# Detecting Collaboration Regions in a Chat Session

Dan Banica, Stefan Trausan-Matu, Traian Rebedea, University "Politehnica" of Bucharest, Department of Computer Science and Engineering, 313 Splaiul Independentei, Bucharest, Romania
Email: dan.banica@cti.pub.ro, stefan.trausan@cs.pub.ro, traian.rebedea@cs.pub.ro

**Abstract:** The paper presents an approach and a software system for the automatic detection of collaboration regions in a chat session. Although there is no unanimous accepted definition of good collaboration regions, they are generally easy to recognize, as their most important properties are known: they contain replies from more participants, the replies are on-topic and the participants elaborate together, they construct starting from the ideas begun by others. This is in opposition to the case of participants discussing in parallel, ignoring each other. Perfectly detecting collaboration regions involves understanding the natural language, which is an AI-complete problem, not solvable for the moment. However, we will show in this paper that good approximations can be done by using some heuristics. We will present a few such techniques, as well as a framework for detecting collaboration regions starting from a Bakhtinian perspective.

## Introduction

Computer Supported Collaborative Learning (CSCL) aims to facilitate interactions between students, or between tutors and students using computers. Numerous tools allow users located far away to communicate, one of the most representatives being the chat systems (Stahl, 2006). In this paper we present an approach for analyzing a chat discussion and identifying regions with a good collaboration. Such regions occur when more participants are involved, discuss on-topic and elaborate together, as opposed to the case when they ignore each other, each exposing only his own ideas.

There are few systems that analyze chat conversations. For example, TagHelper (Rose & all, 2008) uses Natural Language Processing and Machine Learning techniques for tagging utterances, starting from annotated corpora. Dong's approach (Dong, 2006) is in some aspects similar to ours because it analyzes globally the chat for detecting concept formation, but he applies other analysis techniques and it don't detect collaboration regions.

The corpus used for analysis was developed at the University "Politehnica" of Bucharest which consists of chats held in the VMT (Stahl, 2009) environment which has an important advantage – it allows participants to specify to what reply they are answering, using *explicit links*. This is done by clicking another utterance before submitting a reply. The scenario assigned for the chats is the following: each participant must choose a collaborative technology (chat, blog, wiki and forum) and in the first part of the talk he must try to convince the others that his technology is the best. In the second part, the learners must try to reach a consensus discussing how they could integrate the technologies in order to get the best usage scenario in a company.

This paper is structured as follows: in section 2 we will introduce a theoretical framework and the algorithms used by the system. Section 3 describes the heuristics that can be applied in order to estimate the collaboration in a chat, while section 4 presents several results. We end the paper with conclusions and future work.

## Theoretical Framework

As a starting point, we used one of Michael Bakhtin's ideas: "Utterances are not indifferent to one another, and are not self-sufficient; they are aware of and mutually reflect one another. These mutual reflections determine their character. Each utterance is filled with echoes and reverberations of other utterances to which it is related by the communality of the sphere of speech communication. Every utterance must be regarded primarily as a response to preceding utterances of the given sphere (we understand the word <<response>> here in the broadest sense). Each utterance refutes, affirms, supplements, and relies on the others, pre-supposes them to be known, and somehow takes them into account" (Bakhtin, 1986).

It is obvious that implementing a computer program starting from this theory will require some simplifications of Bakhtin's ideas. Nevertheless it offers us all the time a perspective from which to investigate a conversation. According to Bakhtin, each utterance adds some aspects to the discussed topics and in the same time it takes into account the aspects revealed by previous utterances. The extent to which an utterance is based on another varies: it can be an explicit answer, or it can contain only an "echo" of the previous one. Simply because the author is aware of the previous utterance indicates an existing link. From this point of view, between any two utterances there is some degree of collaboration, lower or higher. In our system we modeled this as a complete, weighted graph, utterances being the nodes and the weight of an edge being the degree of collaboration between them. We will call this the *collaboration graph*.

This degree of collaboration (weights of the edges) is estimated using some heuristics. A *zone (region) of the chat* is considered a set of at least two nodes corresponding to consecutive utterances in the chat. The terms *zone* and *region* shall be used interchangeably. We define the *total collaboration of a zone* as the sum of the individual collaborations formed between utterances inside that zone. Figure 1 below illustrates this notion. The total collaboration of region *S* is the sum of weights associated to edges that are completely contained in the rectangle. Although the graph of collaboration is a complete graph, for simplicity, in the figure we did not draw all the edges. Note that the total collaboration is not a measure of how good a collaboration region is. A long region will finally accumulate a large total collaboration, without necessary being a good collaboration zone in the sense we want. To solve this difficulty we defined the notion of *attenuated collaboration of a zone* or simply *collaboration of a zone*, which is the total collaboration divided by some function which increases with the zone's length. We now define a *zone with a good collaboration* as a zone having the attenuated collaboration above a threshold.

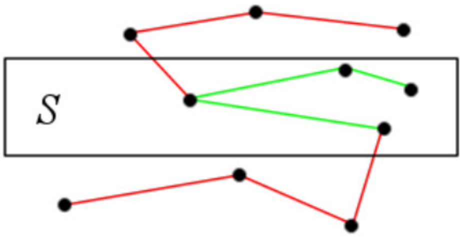


Figure 1. Total Collaboration of a Region.

So far we have defined some notions with the purpose of quantifying a region with a good collaboration. We can at this moment design an algorithm for detecting these zones. First, we will assume that the weights in the collaboration graph are already known, reducing the problem to one involving graphs only. In section 3 we will see methods for actually computing these weights (i.e. estimating collaboration between pairs of utterances).

We analyze the possibility of computing the collaboration for all zones in a chat. This involves computing $\frac{n(n-1)}{2}$ values, where $n$ is the number of utterances, because any two utterances bi-univocally determine a region: the region that is starting from the first utterances in the pair and ends at the second one. Recall that a zone must contain at least two utterances. All these values can be efficiently computed using an algorithm that has a low computational complexity. The key in devising this algorithm is to first compute the total collaboration for each region, pre-computing the values $S(p,q) = \sum_{i=p}^{q} w_{i,q}$. Note that $S(p,q)$ is actually the contribution in terms of total collaboration that utterance $q$ brings to the region starting from utterance $p$ ending at utterance $q$-$1$, when this region is being extended to also include utterance $q$.

Now, that we have the collaboration associated with each zone of the chat, one more aspect remains to be solved. We must select a set of zones in order to present them as high collaboration regions. Simply presenting the zones with collaboration higher than some threshold elicits the problem that overlapping regions will appear. Indeed, starting from a high collaboration zone and removing the last utterance will probably lead to another zone with high collaboration. However, we do not want both these regions to appear in a selection of high collaboration zones. We have used a greedy-type solution, i.e. at each step we choose the best collaboration region that have not been yet chosen or rejected and then reject all other regions that overlap this one. We say that two regions overlap if they have at least one common utterance. In order to implement these ideas, we first sort the regions according to collaboration values, and then we go through them starting from the one with the highest value, for each region checking whether it intersects any of the previous selected ones. Typical values of *k*, the number of selected regions are in the range 10-20. In our corpus chats generally had below 400 utterances. For such values analyzing a chat is almost instant on any computer.

The usage of this greedy approach for selecting regions is not only computationally favorable. We claim that it also makes a good selection of regions. To gain some insight into this, consider the case illustrated in Figure 2. Suppose region *S* has a higher collaboration than each of the regions $T_1$, $T_2$,... $T_n$ which are also good collaboration regions (with collaboration above a threshold) that intersect S. Choosing *S* instead *of* $T_1$, $T_2$, .. $T_n$ seems a good alternative, because while *S* is the zone with the highest collaboration, all the $T_i$ regions are probably good collaboration zones just because they share utterances with *S*.

Figure 2. Example of Using the Greedy Algorithm to Select Collaboration Regions.

## Heuristics for Estimating Collaboration

This section discusses the second part necessary for implementing the automatic detection of collaboration regions. We assumed so far that we know the collaboration score between individual utterances (the weight of the edges in the collaboration graph), and designed the selection algorithms without worrying about these values. In this section, we will present some features used for estimating these collaborations between utterances.

The first feature taken into account is represented by the *explicit links*. During a discussion in the VMT environment, participants can specify before sending a reply that it is a response to a previous one. Therefore, the fact that an utterance $U_2$ has an explicit link to another one, $U_{1,}$ suggests a powerful collaboration between these two replies. Furthermore, if $U_1$ contains an explicit link to another utterance $U_0$, then this link, besides the collaboration between $U_1$ and $U_0$, also indicates some collaboration (however, a lower one) between $U_2$ and $U_0$. This is in concordance with Bakhtin's idea exposed in the previous section. In our implementation, starting from an utterance U, we go back on the chain of explicit links and associate smaller and smaller collaborations between U and utterances found.

Explicit links have some problems that must be treated by other heuristics. Most important, the majority of chat systems don't offer this facility, and even when they do, it is not always used by the participants. If at a given moment of the conversation multiple parallel threads exist, these threads will never be joined by explicit links. It would be useful to determine whether these threads are really independent, or whether they just discuss alternative aspects of the same subject. In the second case, threads are somehow related and overall, the zone should have a better collaboration than it would if the threads were discussing different topics.

When explicit links are missing or just in order to adjust the values offered by them we can use another criterion which should detect links between utterances: *common concepts*. If the same word appears in two different, but nearby replies, then probably there is some link between the two utterances. We don't restrict the search to exactly the same word: if two words have the same stem or are synonyms, they are also considered the same concept, increasing thereby the connection between the utterances.

There can be some zones into a discussion where participants collaborate but outside of the desired topics. These zones are not of interest to us as they are off-topic. In order to increase the bonus for on-topic collaboration, we introduced a criterion which uses a list of keywords. If an utterance contains some of these words then it probably is on topic, therefore the collaborations in which this utterance is involved are increased.

Another feature used to evaluate the degree of collaboration is represented by speech acts and argument models. A speech act represents a function that an utterance possesses in a conversation - some examples of speech acts are greeting, request, complaint, invitation, etc. Generally, utterances that belong to a certain speech act might be formed by a single word or might be arbitrary long. However, for many speech acts some patterns can be identified in which the corresponding utterances fit. By using a pattern matching mechanism, a module of the PolyCAFe system developed under the LTfLL project (www.ltfll-project.org) detects the speech act of each utterance. Elements that belong to the process of argumentation are identified in the same way. These elements correspond (with small variations) to the model introduced by Stephen Toulmin and are the following: claim, ground, qualifier, rebuttal and concession (Toulmin, 1958). Knowing to what category an utterance belongs is useful for our program. For example, an utterance that is a concession probably is involved in a high collaboration with another reply. On the other side, a greeting is not interesting from our collaboration perspective, so although a participant might use an explicit link when greeting another, the collaboration involved should not be taken into account (we only consider on-topic collaboration).

## Results and Heuristics Evaluation

A first important result is obtained by using only explicit links for evaluating the collaboration. A zone obtained using this method is shown in Figure 3. Looking at the graph on the right of the figure (the nodes represent utterances and edges represent explicit links), it can be seen that we have a good edge-density. Although we have shown earlier a series of problems that explicit links have, we mention that if the participants make intensive use of the facility, using this criterion alone for estimating good collaboration regions we obtained very promising results. As stated before, an important problem is that participants don't always use this option.

We have chosen one of the chats from our corpus that contains many explicit links (in this discussion over 75% of the utterances use this facility). We computed the collaboration of every possible region twice:

once using only the explicit-links heuristics, once using other heuristics. Between these two sets of values the Pearson correlation coefficient was computed. If this coefficient is close to 1, then the results obtained by using the second criterion are in concordance with those obtained using only the explicit links. Several alternatives were explored in order to analyze the heuristics, as shown below.

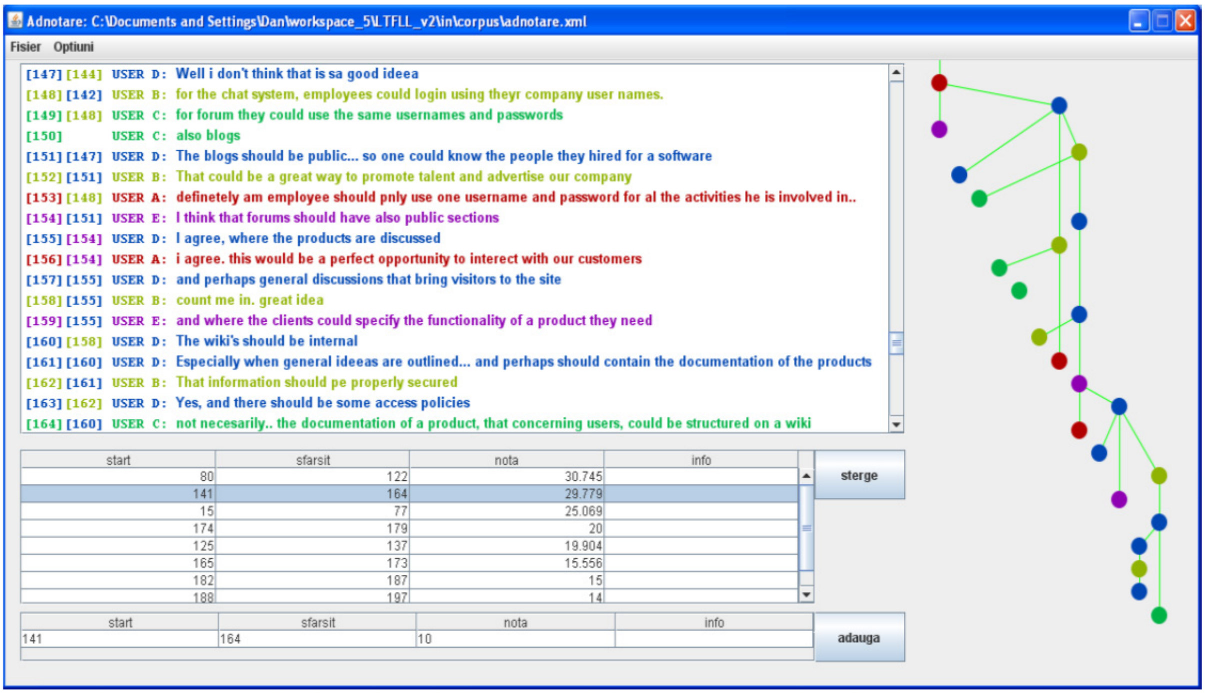
Variant 1: We have used only a restricted version of the criterion "common concepts" defined above, giving a bonus if two replies have a common word – not taking into account neither the distance between the utterances, nor inflationary forms of the same word. We obtained a correlation coefficient of 0.032. This value doesn't indicate the existence of any similarity relative to the values obtained by using only the explicit links.

Variant 2 improves the criterion of common concepts by taking into account the distance between utterances. The intuition is that although two utterances share a common word, if they are far from each other, they are probably not very much related. In order to implement this in our program, the constant bonus accorded for each common word is divided by the distance between the two utterances. This way we have a significant increase in the correlation, which boosted it up to 0.67.

Variant 3: we continue the improvement starting from variant 2, now also using the frequency of the repeated word. In consequence, when a common word is found, the bonus will depend not only on the distance between the two words, but also on the frequency of the word: the more frequent it is in our chat, the less we increase the corresponding value of the collaboration. Thus, we obtained another important increase in the Pearson correlation that was 0.79.

Variant 4: continuing from variant 3, but allowing the words to have the same stem we obtained a small improvement, correlation = 0.80. For this example, the benefits obtained from using stemming are not significant. Further tests should be undertaken in order to more precisely see how useful this feature is.

Variant 5: starting from variant 4 and taking into account some speech acts and argumentation elements, (e.g.: increasing the collaborations that involve a "concession"), we obtained this way a correlation of 0.82. Again, this is a small increase, and the precise effect must be further analyzed. However, overall we obtained promising results, indicating that what can be obtained by using the explicit link structure - which possesses semantic information - can also be obtained using these kind of heuristics.


Adnotare: C:\Documents and Settings\Dan\workspace_5\LTFLL_v2\in\corpus\adnotare.xml
Fisier
Optiuni
sterge
adauga

```
[147] [144]  USER D:  Well i don't think that is sa good ideea
[148] [142]  USER B:  for the chat system, employees could login using theyr company user names.
[149] [148]  USER C:  for forum they could use the same usernames and passwords
[150]        USER C:  also blogs
[151] [147]  USER D:  The blogs should be public... so one could know the people they hired for a software
[152] [151]  USER B:  That could be a great way to promote talent and advertise our company
[153] [148]  USER A:  definetely am employee should pnly use one username and password for al the activities he is involved in..
[154] [151]  USER E:  I think that forums should have also public sections
[155] [154]  USER D:  I agree, where the products are discussed
[156] [154]  USER A:  i agree. this would be a perfect opportunity to interect with our customers
[157] [155]  USER D:  and perhaps general discussions that bring visitors to the site
[158] [155]  USER B:  count me in. great idea
[159] [155]  USER E:  and where the clients could specify the functionality of a product they need
[160] [158]  USER D:  The wiki's should be internal
[161] [160]  USER D:  Especially when general ideeas are outlined... and perhaps should contain the documentation of the products
[162] [161]  USER B:  That information should pe properly secured
[163] [162]  USER D:  Yes, and there should be some access policies
[164] [160]  USER C:  not necesarily.. the documentation of a product, that concerning users, could be structured on a wiki
```

| start | sfarsit | nota | info |
|---|---|---|---|
| 80 | 122 | 30.745 | |
| 141 | 164 | 29.779 | |
| 15 | 77 | 25.069 | |
| 174 | 179 | 20 | |
| 125 | 137 | 19.904 | |
| 165 | 173 | 15.556 | |
| 182 | 187 | 15 | |
| 188 | 197 | 14 | |

| start | sfarsit | nota | info |
|---|---|---|---|
| 141 | 164 | 10 | |

Figure 3. A Good Collaboration Zone Obtained Using Only Explicit Links.

Figure 3 presents a screenshot of the system we developed in order to provide a better visualization of the zones with a high collaboration in a chat. In the upper-left is the chat log. The first number (in square brackets) is the identifier of the utterance, while the second number is the identifier of the explicitly referred utterance. Each participant has a distinct color and their names are automatically anonymized (this option may be disabled in order to show participants' real names).

Below the conversation, we listed the zones detected as having a good collaboration. The first two columns represent the identifiers of the first and the last utterances in the region, while the third column shows the score of the collaboration region as computed by our system. However, the tutor has the possibility to manually edit each of these values, and also to add or remove lines in this table. For convenience we added a fourth column where he can add comments and observations regarding each collaboration zone.

By clicking on a cell of the table, the corresponding region is selected and is represented in the graph displayed on the right part of the image. The nodes correspond to the utterances in the selected region and the edges represent explicit links inserted by participants during the chat. A node has the color of its author, and by moving the mouse above it, the text is shown as a tooltip text.

A verification of the developed system was done by comparing its results with those provided by the PolyCAFe system (Trausan-Matu and Rebedea, 2010, Dascalu, Rebedea and Trausan, 2010) and those manually identified by other persons than the developer. The results of the system were very similar with those of the human and better than those of PolyCAFe.

## Conclusions and Future Work

We implemented a system for automatically identifying the zones of a chat with a good collaboration. First we created a theoretical framework, being guided by an idea stated by the Russian philosopher Mikhail Bakhtin. We considered that between any two utterances there is some degree of collaboration. Starting from this point we derived a few notions leading us to defining what a good collaboration region is and implemented the algorithms that allow us to efficiently extract these regions. Besides this theoretic framework, we also devised a couple of heuristics that estimate the collaboration between a pair of utterances. In this paper we described and analyzed these heuristics. We have shown that the explicit links which are possessing semantic information can be approximated using some heuristics that are based on a rather lexical analysis. These initial results are promising and future research includes testing new heuristics, for example some based on Social Network Analysis (perhaps collaboration appears more between persons with a similar rank). Another improvement would be to take into account more powerful semantic similarity measures that would extend the lexical and WordNet based ones defined in the paper: to this extent, Latent Semantic Analysis, like in Dong's (2006) system and more powerful semantic distances or lexical chains could be used.

## References

Bakhtin, M. (1986). Speech Genres and Other Late Essays. *University of Texas, Austin*.

Dascalu, M., Rebedea, T. & Trausan-Matu, S. (2010). A Deep Insight in Chat Analysis: Collaboration, Evolution and Evaluation, Summarization and Search, *AIMSA 2010*, LNAI 6304, Springer, 191-200.

Dong, A. (2006) Concept formation as knowledge accumulation: A computational linguistics study, *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 20, 1, 35-53

Rose, C. P., Wang, Y.C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., Fischer, F. (2008). Analyzing Collaborative Learning Processes Automatically: Exploiting the Advances of Computational Linguistics in Computer-Supported Collaborative Learning, *International Journal of Computer Supported Collaborative Learning*, Vol.3, nr.3, 237-271

Stahl, G.: (2006). *Group Cognition: Computer Support for Building Collaborative Knowledge*, MIT Press.

Stahl., G. (Ed.). (2009) *Studying Virtual Math Teams,* Springer, Boston.

Toulmin, S., (1958). *The Uses of Arguments*. Cambridge Univ. Press.

Trausan-Matu, S., Dessus, P., Rebedea, T., Mandin, S., Villiot-Leclercq, E., Dascalu, M., Gartner, A., Chiru, C., Banica, D., Mihaila, D., Lemaire, B., Zampa, V., Graziani, E., *Learning Support and Feedback, LTfLL Deliverable 5.2,* http://ltfll-project.org/tl_files/LTfLL-Deliverables/LTfLL_D5.2.pdf, downloaded at November 7, 2010

Trausan-Matu, S., & Rebedea, T. (2010). A Polyphonic Model and System for Inter-animation Analysis in Chat Conversations with Multiple Participants. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (Vol. 6008, pp. 354-363): Springer Berlin / Heidelberg.

## Acknowledgments

The research presented in this paper was partially funded by the projects no. 212578, FP7-ICT-2007-1-4.1 LTfLL (http://www.ltfll-project.org/) and project no. 264207, FP7-REGPOT-2010-1 ERRIC - Empowering Romanian Research on Intelligent Information Technologies (http://erric.eu). The authors would like to thank all the students and tutors at the University "Politehnica" of Bucharest that were involved in the validation experiments.