Method. Section 4 discusses some experimental results on real customer data sets using the Hubbub database. Finally, section 5 concludes a summary of the contributions and future work.

II. HUBBUB AND CYCLONE

A. Hubbub

Hubbub is an online forum for automatic help service, launched for users of the BT Broadband Talk service where customers simply type their questions into an online forum and submit it. Using a set of tag extraction technologies (such as in del.icio.us, Flickr or YouTube) and data analysis technologies, Hubbub conjectures a set of search terms and retrieves relevant discussions automatically which can help the customer to find answers to his/her initial question.

If the answer is not available immediately, the customer can fine-tune the search which the system has automatically created or opt to ask members of the online community, drawing on their combined intelligence to work out the answer, if humanly possible [14]. The best person(s) capable of answering are chosen *automatically* mainly based on their past experience or interest of this topic area.

The difference of Hubbub to other existing online forums is that Hubbub doesn't require the user to understand how a forum works, i.e., how and where to search for a solution among posts. Instead, Hubbub helps to proactively deliver possible answers by bringing together people and their posts that are more likely to provide help and answer the questions to form kind of community. Hubbub is currently serving an average of 600,000 pages a month and an estimated 4,300 users each day [14]. The work described in this paper builds up on Hubbub's unique capability of automatically delivering and aiding the customers in finding the right answers to their questions and aims at improving this search in dealing with inconsistencies and organizing the forum based on overlapping communities with multiple interests.

B. Cyclone

Cyclone, as described in our previous works [11],[15], is an intelligent agent-based visual environment proving a means for the user to exploit, analyze and structure unstructured information into a more manageable form. The main strength of Cyclone is its capability to couple data mining techniques with intuitive information visualization and an adaptive learning system, creating a feedback loop between the user and the system. This enables the user to perform complex data categorization and community generation in fewer steps with less effort than a purely manual interaction, whilst allowing the output to be more accurate than if a completely autonomous technique was adopted.

In the context of this paper, Cyclone is mainly used for visualization and modification purposes allowing the user of the system to illustrate and fine-tune the extracted communities and their characteristics in real-time (such as extending communities as described in the section 3.5) simply through the modification of the keywords set (e.g. adding, extending, amending or removing keywords) associated with each community.

The extracted keywords describing the communities are expected to be highly noisy, inconstant as well as subjective. Thus, enabling the user to interact and manipulate e.g. the keywords association graph in order to reduce the amount of language inconsistencies and ambiguity that the autonomous process alone wouldn't even identify is of the greatest importance to achieve a more reliable and robust means for managing and organizing online forums such as Hubbub based on overlapping communities.

III. OVERLAPPING COMMUNITIES

A. Nomenclature

Definition 1: A community is a composite collection of entities $C = (K, P, U)$, where K is the keyword set describing the community, P the posts member of the community and U the set of users belonging to community C.

Definition 2: In a graph $G = (V, E)$, a community is defined as a set of vertices $C \subseteq V$ and a set of edges between the vertices. Depending on the desired graph, the vertices are represented as the corresponding entity, e.g. for keyword graph construction, the vertices V are represented as the keyword set, etc.

Definition 3: The set of edges of the subgraph induced by a community is $\mathrm{e} = \{(u, v)\ \epsilon\ E \mid u, v\ \epsilon\ C\}$

Definition 4: The size of the community $|C|$ is the number of vertices and $|E|$ is the number of edges in G.

Definition 5: An edge-weighted graph $G_w = (V, E, w)$, is a graph with V defining the set of vertices (e.g. keywords, posts or users), E the set of edges and w the set of edge weights respectively. The edge weight is defined as $w: E \rightarrow \Re$, where $w \in [0,1]$.

B. Keyword Extraction

The first stage of the overlapping community extraction process deals with deriving the most salient words (keywords) from posts submitted to Hubbub. There are several ways to implement this, such as Porter's stemmer to stem words in the posts, TF-IDF to count keyword frequency and inverse document frequency in the collection of posts, etc.

There is in general no limitation in using any of the above mentioned keyword extraction mechanisms (although the results may vary immensely due to the variation of extracted keywords), however in this paper we limit our

explanations to the proposed overlapping community formation approach rather than on the decision of the keyword extraction mechanism. Thus, we employ the Term frequency-Inverse Document Frequency (TF-IDF) method (due to its simplicity and popularity in information retrieval system and text mining applications) to scan the posts in order to obtain keywords by comparing the similarity of the posts in a corpus [16]. The term count $tf_{i,j}$ in a given post indicates the number of a particular term appearing in a post.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_{\mathrm{k}} \mathrm{n}_{\mathrm{k,j}}} \quad (1)$$

where $n_{i,j}$ is the number of occurrences of the considered term (t_i) in a post (p_j).

The inverse document frequency measures the general importance of the term t_i and is given by the following equation:

$$idf_i = log \frac{|P|}{|\{p: t_i \in p\}|} \quad (2)$$

where $|P|$ is the total number of posts in the corpus and $|\{p: t_i \in p\}|$ is the number of posts where the term t_i appears. The weight indicating the importance of a term to a corpus of posts is obtained using:

$$(tf - idf)_{i,j} = tf_{i,j} \times idf_i \quad (3)$$

The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Once the keywords are obtained the keyword processing module is used to construct the keyword association graphs.

C. Keyword Graph Construction

The main purpose of this stage is to identify the inherent semantic correlations of keywords to construct a graph-based keyword association network.

For this, we construct an edge-weighted graph $G(K)_w$ for keywords and utilize the co-occurrence of keywords approach assuming that high co-occurrence words are likely to be used together to describe a certain concept. Hence, it is reasonable to group them together to form a large semantic node.

The keywords graph is defined as $G(K)_w = (K, E, w)$ and forms an undirected weighted graph, where K is the set of keywords and E the set of edges between keywords, with $|K|$ and $|E|$ is the total number of keywords and edges respectively. The edge weights w_{ij} between keywords k_i *and* k_j are calculated using:

$$w_{ij} = \frac{f(k_i \wedge k_j)}{\max\left(f(k_i), f(k_j)\right)} \quad (4)$$

where $f(k_i \wedge k_j)$ denotes the number of posts that contain both words k_i and k_j. The edge weights obviously indicate the correlation degree of two keywordsk_i *and* k_j .

Fig. 1 illustrates an extracted keywords graph visualized using the Cyclone tool (see section 4) where interconnected nodes represent a correlation among keywords.

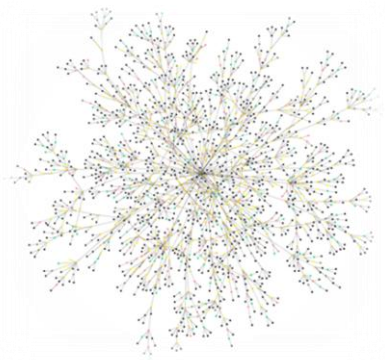

Figure 1: Visualization of Keyword Correlations using Cyclone

D. Keywords Partitioning and Classification

The next stage in the process of overlapping communities' generation is the keywords partitioning and classification. Here the keyword association graph $G(K)_w$ is used to partition and classify keywords so that a series of meaningful clusters which are overlapped to some extent are derived.

There are a number of approaches in graph partition to locate separated groups of nodes, however, only a few algorithms have capacity to find out *overlapping modules*, among which Clique Percolation Method (CPM) [17] has been applied for uncovering community structures of co-authorship networks, protein networks and keyword association graphs.

The basic idea of CPM is that natural and social systems are not fully connected networks because they are inherently noisy. So, a few missing links should be allowed. In CPM, modules are k-clique percolation clusters which are maximal k-clique-connected sub-graphs, i.e. the unions of all k-cliques that share k-1 vertices with a particular k-clique. This naturally allows overlaps between modules because one node can participate in several k-clique percolation clusters. The two k-cliques that share k-1 nodes are called adjacent. All adjacent maximal k-cliques connected together are regarded as a k-clique percolation cluster. This is what CPM uses to find clusters among the posts of the Hubbub forum.

E. Community formation

Based on the previous stage of keyword partitioning, relevant posts and their users are attracted to the clusters if they have enough similarity to the groups of keywords. The similarity of a post to a keyword cluster can be calculated in different ways, in which the cosine is the most often used one and utilized in our approach:

$$\cos(c_x, p_y) = \frac{\sum_i w_{xi} \cdot w_{yi}}{\sqrt{\sum_i w_{xi}^2} \cdot \sqrt{\sum_i w_{yi}^2}}. \quad (5)$$

A post p_y joins a keyword cluster C_X if the cosine similarity $cos(c_x, p_y)$ is larger than a threshold ε.

Accordingly a community is formed by integrating all relevant keywords, posts and users of the posts. While users write posts that are composed of keywords, keyword correlations produce core clusters that accordingly create relevant post and user clusters.

F. Extended Communities

During the experiments (see section 5), we found that CPM is still a very strict method when forming overlapping clusters because a) it requires a node to at least connect to other k-1 nodes in the clusters and b) all k-cliques should be adjacent to each other. This allows the posts which are very similar or relevant to reliably join corresponding clusters but at the same time isolate the ones with less similarity or importance.

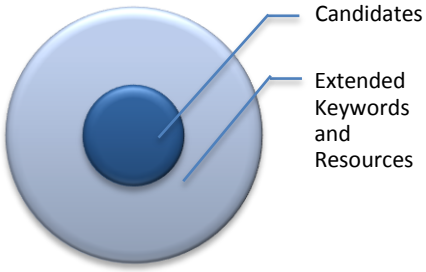

Figure 2: Extended Communities

To solve this problem, we extend a community to include words that have direct connections to the keywords of the community. These include words that may not form an adjacent k-clique in the community but they are all co-occurred with the core keywords of the community with enough correlation (e.g. above $\delta = 2$) (as illustrated in Fig. 2). To distinguish, the keywords that are core to the formation of communities are called 'candidates' and all the other words together with candidates are called 'nodes'. The extended communities therefore include posts that have enough similarity to the nodes of a community, not just their candidates.

The next section describes Cyclone, the visual frontend for the proposed automatic overlapping community generation mechanisms which is used to perform the community extension procedure as described above.

IV. EVALUATION

The evaluation of proposed approach has been conducted on the Hubbub forum data. We collected 2500 posts of 683 users from the forum over a period of 3 months to form the data set for our experiments. The keyword extraction generated 6336 keywords in total. CPM was then used to work on the keyword graph. While *k* (number of clique size) is small, CPM identified more clusters with small sizes (e.g., 290 clusters when k=3), whereas when k is large, only few connected adjacent k-cliques existed (e.g., 2 clusters when k=14) with many nodes left outside of the clusters. Fig. 3 shows the various communities structures with different k values. Here we chose k=11 because a reasonable number of clusters, 12, was generated. These clusters were composed of adjacent cliques, which were closely connected with each other, so they were the cores of the communities to form next. Fig. 4 shows the number of posts, users and keywords in the resulting 12 clusters respectively. Then we extended 12 clusters by including those un-clustered keywords and related users and posts that have close similarity to the existing clusters. Fig. 5 shows the results of the extended communities, where the keywords are those identified candidates in the keyword graph but nodes are the set of the above keywords pulsing other similar keywords. The numbers of users and posts increased obviously in Fig. 5 compared with Fig. 4. Actually the extended communities include all users and posts so no one is left out of the communities. Table 1 lists the possible topics of the 12 communities supported by the core keywords (keywords identified to compose the core communities). The summarized topics/keywords distinguish with each other and represent users' interests/queries from diverse angles. The summarization is also thought as meaningful and useful as told by the Hubbub's system administrators.

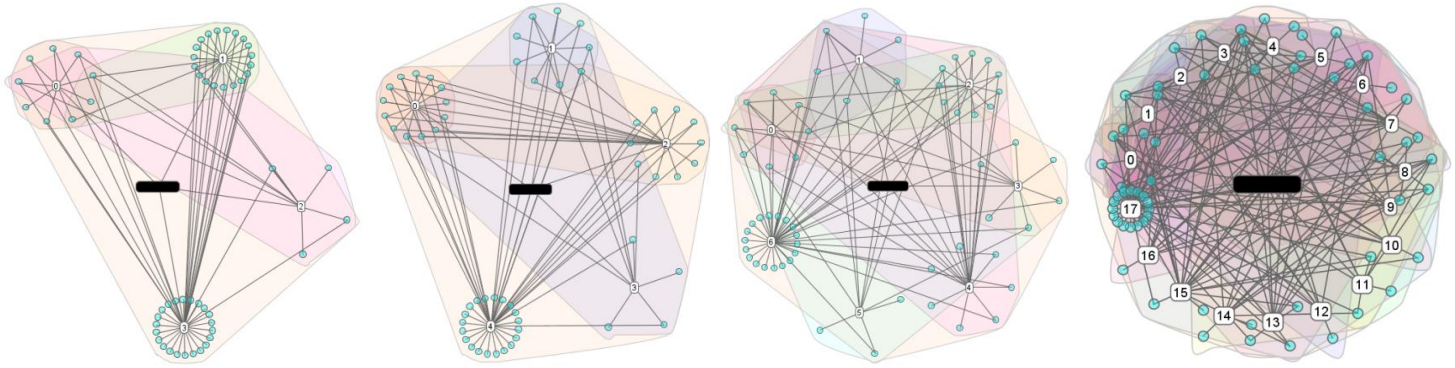

Figure 3: Different overlapping communities structures with varied *k*

V. DISCUSSIONS AND FUTURE WORK

This paper introduces the formation of overlapping communities in online forums (e.g., Hubbub) so as to capture users' diverse angles of interests. Within communities, users can easily share and seek information with other similar-minded people. The communities are also helpful to system administrators for easier and more efficient management. So the formed overlapping communities provide a novel means to support various services on the forum such as recommendation, alerting and profiling customer support personnel.

Recommendation. When a new post is submitted to the forum, the post will first be examined to see which communities it may belong to. These communities also indicate the possible interest topics the new post may be related with. Accordingly, the user of the new post will be recommended with the useful information in relevant communities, which may interest the user. The useful information of a community could be the most popular posts in the community, the posts with the highest weights (i.e., posts that have highest relevance to the community topic), the most similar posts in the community to the new post, and so on. Table 2 shows an example of the recommendations made by the overlapping communities and the current Hubbub forum. To a post regarding phone charging, communities-based recommendations provide the most relevant communities (0, 5 and 6) and the most useful posts in the communities (the contents of these posts are not listed here due to the limitation of space). These communities/posts are of the topics of phone charging and related issues such as handset registration. On the other hand, the Hubbub forum provides a few posts such as dial tone, phone hub, and caller ID which are not so pertinent to the new post A. This is probably because the current Hubbub weights high those popular words including hub, caller, dial, so the posts containing popular words overwhelm those with less popular words but appropriate to the new coming post. This example and some other tests on a few posts have verified that the information recommended by overlapping communities is more useful than those of the current Hubbub.

TABLE I. EXTRACTED TOPICS OF 12 COMMUNITIES

Community	Topics (Core keywords)
0	Battery, charging (posting, charging, battery, breach, consumer, trading, laws, organisation, dragging, feet, bearing,)
1	Video calls (picture, upsidedown, 100, camera\'s, greyed-out, cosequently, horizontal/vertical, video, webcam, camera)
2	Video calls + call charges (payg, subscribe, softphones, merge, locations, videophone, geographic, dupped, indicated, skypein, video)
3	Video chat + BT softphones (btsoftphone, bought, behalf, videophone, dupped, oz, softphone..after, hype, landline..anyone, this..seems, video,)
4	Cost (5, evenings, weekends, costs, landlines, ntl, determined, supplies, telewest, 5p, refer,)
5	Caller display, registration, configuration (register, handset, pin, association, flash, 0, ready, display, registering, registration, mode,)
6	Sound (card, mac, os, apple, runs, knowledge, parallels, bridge, handset, media, headset, audio,)
7	Firmwell settings (answered, ie7, ie6, closing, add-on, sneaked, ie/tools/manage, add-ons, sombody, disable)
8	Text, sms, router (provide, turns, controls, beta, browsers, image, apparently, safari, submit, ignore, text)
9	Failure, problems (secure, fails, firewalls, wi-fi, cd, channel, physical, networking, bother, word, green, save, mode,)
10	Parental control (posting, parental, controls, relating, beta, locked, wds, menus, dyndns, love, cookies, contacts, quiet, cleared)
11	Secure, Norton, linux (secure, norton, hard, os, knowledge, installation, linux, platform, latest, compatibility, route, browsing, guid, practice,)

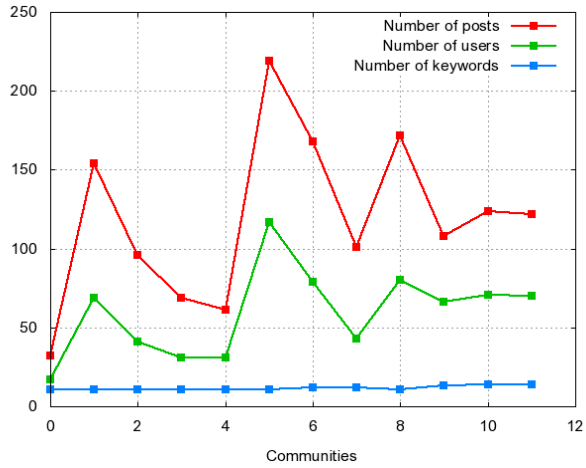

Figure 4: Experiment 1 (Communities)

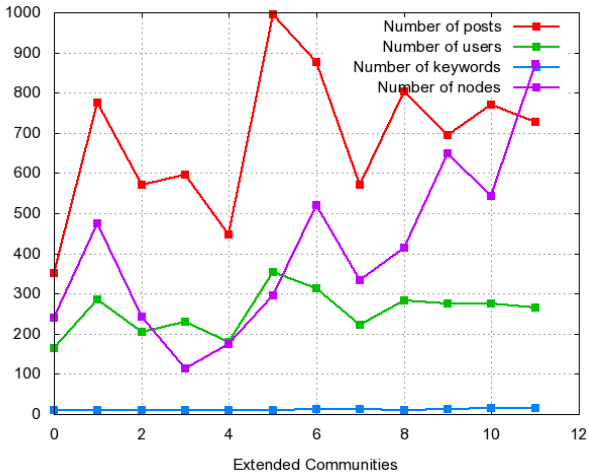

Figure 5: Experiment 2 (Extended Communities)

TABLE II. COMPARISON BETWEEN OVERLAPPING COMMUNITIES-BASED RECOMMENDATION AND THE ORIGINAL HUBBUB RECOMMENDATION

Post A	**Should my phone still be charging after 36 hours?** Many people have reported that their phones always display a slight charge even when fully charged, mine also and it does not seem a problem. When first charged (3 or 4 hours) the display runs fully across the battery then only slightly at the L/H side, this seems to be normal.
Recommended Communities	0 (posting, charging, battery, breach, consumer, trading, laws, organisation, dragging, feet, bearing,) 5 (register, handset, pin, association, flash, 0, ready, display, registering, registration, mode,) 6 (card, mac, os, apple, runs, knowledge, parallels, bridge, handset, media, headset, audio,)
Hubbub Recommendations	*a) 1. no dial tone* ***no dial tone*** *2. I can not get my BT Hub Phone 1010 to register to the BT Hub* ***I can not get my BT Hub Phone 1010 to register to the BT Hub*** *3. Caller ID on Hubphone doesn't work* ***I never see caller Id on the hubphone. The land line phone does display it.*** ……

Alerting. When a new post is submitted into the forum, this post will be sent to relevant users to alert them of new useful information. Again, the new post will be clustered into relevant communities and the users of all relevant communities will receive the alert. For example, when the new post A as shown in Table 2 is submitted to the forum, users of communities 6, 5 and 0 will be informed of the new coming post regards battery charging, if these users have set up their "alert" option to be true.

Customer Agent Profiling. In the hubbub forum, there are a few human helpers to help investigating the posts in the forum and answering un-answered queries when necessary. These helpers are not fixed because new helpers may join in at any time. For the moment, it is the forum administrator to define categories of help and allocate new helpers to suitable categories by hand. As there are now more than 40,000 posts in the forum with varied questions and topics, it was quite a challenge to manually organise the helpers in a reasonable way. By analysing the posts and their co-occurred keywords, the overlapping communities generated in this paper provide a good suggestion to recommend categories for helpers to choose. This work will be implemented in the near future.

ACKNOWLEDGMENT

We would like to thank the members of the Centre for Information & Security Systems Research, in particular the Future Technologies and Intelligent Systems Groups at BT UK for their invaluable contributions and support. We are also grateful to the anonymous reviewers for their valuable suggestions and comments.

REFERENCES

[1] A. Das, M. Datar, A. Garg, "Google News Personalization: Scalable Online, Collaborative Filtering", WWW2007/Track: Industrial Practice and Experience, Banff, Alberta, Canada, 2007.

[2] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon and J. Riedly, "GroupLens: Applying collaborative Filtering to UseNet news", Communications of the ACM, 40(3), pp 77-87, 1997.

[3] M. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann and M Gerstein, "Genomic analysis of regularity network dynamics reveals large topological changes", Nature 431, pp 308-312, 2004.

[4] S. Wuchty and E. Almaas, "Peeling the yeast protein network," Proteomics 5, pp 444-449, 2005.

[5] S. Seufert, U. Lechner, K. Stanoevska, "A reference model for online learning communities," International Journal on e-learning, pp 43-55, 2002.

[6] L. Talavera and E. Gaudioso, "Mining student data to characterize similar behavior groups in unstructured collaboration spaces," Workshop on Artificial Intelligence in CSCL, European Conference on Artificial Intelligence, pp 17-23, 2004.

[7] M. Khambatti, K. D. Ryu and P. Dasgupta, "Structuring Peer-to-Peer Networks using Interest-based Communities, Databases, Information Systems and Peer-to-Peer Computing," International Workshop on DBISP2P , pp 48-63, 2004.

[8] A. Iamnitchiand, M. Ripeanu and I. Foster, "Small-World File-Sharing Communities", International Conference of the IEEE Communications Society (InfoCom), 2004.

[9] E. Ogston, B. Overeinder, M. van Steen, B. Brazier, "Group Formation among P2P Agents: Learning Group Characteristics", International Workshop on Agents and Peer-to-Peer Computing (AP2PC), pp 59-70, 2003.

[10] F. Wang, "Multi-interest communities and community-based recommendations", 3rd International Conference on Web Information Systems and Technologies, Barcelona, Spain, 2007.

[11] H. Duman, A. Healing, R. Ghanea-Hercock, "An Intelligent Agent Approach for Visual Information Structure Generation", 2009 IEEE Symposium on Intelligent Agents, 2009.

[12] A. K. Jain, M. N. Murty, P. J. Flynn, "Data clustering: a review", ACM Computing Survey 31, 264-323, 1999.

[13] B. King, "Step-wise clustering procedures", Journal of the American Statistical Association 69, pp 86-101, 1967.

[14] D. Nguyen, S. Thompson, C. Hoile, "Hubbub – An innovative customer support forum", International Conference on Business Information Systems, 2008.

[15] H. Duman, A. Healing, R. Ghanea-Hercock, "An Adaptive Visual Clustering for Mixed-Initiative Information Structuring", 13th International Conference on Human-Computer Interaction, 2009.

[16] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval", Information Processing & Management, 24, pp 513-523, 1988.

[17] G. Palla, "Uncovering the overlapping community structure of complex networks in nature and society", Nature, 435, pp 814-818, 2005.