Towards Identifying Collaborative Learning Groups Using Social Media

http://dx.doi.org/10.3991/ijet.v7iS2.2325

S. Softic
Graz University of Technology, Graz, Austria

***Abstract*—This work reports about the preliminary results and ongoing research based upon profiling collaborative learning groups of persons within the social micro-blogging platforms like Twitter[1] that share potentially common interests on special topic. Hereby the focus is held on spontaneously initiated collaborative learning in Social Media and detection of collaborative learning groups based upon their communication dynamics. Research questions targeted to be answered are: are there any useful data mining algorithms to fulfill the task of pre-selection and clustering of users in social networks, how good do they perform, and what are the metrics that could be used for detection and evaluation in the realm of this task. Basic approach presented here uses as preamble hypothesis that users and their interests in Social Networks can be identified through content generated by them and content they consume. Special focus is held on topic oriented approach as least common bounding point. Those should be also the basic criteria used to detect and outline the learning groups. The aim of this work is to deliver first scientific pre-work for successfully implementation of recommender systems using social network metrics and content features of social network users for the purposes of better learning group communication and information consumption.**

***Index Terms*—Educational Data Mining, Micro-Blogging, Social Network Analysis; Recommender Systems**

I. INTRODUCTION AND RELATED WORK

Recent decade with the phenomenon of Web 2.0 has brought the concept of user generated content, social networks and as a part of it a phenomenon of micro blogging. Either through deliberate or incidental contribution increasing number of people has created a huge set of data that gives us millions of potential insights into user experience, marketing, personal tastes, and human behavior in general.

Especially micro blogging platforms as Twitter gained strong importance in recent years. Today Twitter is generating 200 million Tweets and 1.6 million search queries each day. According to recent statistics[2] (2012), Twitter has over 250 million users. As such platform it implies daily numerous social interactions based upon interest sharing, opinion and experience exchange. Recent research has shown that social interactions with people who share the same affinities can contribute progress in research and learning [1].

Another trend is that many of them blog and tweet about events, like conferences, especially in communication and technical research communities [2] [3] [4]. Lately also universities started to use the advantage of fast information exchange in micro blogs to consolidate the information sharing and discussion across courses lead by the idea of technology based and collaborative learning.

This creates huge opportunities for profiling [6]. The attendees tweet about what they notice, what they remark as interesting according special topic of matter. In the focus of lecture support this could be a special lecture or topic related to it.

However many of the content generated by the people a user follow does not offer focused view on a special interest and it is still noisy and unstructured.

Micro blogger assign topics, links and media artifacts to their user generated content. Focused view on heterogeneously disseminated information resources like this accommodated to personal preferences and learning goals offers the possibility of spontaneous involvement and initiation of collaborative learning tasks based upon the matter of content. Interacting on same topic targeted on learning process generates opinion exchange and knowledge aggregation.

What if these users could be clustered into sub-networks of main topic based upon their interest using this information? What if science could contribute to these users to receive filtered view on information generated in their micro sub-networks? Which methods or technologies would be suitable for this challenge? What are the metrics that can be used to achieve this distinction?

These are the questions this paper is trying to address in a specific area of collaborative learning. Efforts described here will not be able to offer answers to all the questions, but it tends to report a preliminary study on possibilities offered through science how to detect and cluster people with similar interests inside the social networks and let them communicate on purpose with each other in the boundaries of their interest. Such awareness delivers many appliances like in the area of recommender systems for e.g. collaborative learning and technology enhanced learning or for interconnecting the interest groups like learn and research communities [5]. Further then that this work is interesting for areas like viral marketing and market research for placing offers and materials a certain group of users would consume [7].

Processes that happen spontaneously are mostly initiated by adequate stimuli. As necessary precondition for stimuli of this kind as fundamentally important indicator a familiar ambience will be assumed. All methodologies represented in following subsections will use this hypothesis as preamble

[1] http://www.twitter.com
[2] http://thesocialskinny.com/100-social-media-statistics-for-2012/

II. METHODOLOGY

Thinking in manner of solving such complex task as collaborative learning content consummation inside of heterogeneous information networks as Social Networks are, the first task that has to be solved is to identify the information stakeholder relevant for the process of collaborative learning with respect to information consumer. In order to achieve this first task area of semantically-lexical analysis combined with NLP and data mining can deliver the proper tools and techniques.

However before the clustering process can be done, data has to be pre-processed and formed in a manner acceptable for common clustering algorithms. Then significant features of content should be used to determinate least common relation. In the case of Twitter this would be mentions denoted in micro text fragments with "@someusername" and hash tags denoted using "#sometopic" .Hash tags are expected to contribute content related clustering while mentions will be used to discover relatedness in social context.

This methodology follows the logic of item based filtering of recommender systems design. To the best of authors knowledge no similar comparison or evaluations has been done so far in the area on similarity measures as preamble of item based recommendation of learning groups.

Using these two common features as base for clustering and identification of potential collaborative groups makes sense since the persons who communicate about same topic and persons belong potentially to the same interest area. On the other hand persons mentioning the same communication actors also share implicitly an interest on the content generated from particular source.

Tweets as small as they are, brought into a proper context can delivery astonishing results. Their usage as "social sensors" is applicable for several purposes. Lately some work on tracking the sentiment inside the "electronic word of mouth" as tweets were described has been published with respect to e-commerce area of appliance [7].

This is a pre-assumption that has to be necessarily done before the context of learning groups in the manner of E-Learning can be considered. Therefore for now the focus of this paper remains on this pre-condition. Aim in this realm was targeted primary at evaluation of similarity measures needed for clustering of collaborative groups.

To the best of authors knowledge no similar comparison or evaluations has been done so far in the area on similarity measures as preamble of item based recommendation of learning groups.

III. ACQUISITION OF DATA

As data source serves the database of Grabeeter[3] tool which includes the tweets from around 1600 users from mostly educational and research area. This tool developed by Social Learning Group at Graz University of Technology simply grabs the user timeline via the regular Twitter API[4]. Therefore potentially every person or institution that owns a Twitter account can grab his/her/its own Tweets using the Grabeeter. These tweets are then preserved in

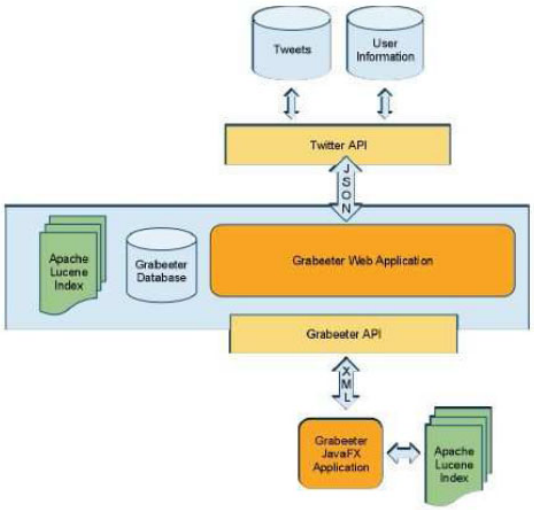

Figure 1. Architecture of Grabeeter

the local database of the software and can be searched by web interface or by a JavaFX based client. Alternatively Grabeeter offers a rudimentary REST API[5] with export possibility of timeline to XML or JSON format. For local search with Java client tweets exported to file system has to be indexed by local Apache Lucene[6] engine.

Grabeeter serves primarily as tweet storage. In contrast to Twitter API which allows the insights on only last 300 tweets, Grabeeter provides all stored tweets and makes no restriction over time. At the moment of writing this paper Grabeeter database contained approximately 4.700.000 tweets, which makes it a very reliable source.

IV. DEFINITIONS, DATA SET AND MEASUREMENT

A. Definitions

Considered as simple concept a Collaborative learning group can be primary treated as a "Interest Group". Let us define a potential "Interest Group" in a more formal way:

Let **G** be the a set of "Group Candidates" defined as follows:

$$G = \{ G_i \} \text{ where } i = 1..\ n \text{ and } n \subseteq N \qquad (1)$$

And a single member of this set $\boldsymbol{G_i = \{ C_j , L_k \}}$ is a pair of items where $\boldsymbol{C_j}$ is a vector of top content items and $\boldsymbol{S_k}$ a vector of top social references (where $\boldsymbol{j , k = 1 .. n}$ and $\boldsymbol{n \subseteq N}$, and where $\boldsymbol{j \neq k}$). Items of both sets can be either single values or tuple of values.

In current observation single values and value pairs depending on similarity function will be used (e.g. #hash-tag or {#hash -tag , 2} where 2 represent the occurrence). Also $\boldsymbol{j}$ and $\boldsymbol{k}$ indexes are of the same length, which means that we assume $\boldsymbol{j = k}$.

Let **H** be a single reference "Reference Candidate" of type "Group Candidate " as previously defined

[3] http://grabeeter.tugraz.at
[4] https://dev.twitter.com/
[5] http://grabeeter.tugraz.at/developers
[6] http://lucene.apache.org

$$H = (Cr, Lr) \quad r = 1 .. n \text{ and } n \subseteq N \qquad (2)$$

Note that indexes ***j*** , ***k*** and ***r*** are the same length! Further **T** a pair of real value thresholds between 0 and 1 will be defined as follows:

$$T = \{ t_C, t_L \subseteq R \mid 0 \le t_C \le 1 \text{ and } 0 \le t_L \le 1 \} \qquad (3)$$

Intersection between the corresponding item sets ***Cj*** , ***Cr*** and ***Lr*** , ***Lk*** delivers a subset ***μ***:

$$\mu = H \cap G_i = (C_\mu, L_\mu) \qquad (4)$$

This subset delivers input for a similarity ration function **α**. This function delivers either correspondence ratio in percent between significant content or social reference items from intersection set ***μ*** respectively the "Group Candidate" vectors as a value between 0 und 1.

$$a(\mu) = \{ x \subseteq R \mid 0 \le x \le 1 \} \qquad (5)$$

As final step a threshold based clustering function **δ** is applied on ***a*** to determinate whether a "Group Candidate" **Gi** belongs to an "Interest Group" or not.

$$\delta(\alpha, T) = \begin{cases} 1 \text{ if } 0 \le t_C \mid t_L \le a \le 1 \\ 0 \text{ otherwise} \end{cases} \qquad (6)$$

Hence "Interest Group" **I** is defined through following factors:

$$I = (G, H) \text{ where } \delta(a(\mu), T) = 1 \qquad (7)$$

For the matter of evaluation one additional measure will be defined called **λ** or "acceptance ratio". This is a ratio between the count of accepted and considered "Group Candidates".

$$\lambda = \#accepted\ G_i \,/\, \#considered\ G_i,\ 0 \le \lambda \le 1 \qquad (8)$$

As similarity function in the context of group detection *Cosine Similarity* was used for single valued vectors while *Euclidian Distance* was used as pair value vectors similarity measure.

1) Cosine Similarity

This ratio can be used as a similarity measure between any two vectors representing documents, text fragments, snippets or the like. Cosine Similarity represents the angle between two vectors that reflects their diversity. As the angle between the vectors becomes shorter, the cosine angle approaches the value of 1, which means that the two vectors are getting closer regarding their similarity. Total diversity is represented through 0. Cosine Similarity is defined as:

$$\text{Sim}(A, B) = \cos\theta = \frac{A \bullet B}{|A| \cdot |B|} \qquad (9)$$

2) Euclidian Distance

Euclidian Distance is base for many similarity measures. The distance between the vectors A and B is defined as follows:

$$d(A, B) = \sqrt{\sum_i^n (Ai - Bi)^2} \qquad (10)$$

Euclidean distance is most often used to compare profiles of respondents across variables. In other words, Euclidean distance is the square root of the sum of squared differences between corresponding elements of the two vectors. Note that the formula treats the values of X and Y seriously: no adjustment is made for differences in scale. In order to hold the scaling convention some correlations and scaling for the purposes of evaluation respectively expressing the similarity in percent as value between 0 and 1 has been made.

B. Data set preparation and measurement process

As reference data for evaluation set top 100 results for persons from Grabeeter accounts register who used *"elearning"* or *"e-learning"* keyword in their tweets were taken. This is done in order to compare the ratio of similarity respectively the size of candidate group.

For evaluation purposes always the last 250 tweets of a specific user hast been taken into account. Out of them top 5, 10 and 20 hash tags and mentions per each user were generated and compared using similarity measures: Cosine Similarity and Euclidian distance. Vectors are all of same length. Dynamical vector size adjusting was intentionally left out since the main point of matter rather whether the approach delivers promising results than the scalability of algorithm presented here.

All measurement made respectively the detection of potential "Interest Groups" were made using a specially designed Similarity API based upon Grabeeter tool. Similarity API was implemented in PHP[7] using the Grabeeter database as primary data source. Results are delivered in JSON[8] (Fig.2) Format and finally processed into results using the statistic functions inside the API.

Cosine Similarity and Euclidian Distance were used as similarity measure since they has been shown in various research works before [8] as reliable indicators for detection of text based similarity. Distances used here belong in two different groups. Cosine Similarity uses only simple items to calculate the similarity angle among two text terms while Euclidian distance is calculated using the text item and their occurrence.

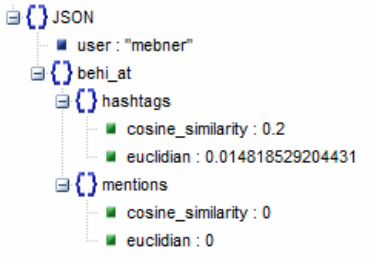

Figure 2. Similarity API in action

[7] http://www.php.net/
[8] http://www.json.org/

Upon these results clustering using simple thresholds in percent in the range from 10% and 20% has been applied on similarity results. As a reference candidate for target learning group @mebner account was used since this account can be considered as one of the key competence bearer for E-Learning area.

Each simulation consisted as described above out of similarity calculation and calculation of δ ratio function which checks if the result which is calculated for similarity reaches the threshold. "Interest Group" potential was reflected by the number of acceptable group candidates respectively the number of observed group candidates or as defined in Def. (8) as λ ("acceptance ratio") .

Values presented in the results section represent a median value of retrieval ratio. To get a deeper insight also the number of top hash tags and mentions was varied from 5 to 10 to 20 in order to evaluate how the length of parameter vector s influence the result.

Expectance of presented measurement relies on the thought that comparison different similarity measures should deliver first hints on building the collaborative groups techniques and an evaluation which of the measure fits in the best way for proposed effort. It has to be considered that the test group was quite small but as it will be shown in the result section it delivers very encouraging results. Hereby has to be mentioned that the choice of keywords for filtering the users for candidate group as well as choice of reference candidate had a decisive influence on similarity level ratio as most important clustering criteria.

V. PRELIMINARY RESULTS AND DISCUSSION

At the beginning of this section it has to be mentioned that all of the observation made respectively simple clustering of the potential "Interest Group(s)" are aiming at the evaluation of proposed methodology and system dynamics more than at qualitative analysis of retrieved results. "Interest group" detection is meant to be as prestep for building the qualitative "Collaborative groups". Described methodologies in this paper are meant to act as "sieves" and can be used as tools to simplify the task of building "Collaborative groups" by reducing the number of potential candidates.

A. Single valued measurement results with Cosine Similarity

1) Evaluation of "hash tag" vectors

Evaluation results for Cosine Similarity measure applied on "hash tags" vectors of different length (5,10,20) with thresholds of 0, 1 (10%) and 0, 2 (20%) can be seen in Fig. 3 and Fig. 4:

The 10% threshold can be easily reached result retrieved in Fig. 3 and their diffusion between 0 and 0, 35 (or 0% and 35%) does not come surprisingly. The same can be observed for 20% threshold (Fig. 4) according the dynamics, although test with 10% boundary drifts more stable hand in hand with candidate group size, both of them tend to converge against a median value. Linear behavior of both systems relies on distribution of correspondences across the test set and on the nature of similarity function. Threshold with 10% is reached easily and causes less oscillation. Real nature can be recognized for candidate sets n > 80. Systems ten to stability as the candidate group increases and acts linearly number of hash tags. In Fig. 3 there are some deviations for vectors of size 20. The reason is the structure of data set and its potential regarding the variation of vector size.

Same can be said for Fig. 4. And 5 "hash tags" sized vectors. It is obvious that significantly corresponding hash tags in test data set are placed at top 5 positions.

2) Evaluation of "mentions" vectors

Fig. 5 and Fig 6. reflect the results of appliance of Cosine Similarity on "mentions". Sam as in the case with "hash tags" the size of vectors was varied starting by 5 over 10 up to 20.

For 10% matching threshold however the values of λ ("acceptance ratio") seem to perform better than for "hash tags" (0, 05 < λ <0, 45). This fact points to the consistence better distribution and quality of "mentions" retrieved from test data. Same case can be observed also for the 20% threshold (0 ≤ λ ≤ 0, 3). This is also reflected in the trend of λ which changes consistent together with the growth of number of group candidates.

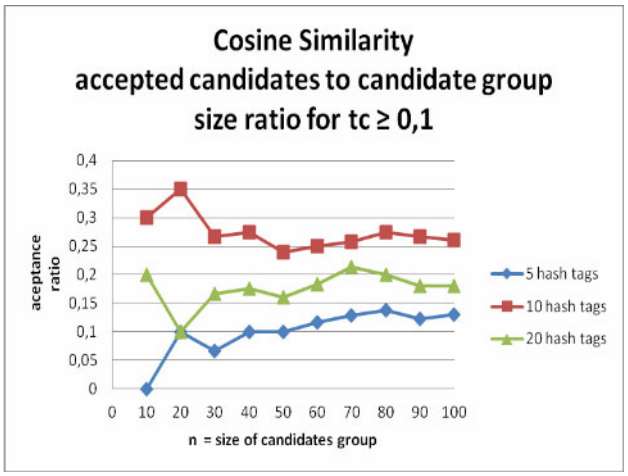

Figure 3. Cosine Similarity – ratio that reflects the percentage of accepted number of candidates to the total number of evaluated candidates for $tC \geq 0,1$

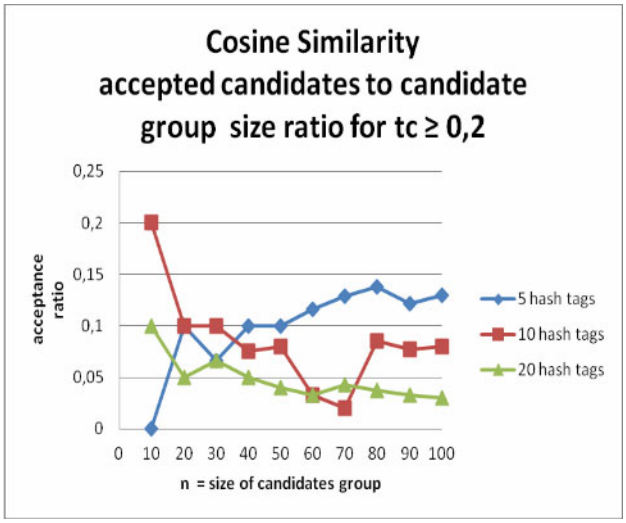

Figure 4. Cosine Similarity – ratio that reflects the percentage of accepted number of candidates to the total number of evaluated candidates for $tC \geq 0,2$

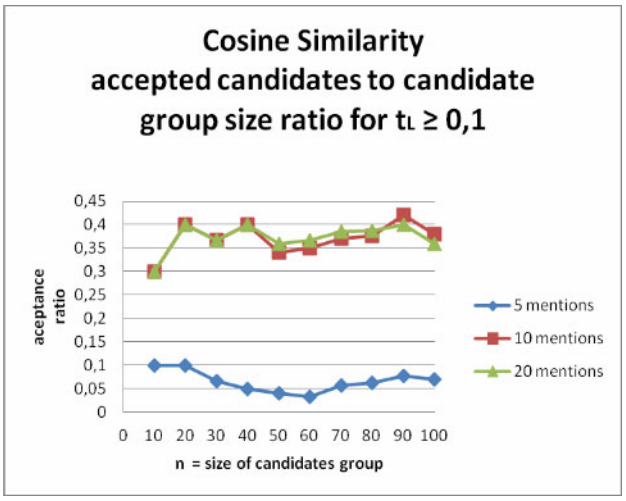

Figure 5. Cosine Similarity – ratio that reflects the percentage of accepted number of candidates to the total number of evaluated candidates for *tL ≥ 0,1*

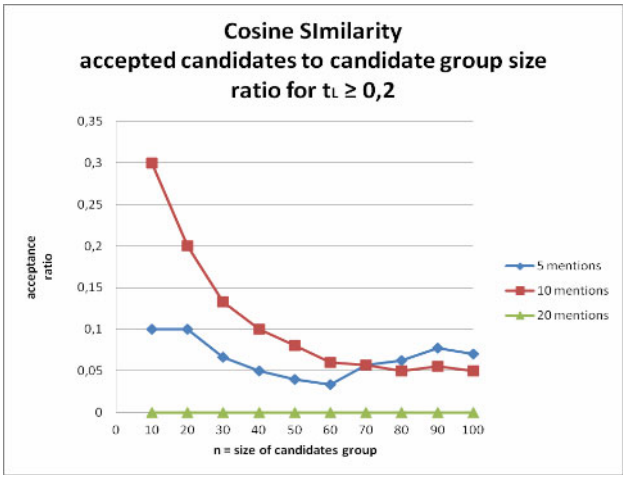

Figure 6. Cosine Similarity – ratio that reflects the percentage of accepted number of candidates to the total number of evaluated candidates for *tL ≥ 0,2*

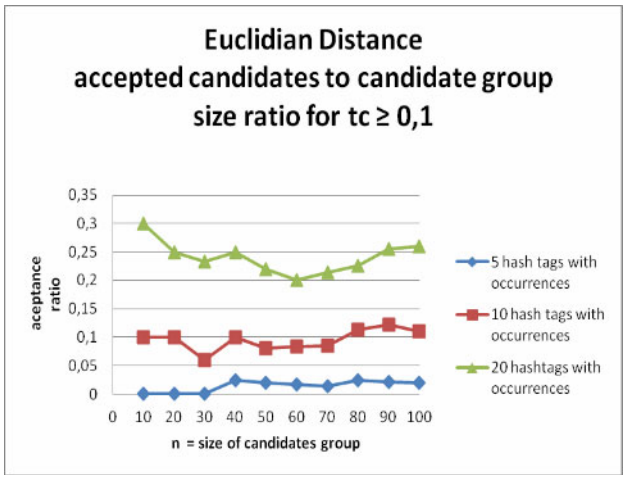

Figure 7. Euclidian Distance – ratio that reflects the percentage of accepted number of candidates to the total number of evaluated candidates for *tC ≥ 0,1*

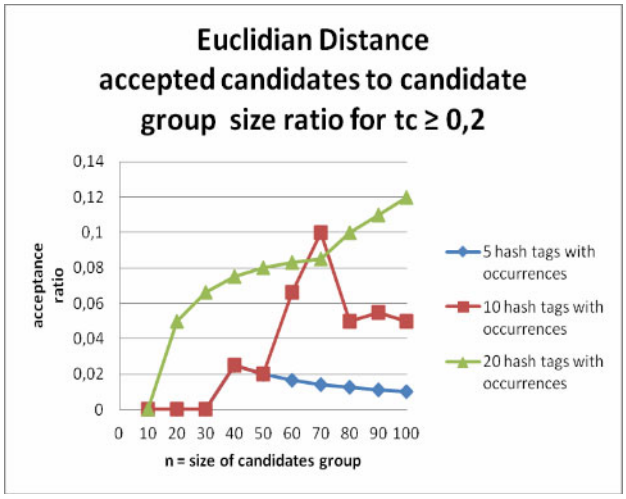

Figure 8. Euclidian Distance – ratio that reflects the percentage of accepted number of candidates to the total number of evaluated candidates for *tC ≥ 0,2*

The same observation as for the „hash tags“ can be concluded for the appliance of Cosine Similarity on the mentions in the case of linear dependency of „acceptance ratio“ from the candidate group size.

Dynamics of the system as already mentioned relies of distribution of interesting „mentions“ and on the nature of similarity function. Deviation regariding the vector size are caused as in the case of „hash tags“ by the placement of relevant „mentions“ inside the vector. Interpreting the course and form of „acceptance ration“ it can be easily consluded that in observed data se the mentions are distributed more equally all-over the data set.

B. Pair valued measurment results with Euclid Distance

1) Evaluation of "hash tag" vectors with occurrences

In following figures results based upon Euclidian Distance will be presented. Additionally to sole "hash tags" also their occurrences are taken into account by calculation of Euclidian distance. Occurrence as it will be shown contributed to more stable behavior of "acceptance ratio" course.

Fig.7. and Fig.8. are representing the results for thresholds of 10% and 20%. It is significant that larger number if "hash tags" in vector for the the case of 10% threshold also increases the "acceptance ratio" ($0 \leq \lambda \leq 0,3$). For the 20% thresholds this happens after the size of candidate group exceeds the count of 70 with approximately half lesser "acceptance ratio" ($0 \leq \lambda \leq 0,12$).

Except two deviating values for n = 50 and n = 60 observations made by 10% thresholds mainly correspond with the 20% case. It is also evident especially for the 10% that when a "acceptance ration" reaches its nearly median value it hardly deviates heavily. Depending obviously on threshold this convergent behavior is reached at different count of candidates.

2) Evaluation of "mentions" vectors with occurences

Hardly different behave threshold based clustering based upon Euclidan Distance for input vectors consisting out of "mentions" and their occurrences which is clearly depicted in Fig. 9 and Fig. 10. Once again size of input vector here filled with "mentions" and occurrence counts influences the rate of "acceptance ratio".

For threshold of 10% "acceptance ration" varies in dependence on size of vectors between 0 in single case of 10 candidates and 10 "mentions" up to high rate of 0,4. Same characteristics are also measured by the 20% threshold.

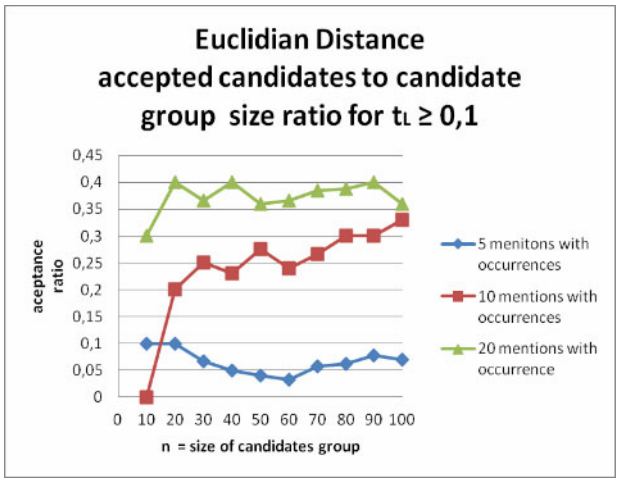

Figure 9. Euclidian Distance – ratio that reflects the percentage of accepted number of candidates to the total number of evaluated candidates for *tL ≥ 0,1*

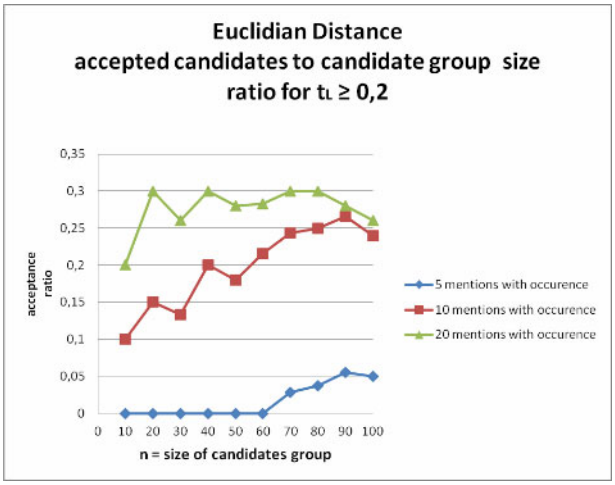

Figure 10. Euclidian Distance – ratio that reflects the percentage of accepted number of candidates to the total number of evaluated candidates for *tL ≥ 0,2*

However here is the highest "acceptance ratio" value by 0,3.

In comparison to the "hash tags" Euclidian Distance measurements with same clustering threshold "acceptance ratio" does not decreases by the same coefficient. The reason for this behavior relies most probably on more equally dissemination of relevant vector items ("mentions") in test data set than the one of "hash tags" as in the case of Cosine Similarity for the same observation. Same as in the case of "hash tags" here even more evident the course of "acceptance ratio" values deviates lesser as the number of candidates increases.

VI. CONCLUSION AND FUTURE WORK

Concluding the measurements some significant observations has been made worth outlining as the results. First of all despite of very small test set including only 100 candidates and one reference candidate conclusion dynamics of similarity measures based threshold driven clustering could be evaluated and observed with some valuable answers. Although no qualitative evaluation has been made, and "acceptance ratio" as such is clearly inaccurate indicator of the precise distinction of discovered "interest groups", it was sufficient to approve the significance of the intention behind the usage of similarity based approach for organizing and steering of targeted information exchange between the persons that have same interests participating in Social Networks as Twitter.

Results presented in previous section are showing us that this approach looks promising even on very small data sets, which is encouraging for future works. The choice of parameters approved the initial expectance of setting the first steps in right direction. Further it made possible the comparison of two approaches.

Details from measurement also clearly outlined the facts about the stability of single measures. Euclidian Distance performed more stable and consistent in comparison to Cosine Similarity at least according the presented measurement. Some instability characteristics of Cosine Similarity can be explained by not equally dissemination of relevant matching items across the data set, however this measurement demonstrates because of that even more realistic circumstances.

It would be too optimistic to claim that the presented approach could be the end concept towards building collaborative learning groups however it seems to be a small step in right direction.

It would be more interesting for future work to extend the measurement on more appliance cases and reference users from different areas. Additionally in order to enable more accurate and qualitative evaluation of clustering single matching similarities should be considered, clustered and re-evaluated more precisely during the measurement process. Also some other approved similarity measures like Pearson or Jaccard could be considered as extension to the current experiment setup. In this way it could be possible to determinate the level of quality of each single similarity method. Such extension of presented approach would contribute the reliability of the initial idea. Improvements towards preparation of more extended test data set are aimed to be done with expectation to reapprove the results.

Nevertheless presented results confirm the basic intention of the current work made by author and other researchers towards improving organized collaboration and information placement and exchange in Social Networks and underlines the claims that such effort is based upon realistic expectations.

Most encouraging about this approach is awareness that current scientific technologies, methods and techniques can be used to deliver complete solutions and answers to addressed challenges in a very near future.

ACKNOWLEDGMENT

I would like to thank Dr. Ebner and to his Social Learning Group crew from Graz University of Technology, who provided the Grabeeter as valuable data storage for my experiment and many useful suggestions regarding the implementation of Similarity API. My special thanks also to my colleague and my dear friend Behnam Taraghi who always delivered constructive ideas to my research work.

REFERENCES

[1] U. Mejías, "A nomad's guide to learning and social Software", in The Knowledge Tree, Edition online, 2005.

[2] W. Reinhardt, M. Ebner, G.Beham, and C. Costa, "How people are using Twitter during conferences," in V. Hornung-Praehauser,

M. Luckmann, Eds.: Creativity and Innovation Competencies on the Web. Proceedings of the 5th EduMedia 2009, pp. 145-156, Salzburg, 2009.

[3] M. Ebner, H. Mühlburger,S. Schaffert, M. Schiefner, W.Reinhardt, and S. Wheeler, "Get granular on Twitter - tweets from a conference and their limited usefulness for non-participants, " in Key competences in the knowledge society (KCKS 2010), pp. 102-113, 2010.

[4] R. Klamma, M. C. Pham and Y. Cao: "You never walk alone: recommending academic events based on Social Network Analysis," in Proceedings of the First International Conference on Complex Science (Complex'09), pp.23-25, Shanghai, China, Februar 2009.

[5] M. Ebner, T. Altmann and S. Softic "@twitter analysis of #edmedia10– is the #informationstream usable for the #mass," in Form@re Open Journal 11(74) (74), 2011.

[6] S. Softic, M. Ebner, H. Mühlburger and T. Altmann. "@Twitter mining #microblogs using #semantic technologies," in Proceedings of 6th Workshop on Semantic Web Applications and Perspectives SWAP, pp. 1-12, 2010.

[7] B. J. Jansen, M. Zhang, K. Sobel and A. Chowdury. "Twitter power: Tweets as electronic word of mouth", J. Am. Soc. Inf. Sci., 60: 2169-2188, 2009. http://dx.doi.org/10.1002/asi.21149

[8] A. Huang. "Similarity measures for text document clustering". in Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), pp. 49-56, Christchurch, New Zealand, 2008

AUTHOR

S. Softic works as senior researcher and he is a PhD student at Social Learning Group, Graz University of Technology, Austria. His specialties are: Social Networks Analysis, Educational Data Mining, Semantic Web, Linked Data, E-Learning (e-mail:selver.softic@tugraz.at).

Manuscript received 23 Octobver 2012. Published as resubmitted by the author 7 November 2012.