

Using text mining to understand the call center customers' claims

G. M. Caputo, V. M. Bastos & N. F. F. Ebecken
COPPE – Federal University of Rio de Janeiro, Brazil

Abstract

This work presents a study related to the methods of text mining applications, more specifically, data clustering, in call center's databases, whose texts are in the Portuguese language. The main objective is to identify new and useful knowledge, based on customers' claims. Through the information agreement, it will be possible to identify better ways to help the customer, increasing their satisfaction with company services as well as supplying the call center staff and other related areas with a set of procedures and information concerning the most common customers questions.

Keywords: text mining, clustering, call center.

1 Introduction

Many companies who give telephonic attendance services for its customers need solutions guided to the information technology that make possible the register of their customers contacts, the aid and support to the attendants. For such, new access technologies to the knowledge bases and flow of information management can increase the productivity of the attendants and, as consequence, improve the quality of the given services.

Many times, the attendance can be made by the Internet, bringing some advantages such as: costs reduction with attendance, user satisfaction increasing and knowledge of the customers interests, gotten through the automatic storage of the typed information.

Text mining tools directed to knowledge bases creation, support and operation, and for attendance centers are applicable for the Web cases email or telephone attendance.

The technologies presented in these tools make possible to create knowledge bases for queries through the search of one or more words in documents, allow to



combine resources of search in texts with resources of interaction with support (attendance in the Web: email, chats, virtual rooms, etc), and make possible the creation of knowledge bases for queries in natural language.

Thus, the objective of this work is to show some methods of text mining applied in call center's databases, especially in customers' complaints of a electrical energy company, where the results serve to show that techniques are extremely useful during the company services evaluation processes, as also to be used in decision taking. For this, a text mining tool called Insight Discoverer Clusterer by Temis [1] was used, that contains the documents grouping functionality or *clustering*, allowing the classification of the same ones.

With this study we have created conditions that can to evidence problems that are occurring with frequency, as for example problems of invoicing, financial income, supply and attendance quality, etc, who are gotten by the analysis of the results.

2 Text mining

Text mining [2] is the hidden knowledge extraction process in a great set of not structuralized literal documents, using advanced technology.

The use of text mining tools, identified in [3], facilitates the manipulation and understanding of the available electronic documents increasing volume not only in the Internet, as in companies, providing as resulted a set of information that will go to assist in the decision taking processes and call centers automation.

In the call center case, the data are gotten from customers' calls registers that desire to complain on the service. Each call reason summaries are typed and stored in databases. Text mining applied to customers' claims data search specific standards in the reasons for which call occurs. The objective is to group the similar claims and, through the results analysis, get a summary on the claims main reasons.

The clustering process occurs in two distinct steps: pre-processing and post-processing.

Pre-processing is the step where the base preparation is made to identify the excellent information and to eliminate the data that does not add value in the content. Moreover, in the call centers great majority, the calls register typing is done manually by the attendant, after the call end. Therefore, the base is susceptible to typing errors, grammatical errors and language vices, as abbreviations and acronyms. These errors and language vices can be interpreted as a Call Center Metalanguage. Pre-processing also is responsible for such errors correction. This process corresponds to:

- Stop words elimination, or either, words elimination that are not excellent for the analysis process.
- Stemming algorithms [4] application, the words reduction to its radical.
- Synonymous conversion, grouping different words, but that they possess the same meant semantic. This stage includes the related abbreviations and other terms conversion.

- Metalinguage correction, eliminating typing errors and grammatical errors converting to the correct form.

After pre-processing, the data is prepared to pos-processing. This means calls registers grouping, where, through the clustering and results analysis, it is possible to identify existing patterns in the base and to get a real knowledge on the main customer's claims reasons.

For accomplishment and results interpretation easiness, some text mining tools had been developed and offer diverse options. Amongst most used are: Temis, Megaputer Intelligence, SAS, SPSS, Synthema [5] and others are the best popular.

Some tools present many languages support, between them, the Portuguese Language. The IDC [6] module (Insight to Discoverer to Clusterer) developed by Temis, is a server that groups documents according to their semantic similarity. Documents that are similar can share one or more topics inside the clusters. As output, this module presents a visualization that shows clusters, the topics inside each cluster, and subclusters, connected to the main clusters, providing the hierarchical structure of documents within clusters [7, 8].

3 The database

One of the text mining objectives is the discovery of stored knowledge in historical databases. It has as information source the generated and stored literal data throughout all the process. A call center generates daily a great amount of data that include since the customer attendance phase, until the end of solicitation of a determined task. To focus the customers and understand their necessities, doubts and suggestions it is necessary for continuous customer services improvement.

The sample analyzed is referring to the attendance process during all year of 2004 and is classified in accordance with three types of customers' contacts: Request, Consultation and Claim.

Thus, the analysis was made on data generated from 177063 claims presented by the customers, through communication channels that include: virtual agency, telephonic call, chats, mailing, manifestations book, among others.

When the customer searches one of these attendance canals to carry through its claims, a register in the database is created, which store among other informations, the reason detailing the information.

In accordance with the reason presented for the customer, the attendant classifies the register as being: financial income, customers' attendance, invoicing, insolvency, supply quality and attendance quality.

The methodology adopted as the solution of the decision taking problem, aims to adopt the best strategy for the intelligent use of call center data. For this, it is necessary to test data treatment options, based on text mining technique, which best satisfies the problem.

Some steps must be considered during the study of the database, to understand the information contained on it and to take care the information recovery requirements.



As the database created in call center consists of information inserted for the customer attendance staff and they are typed during the telephone call, the database is entirely susceptible to have as many typing errors, abbreviations and vices of language.

3.1 Metalanguage pre-processing

The existing vocabulary in the call center database is directly proportional to the type of carried through attendance, or either, sufficiently summarized. In this case, the occurrences of terms in the text or are restricted to a group of terms very used by many attendants or are specific for each type of request, making the frequency very high for some terms or very low for others. In general, more than 60% of existing terms in the database have linear frequency between 1 and 2.

In this way, the work executed in the pre-analysis of the information was cleaning and correction of metalanguage format so that text mining algorithms are applied. In the cleaning process, more than 3000 proper names, users logins, separators used by the attendants, as for example, “xxxx”, typist errors and terms that do not have added knowledge, as for example, prepositions and articles had been discarded and are part of *stop words* list. In the correction process, around 200 entrances were created for a terms conversion dictionary, that appear in the text of some form or typed incorrectly (for ex: pagto, pgto and pg for pagamento - payment).

In order to adjust the data format to the Insight Discoverer Clusterer tool, some text archives were generated, each one had been created corresponding to a request.

3.2 Clustering process

For this process the sample was divided in many archives, where each one contains a unique request made by the customer. Moreover, the tool executes previously a syntactic and semantic classification for each term of the archives, identifying if the term is a name, a verb or an adjective. The obtained results presented a clustering with strong influence of the classification used for the company during the customer attendance process, presenting the terms related with the classifications that have the biggest occurrences. After, the obtained results were compared with some classes of requests.

The clustering procedure, using some archives of data, identified 10 different groups that are almost the some classification criteria adopted by the company practice.

4 Case study

In the present study, three call reasons categories had been considered. The categories had been clustered separately aiming to understand the claims call general reasons. The clustering result is shown in the following.

4.1 Financial income

In the power electric industry there are several complained problems related to Financial Income. In these topics, some can be detached as more important and more common. Figure 1 illustrates the customers' claims database clustering in Financial Income area. With these clusters, the problem can easily be observed, and consequently, detected.

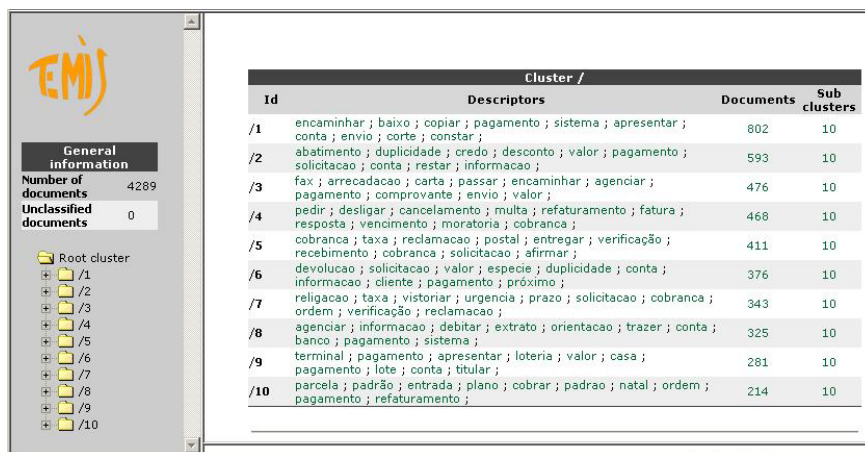


Figure 1: Output for category financial income.

Observing the generated clusters, it can be noticed that the main detected contents are:

- Payment – it appears in all clusters as demonstrated by: “abatimento” (reduction), “conta” (bill), “comprovante” (receipt), “cobrança” (collection), “fatura” (invoice);
- Bill duplicity – related to cluster 2 by: “duplicidade” (duplicity), “pagamento” (payment), “desconto” (discount).

4.2 Invoice

The customers' bills invoicing is, probably, one of the problems that more receive calls in all the existing topics in a company from the power electric industry.

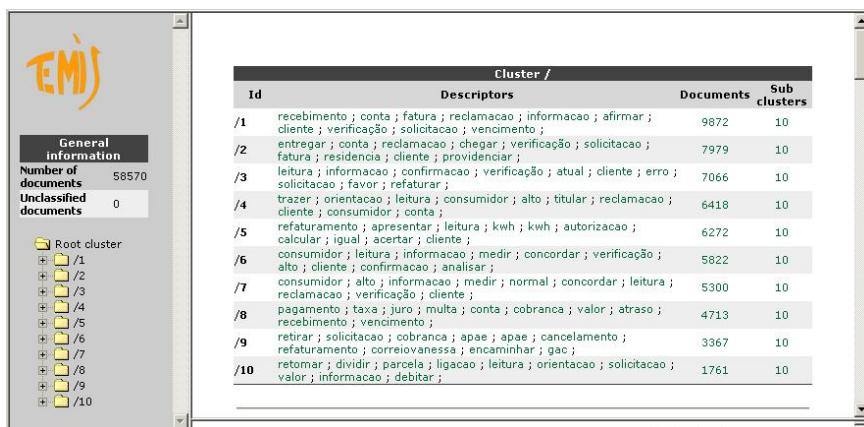
Figure 2 shows the database clustering result of the Invoice category carried through ten clusters.

From all clusters, three are considered initially more important:

- Bill value claim – as demonstrated in: “leitura” (lecture), “valor” (value), “conta” (bill), “reclamação” (claim), “alto” (high), presented in all clusters;
- Ask for a bill copy – related in: “reclamação” (claim), “conta” (bill) and “solicitação” (request), but in subclusters of cluster 1 it is more visible, by the term “2ª via” (copy);

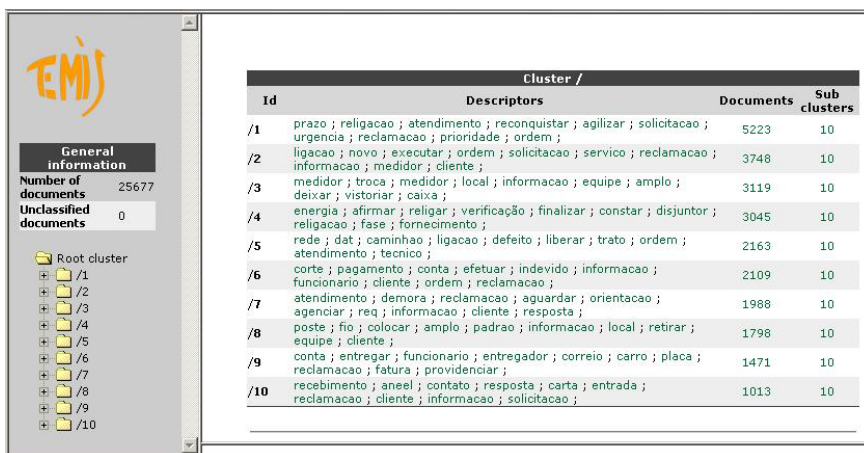
- Payment delay – explicit in: “atraso” (delay), “multa” (fine), “juros” (interest). It is shown in subclusters of cluster 8;
- Invoice – it occurs frequently in clusters 3,5,9 and all other subclusters.

Many of the keywords founded in the invoicing are common to those finding in Financial Income. This must be caused by the fact that both categories be indirectly linked for the subject “payment bill”.



Cluster /			
Id	Descriptors	Documents	Sub clusters
/1	recebimento ; conta ; fatura ; reclamacao ; informacao ; afirmar ; cliente ; verificacao ; solicitacao ; vencimento ;	9872	10
/2	entregar ; conta ; reclamacao ; chegar ; verificacao ; solicitacao ; fatura ; residencia ; cliente ; providenciar ;	7979	10
/3	leitura ; informacao ; confirmacao ; verificacao ; atual ; cliente ; erro ; solicitacao ; favor ; refaturar ;	7066	10
/4	trazer ; orientacao ; leitura ; consumidor ; alto ; titular ; reclamacao ; cliente ; consumidor ; conta ;	6418	10
/5	refaturamento ; apresentar ; leitura ; kwh ; kwh ; autorizacao ; calcular ; igual ; acertar ; cliente ;	6272	10
/6	consumidor ; leitura ; informacao ; medir ; concordar ; verificacao ; alto ; cliente ; confirmacao ; analisar ;	5822	10
/7	consumidor ; alto ; informacao ; medir ; normal ; concordar ; leitura ; reclamacao ; verificacao ; cliente ;	5300	10
/8	pagamento ; taxa ; juro ; multa ; conta ; cobranca ; valor ; atraso ; recebimento ; vencimento ;	4713	10
/9	retirar ; solicitacao ; cobranca ; apae ; apae ; cancelamento ; refaturamento ; correioanessa ; encaminhar ; gac ;	3367	10
/10	retomar ; dividir ; parcela ; ligacao ; leitura ; orientacao ; solicitacao ; valor ; informacao ; debitar ;	1761	10

Figure 2: Output for category invoice.



Cluster /			
Id	Descriptors	Documents	Sub clusters
/1	prazo ; relacao ; atendimento ; reconquistar ; agilizar ; solicitacao ; urgencia ; reclamacao ; prioridade ; ordem ;	5223	10
/2	ligacao ; novo ; executar ; ordem ; solicitacao ; servico ; reclamacao ; informacao ; medidor ; cliente ;	3748	10
/3	medidor ; troca ; medidor ; local ; informacao ; equipe ; amplo ; deicar ; vistoriar ; caixa ;	3119	10
/4	energia ; afirmar ; religar ; verificacao ; finalizar ; constar ; disjuntor ; relacao ; fase ; fornecimento ;	3045	10
/5	rede ; dat ; caminho ; ligacao ; defeito ; liberar ; trato ; ordem ; atendimento ; tecnico ;	2163	10
/6	corte ; pagamento ; conta ; efetuar ; indevido ; informacao ; funcionario ; cliente ; ordem ; reclamacao ;	2109	10
/7	atendimento ; demora ; reclamacao ; aguardar ; orientacao ; agenciar ; req ; informacao ; cliente ; resposta ;	1988	10
/8	poste ; fio ; colocar ; amplo ; padrao ; informacao ; local ; retirar ; equipe ; cliente ;	1798	10
/9	conta ; entregar ; funcionario ; entregador ; correio ; carro ; placa ; reclamacao ; fatura ; providenciar ;	1471	10
/10	recebimento ; anel ; contato ; resposta ; carta ; entrada ; reclamacao ; cliente ; informacao ; solicitacao ;	1013	10

Figure 3: Output for category attendance quality.

4.3 Attendance quality

Many are the claims calls in relation to the employee’s attendance and to the given services. Figure 3 presents the call registers clustering result.

The problems that must be considered are three:

- Meter problems – related to company equipments defects, as shown by: “medidor” (meter), “disjuntor” (circuit breaker), “defeito” (defect), “vistoriar” (inspect), “fio” (wire), “poste” (lamppost), “rede” (grid) e “fase” (phase);
- Service claim – related to customers attendance problems. As noted in: “trato” (treat), “funcionário” (employee), “reclamação” (claim), “atendimento” (attendance) e “técnico” (technical);
- Urgency in the attendance – associated to: “urgência” (urgency), “prioridade” (priority), “atendimento” (attendance), “demora” (delay), “aguardar” (wait), “agilizar” (streamline), dentre outras.

5 Conclusions

The performance of the tool and the quality of the obtained results was improved using a specific cartridge to Portuguese language. This facility contains particular properties that are very helpful.

In accordance with the obtained results, it must be stressed that the sample of data, as well as the activities carried through the pre-processing step and clustering process are of fundamental importance to the knowledge extraction.

In our experiments the traditional categorization process built during the company practice was confirmed, and could be validated. Although promising results have been achieved in this work, there are some issues that can be further investigated by the call center experts. But we can conclude that now is possible to implement an automatic classification system to online monitor the service quality.

Finally it can be concluded that the obtained text mining results can be used in the power electric industry to:

- Understand the needs of customers while accessing preferred customers;
- Provide service menus for locking in preferred customers, such as offering discount charge menus and added-value services;
- Analyze how both electric power income and related income change, and investigate the contents of the services;
- Measure whether customers offered services are satisfied with those services and reflect the results in the service menu planning.

Acknowledgement

We are grateful to the Brazilian Research Agencies CNPq and FAPERJ for their financial support.

References

- [1] Insight Discoverer Clusterer (IDC) – Developer’s Guide, Temis Company, 2002.



- [2] Spinakis, Dr. Antonis, Text Mining: A Powerful Tool for Knowledge Management, Managing Director of QUANTOS SARL.
- [3] Lopes, Maria Célia S., 2004. *Mineração de Dados Textuais Utilizando Técnicas de Clustering para o Idioma Português*, Tese de Doutorado, COPPE/UFRJ.
- [4] Porter M., An algorithm for Suffix Stripping, *Program*, v. 14(3), 1980, 130-137.
- [5] ZANASI, A., 2005, *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*. 1 ed. Great Britain, WIT Press.
- [6] Insight Discoverer Clusterer (IDC) – Developer's Guide, Temis Company, 2002.
- [7] Zanasi, A., Web Mining through the Online Analyst, *Data Mining II*, N. Ebecken & C.A. Brebbia, 2000.
- [8] Zanasi, A., Text Mining: the new competitive intelligence frontier. *In VST2001 Barcelona Conference Proceedings – IRIT*, Spain, 2001.

