# Finding the real subject: the application of categorisation methods to forum messages

G. La Rocca
*Department of Demographic, Sapienza, University of Rome, Italy*

## Abstract

The growth of interactive communities and multimedia nets seems to be uncontrollable to the extent that both human and global communication are impossible today without the Internet. Nowadays Computer Mediated Communication has created and developed on-line communities that well in digitalized territories. This new segment of social science is mostly built up of "words" and the necessary strategy and tools should be aimed at enabling the scanning, the attentive reading, the understanding and the explanation of such segment. On-line communities are perfect places to study linguistic behaviours and mass interactions. The virtual space in which people can speak are chat lines, mailing lists, forums, etc. In a forum we have the possibility to send one or more messages, participating to the discussion of a specific topic, in this case wine. This forum gathers specialists, technicians and ordinary people who end up becoming a real community, where information and news about wine are frequently swapped. The study of a forum is a method normally used to verify customer satisfaction; in our case we want to find out if there is a parallelism between the titles of the subject and the content of the messages. Because we frequently answer messages without changing their title, to monitor only these titles is not enough for marketing or communication specialists if they want to obtain an exhaustive description of customer communication. The purpose of this work is to obtain a new list of subjects for each message. For this reason we use Text Mining techniques which allow us to look for sets of words inside the texts. TaltaC² "Entity Search" utility has been used in order to search a distinctive word sequence inside fragments by using complex queries with regular expression.

*Keywords: text mining, automatic categorization, entity search, computer mediate communication, forum.*

WIT Transactions on Information and Communication Technologies, Vol 40, © 2008 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA080121

## 1 Communities and interviews

Communication technologies have left their mark in the history of human development. The Medium Theory [8] considers the media as «environments» where interactions among individuals take place, contributing to the creation of "technologically structured communities". This is in contrast with the idea that means of communication are passive containers inside which information travels. The new media must be considered as platforms where social groups can face each other creating forums for the discussion of crucial social life issues [7]. The introduction of new media gives social groups the opportunity to re-examine and question the models used by the old media and which are the heart of social exchange. The history of the medium itself starts from the organization of a variegated public around these tools; the consequence is that the new customs, the new behaviours of the public represent a change of the old habits no longer functional in the new contexts. To observe the environment built by the Internet, one can analyze different areas of interest, among which the study of non-verbal skills, of "spatial" behaviours, of linguistic and extralinguistic behaviours. With the current technological devices only one of these areas can be analyzed through the Internet: the linguistic behaviour. Paccagnella [9] has identified some cases in which the quantitative analysis of on line language hides distortions caused by the very observation of the dialogues and he has underlined that there can be a different dynamic in chatting by CMC (Computer Mediated Communication) towards real conversations. The time required to type and the delays in answering, when chatting, may alter the style of the conservation, distorting it. This consideration is often not considered sufficiently in analyses of this kind. In addition, written messages are often "decontectualized" from the place where the conversation is held: one sits in front of the keyboard in his/her room and can communicate with the entire world. The participants are not always really involved in the interaction (sometimes they only play a part). Much more than in real conversations, communication by CMC loses perhaps some of its sense and of its meaning when messages are re-read and then changed. This modifies both the intention and the sense of involvement of the participants. CMC, however, offers researchers a wide range of linguistic observation possibilities.

## 2 The forum

The virtual places, where there can be communication between two or more individuals, are virtual communities, chat lines, mailing lists and alike. Object of this analysis is a forum. In forums one has the possibility to send one or more messages participating to the debate going on in one section of the forum itself or participating to various debates if the forum is structured in different topics. Naturally, it is possible to interact with other users only after registration, that is after the login. Forums are real «moments» of debate on prearranged topics between two or more people who have an interest in common and decide to exchange ideas, opinions and experiences on it in order to increase their knowledge and share this knowledge with others. Everyone is free to voice

his/her opinion and to exchange ideas with a virtual interlocutor, but, if necessary, the moderator can silence an idea, a thought, an opinion of any kind without giving importance to its provenance. The opinions expressed by «wine and food opinion makers» are known to be «enthusiastic» and they represent ideas that some «experts», strongly interested in the topic, leave in a forum to let their voice be heard. The expressive form remains, however, a sort of written orality. Who sends a message, does it to voice an opinion, to open a debate on a particular aspect of a specific topic. For who gives it, this is the opinion of an expert: here emerges the concept of self-expression, talked about by Atkinson and Silverman [1]. The opinion is "visible, open and strongly felt", and it develops a passionate communication about a specific topic. In this paper we have chosen to analyze a forum on wine in the section "Disputes and Opinions". Wine is a drink which in recent years has gained significant importance with reference to local development phenomena, but also in terms of business.

The analysis of the 181 messages sent to the forum aims at:

- outlining the pre-treatment procedures of the messages downloaded from a virtual place and therefore full of noise;
- showing the course of the debate about the "product" wine, as it occurred in an arena filled with enthusiastic and/or qualified people;
- extracting the textual categories connected to each topic. Often inside a topic in fact – with reference to a specific subject – the debate starts off moving away from the initial input given by who sent the first message.

## 3 First analysis operations

The corpus was built downloading the texts of the messages contained in the pages of the forum chosen (we prefer don't write the name of the forum, because we sent an email to the editorial staff team asking for their collaboration in the retrieval of the texts. We did not, however, receive a reply.). Viewing the discussion pages, we noticed a lot of noise in the messages, the particularity of the language used and the presence of orthographical mistakes. To work on a HTML text of this kind, we must follow a serious of significant preliminary operations. At the top of each forum page there are three commands which allow us: to mark out the topic – like in a work flow system –, to post the topic via e-mail and to print it. The printed version of the topic has a more reduced level of noise; in fact the images disappear with the emoticons contained in the digital signature and the quoting seems easier to manage. The next operation consists in saving the message directly as text file and then in removing from the files all the Internet tabulations. In this first procedure, we used a specific software, TexPad, thanks to which it was possible to eliminate some forms of noise. The use of TexPad allowed us:

- to identify and eliminate tabulations;
- to eliminate the empty lines;
- to introduce keys useful when passing in TaltaC².

The txt format file so obtained has an extension of 7,82 MB and it is structured in 181 keys, the number of the topics downloaded from the forum. The file was inserted in TaltaC² by creating a new work session. Entering from the menu file, you can open an existing session or create a new one. By default the software creates a session folder inside the program folder, but you can create one in any other free space of your pc. Once the folder has been created, you can insert the file you want to work with using the program. The next operation consists in carrying out the parsing, the tagging and the normalization. From the vocabulary created thanks to TaltaC², it is possible to identify the first corpus lexicometrical measurements.

Table 1: Lexicometrical measurements.

| | |
|---|---|
| Words Token (N) | 890.648 |
| Words Type(V) | 39.034 |
| Type/Token Ratio (V/N)*100 | 4,383 |
| Hapax % | 30,748 |
| V/√N – Giraud's Index | 41,361 |

Giraud's index (41,3%) and the hapax percentage (30,7%) connote the text as a rich one; this depends on the particularity of the language used, but also on the topic considered and on the freedom of expression one has when discussing online.

## 4 The debate on wine

After carrying out the pre-treatment, we started exploring the text using the traditional strategies of lexico-textual analysis. We therefore proceeded with a lexical analysis evaluating the particular language by taking out the peculiar words and by analyzing the repeated segments. The fulfilment of these first analysis procedures gave us useful information about the words used in the text and about their peculiarity within the sub-corpus. After entering the headwords we compared them intersecting the vocabulary with a frequency lexicon: the Polif2002 [A list of graphical forms (occ>1] available in TaltaC². This way we were able to extract the characteristic forms: words which are over or under used in comparison to the reference model. The peculiarity was calculated in terms of intrinsic positive and negative specificity using a standard relative frequency spread (the spread is expressed by the formula $Z\,t = (ft - ft^*)\backslash (ft^*)^{1/2}$ in which ft is the number of the normalized occurrences, ft* is the correspondent value in the reference lexicon and $(ft^*)^{1/2}$ is the relative frequency average quadratic spread [3]). By calculating this index, we divided the graphical forms of the vocabulary in key words – over\underused – and trivial words, which have a spread proximate to zero, and are used therefore with the same frequency both in the text body and in the reference model. We considered trivial those words with a statistical spread between or equal to + 0 – 0.9. We considered underused

words those with a spread lower than –0.9. Finally, we considered overused words those with a spread higher than +0.9 [3].

Table 2: Distribution of the graphical forms according to trivial, overused and underused words.

| | |
|---|---|
| Trivial words | 718 |
| Overused words | 15.973 |
| Underused words | 4.466 |
| Total | 21.157 |

The results have pointed out a clear prevalence of underused and non trivial words in comparison with the frequency lexicon. This first distribution indicates immediately the specificity of the text considered: this depends on the characteristics of the means of communication and on the message diffused by it. To identify the great absentees in the wine debate, we must go beyond the twenty-first position (the order follows the growing spread) of the underused words. The presence of many empty words can depend on the topic discussed and on the communication modalities. It is only after the twenty-first position that we notice the presence of full graphical forms, among which there are neither verbs nor adjectives and the near totality of nouns refer to the social-economic sphere.

Table 3: Full graphical forms in the first 190 positions (the order follows the growing spread) among these underused words.

| **Type words** | **Deviation on the occurrences** |
|---|---|
| development | -8,691024 |
| society | -8,540772 |
| world | -8,028287 |
| law | -7,15863 |

From the analysis of these graphical forms clear is the absence in the debate of the social-economic dimension of wine, confirming therefore the orientation of the forum as a place for wine-lovers. Among the underused terms there is also the importance of wine as a factor of “development”. To find the protagonists of the on-line debate, we consider instead the interpretation of the graphical forms with a spread higher than 10. What results first from the reading of the list is the high spread for the words “wine”, “memory”, “bottle”. The topic words are mixed up with the key words and the result of the spread for “wine”, the subject of the forum, is predictable. However, the presence in the first positions of the word “memory” shows how wine tasting is set in our sensorial, perceptive sphere and it remains an intimate experience. To sip wine appears to be a moment of rapture, connected to the “quality” of the wine itself; this perceived by the sense of “smell”, by the “taste”, which allows us to notice the “scent” of “wood” in the “glass”.

Table 4: Full graphical forms with a spread higher than 10 among the overused words.

| Type words | Deviation on the occurrences |
|---|---|
| wine | 198,792 |
| memory | 85,712 |
| bottle | 85,474 |
| canteen | 81,995 |
| euro | 77,806 |
| nose | 52,764 |
| glass | 34,442 |
| scent | 31,791 |
| wood | 31,187 |

So far our knowledge of the subject has allowed us to identify the general topics related to the topic wine. But we do something different when we start looking for "fragments" or tags capable of parcelling out the topics in activities, tastes and preferences.

## 5 The categorisation procedure

When carrying out a categorisation procedure, the use of these techniques must be considered almost compulsory, because although they tell us very little about the achievement of the categorization objective, they are essential if you want to reach it. In fact, in the creation of categorial variables we must take into account the words present in the text and the local contexts in which they are inserted, if we want to be pertinent and reduce the manual work of revision. This to create very effective tags when summarizing a text. The objective of this paper is to dispose of a list of topics – some of them standardized – discussed online, to try and see in what way a subject-matter talked about online differs from the initial topic. In order to identify and categorize the activities inside each fragment, we must have at our disposal Text Mining tools which allow us to look in the text for sets of words (entities). We were able to use this strategy thanks to the new functions present in TaltaC².

The starting point of the categorization procedure is the list of repeated segments. TaltaC² identified over 80.593 segments, counting also those with just one occurrence. From this list we took out only the segments which could be traced back to a specific topic; in other words, those which contained a topic idea, for example, "eating well" or "wine production", distinguishing them from other segments, always useful but not specific enough to identify the subjects of the discussion, such as, for example, "red wine" or "brunello di montalcino".

The list created following this procedure, with something like 6.223 segments, is the starting point of the categorization of the topics using the entity search utility supporting regular expressions existent in TaltaC².

This function considers as unit analysis not the single textual expressions, but the entire fragment, looking for a particular expression or combination of

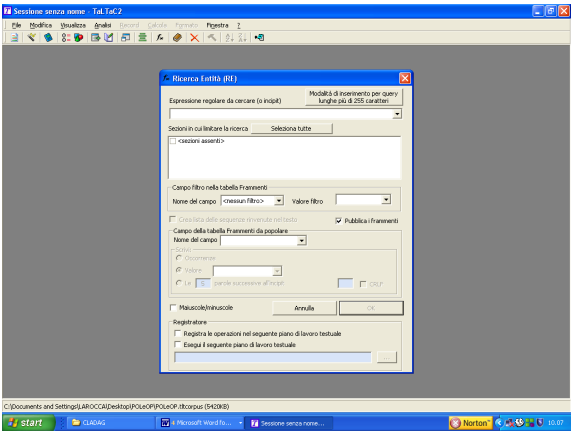


Figure 1: The TaltaC² "entity search" utility.

expressions and writing the result obtained by the researcher in a new personalized variable that is added to the initial corpus. The search in the text of useful information capable of feeding this new variable is carried out thanks to a query defined by *regular expressions* [2]. These use:

- Boolean operators (and, or, and not),
- Lexicomatic reductions (thanks to the jolly characters * and ? like, for example, produ* and vin?),
- LAGgxx distances between words connected to each other. For example, if we work on "produ* LAG3 vin?", we are able to identify different strings such as: "produrre il vino" ("producing wine"), "produzione vinicola" ("wine production"), "produttori di vino" (wine producers"). For these three strings we could insert "produzione vinicola" ("wine production") as the new value for the new variable "Subject". It is clear that the logical function of categorization is to subsume in a very wide category variables with groupable characteristics.

The automatic tagging – which allows us to use TaltaC²'s RE function – facilitates the "reduction of complexity" as well as the construction of a list/archive of useful elements capable of summarizing and cataloguizing the information. The automatization of the process allows us in fact to tag more strings at the same time. We must also point out that this procedure has been used in other works [4], obviously with different research objectives. It was then that procedure applied to the rest of the texts.

For example, starting from just 156 fragments, this tool made it possible to tag 4.107 lines, that is 0,9% of all the fragments.

Before considering the work as it appears finished, we must reconstruct the text replacing the initial verbal forms with the headword expressed in the infinitive. When we export the file, we can then work on a matrix containing a new variable (Subject), inside which we find codified the new tag obtained thanks to the categorization procedure.

If we export the file to a worksheet, such as excel, we can clean the variable Subject codified manually by the researcher who has to evaluate the pertinence of the category identified with the text fragment it refers to. The excel work file is therefore configured as follows: sub-corpus code, forum topic, identity of the person who posted the message, the category "Subject" which is automatically applied, a new variable called the real subject and, finally, the text it refers to. The researcher will then have to evaluate the pertinence of the category created and validate it using a semi-automatic procedure.

## 6 Final considerations

Although the text requires a manual control system, at the end of the manual process are obtained 4.653 real subject compared to the initial 181. Obviously some of these have a high frequency. Now, the results of the categorization procedure allows us to parcel out and/or divide the topics in discussion, that is to identify the dimensions of a macro-container. If we gather in two areas the tags identified – in the same way in which indicators are expressed in indexes – we obtain two macro-categories: talking about wine and sociability. Within the first category fall 42,7% of the tags created while 57,3% belong to the sphere of sociability. The macro-category sociability is divided as follows: greetings/salutations (15% of the total), personal opinions (19%), wine tasting capabilities (20%), reserves (25%), production (11%), festivals/fairs (10%). Talking about wine: organoleptic characteristics (15%), production cost (20%), area of production (15%), pictures (20%), structures and accessories (18%), matching food and wine (12%).

Table 5: The List of macro-category and category.

| **Macro-Category** | **%** |
|---|---|
| Sociability | 57,3 |
| **Category** | **%** |
| greetings/salutations | 15% |
| personal opinions | 19% |
| wine tasting capabilities | 20% |
| reserves | 25% |
| production | 11% |
| festivals/fairs | 10% |
| **Macro-Category** | **%** |
| Wine | 57,3 |
| **Category** | **%** |
| organoleptic characteristics | 15% |
| production cost | 20% |
| area of production | 15% |
| pictures | 20% |
| structures and accessories | 18% |
| matching food and wine | 12% |

## References

[1] Atkinson, P., Silverman, D. "Kundera's Immortality: The interview Society and the Invention of Self", in Qualitative Inquire, 3 (3), pp. 324–345, 1997.
[2] Bolasco, S. *Analisi dei diari giornalieri con strumenti di statistica testuale e text mining*, in: ISTAT, *I tempi della vita quotidiana*, on proceeding, 2007.
[3] Bolasco, S. *Analisi multidimensionale dei dati,* Roma: Carocci, 1999.
[4] Castelles, M. *The Information Age: Economy, Society, Culture*, Oxford: Blackwell Publishers Ltd. 2000.
[5] della Ratta, F., Lorè, B., & La Rocca, G. "Textual analysis perspectives on categorisation of activities in Istat survey on occupations", on *Book of Short Paper. Meeting of the CLAssification and Data Analysis*, Macerata: EUM pp. 263–266, 2007.
[6] Levì, P. *L'intelligence collective. Pour une anthropologie du cyberspace*. Paris: Éditions La Découverte, 1994.
[7] Marvin, C. *When Old Technologies Were New*, 1988 (it version *Quando le vecchie tecnologie erano nuove*, Torino: Utet Libreria, 1994).
[8] Meyrowitz, J. "Medium Theory", in Crowell D., Mitchell D. (eds.), *Communication Theory Today*, Cambridge: Polity Press, 1994.
[9] Paccagnella, L. *La comunicazione al computer. Sociologia delle reti telematiche*, Bologna: Il Mulino, 2000.