Opinion Mining: Taking into account the criteria!

Pascal Poncelet

Opinion Mining: Taking into account the criteria!

Pascal Poncelet
LIRMM - UMR 5506
Campus St Priest - Bâtiment 5
860 rue de Saint Priest
34095 Montpellier Cedex 5 - France
and
UMR TETIS Montpellier - France
Pascal.Poncelet@lirmm.fr

Abstract

Today we are more and more provided with information expressing opinions about different topics. In the same way, the number of Web sites giving a global score, usually by counting the number of stars for instance, is also growing extensively and this kind of tools can be very useful for users interested by having a general idea. Nevertheless, sometimes the expressed score (e.g. the number of stars) does not really reflect what it is expressed in the text of a review. Actually, extracting opinions from texts is a problem that have been extensively addressed in the last decade and very efficient approaches are now proposed to extract the polarity of a text. In this presentation we focus on a topic related with opinion but rather than considering the full text we are interested with the opinions expressed on specific criteria. First we show how criteria can be automatically learnt. Second we illustrate how opinions are extracted. By considering criteria we illustrate that it is possible to propose new recommender systems but also to evaluate how opinions expressed on the criteria evolve over time.

1 Introduction

Extracting opinions that are expressed in a text is a topic that have been addressed extensively in the last decade (e.g. (Pang and Lee, 2008)). Usually proposed approaches mainly focus on the polarity of a text: *this text is positive, negative or even neutral*. Figure 1 shows an example of a review on a restaurant.

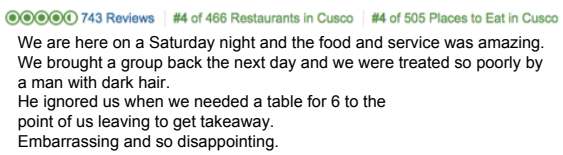

Figure 1: An example of a review

Actually this review has been scored quite well: 4 stars over 5. Any opinion mining tools will show that the review is much more negative than positive. Let us go deeper on this exemple. Even if the review is negative it clearly illustrates that the reviewer was mainly disappointed by the service: he was in the Restaurant and found it amazing. We could imagine that, at that time, the service was not so bad. This exemple illustrates the problem we address in the presentation: we do not focus on a whole text rather we would like to extract opinions related to some specific criteria. Basically, by considering a set of user-specified criteria we would like to highlight (and obviously extract opinions) only on the relevant parts of the reviews focusing on these criteria. The paper is organized as follows. In Section 2 we give some ideas on how to automatically learn terms related to a criterium. We give also some clues for extracting opinions to the criteria in Section 3. Finally Section 4 concludes the paper.

2 Automatic extraction of terms related to a criterium

First of all we assume that the end user is interested in a specific domain and some criteria. Let us imagine that the domain is movie and the two criteria are actor and scenario. For each criterium we only need to have several keywords or terms of the criterium (seed of terms). For instance in the movie domain: **Actor=** {actor, acting, casting, character, interpretation, role, star} and **Scenario=**{scenario, adaptation, narrative, original, scriptwriter, story, synopsis}. Intuitively two different sets may exist. The first one corresponding to all the terms that may be used for a criterium. Such a set is called a *class*. The second

one corresponds to all the terms which are used in the domain but which are not in the class. This set is called *anti-class*. For instance the term *theater* is about movie but is not specific neither to the class actor nor scenario. Now the problem is to automatically learn the set of all terms for a class. Using experts or users to annotate documents is too expensive and error-prone. By the way there are many documents available on the internet having the terms of the criteria that can be learned. In a practical way by using a research engine it is easy and possible to get these documents. For instance, the following query expressed in Google: "+movie +actor -scenario adaptation narrative original -scriptwriter story -synopsis" will extract a set of documents of the domain movie (character $+$), having actor in the document and without (character $-$) scenario, adaptation, etc. In other terms we are able to automatically extract movie documents having terms only relative to the class actor. By performing some text preprocessing and taking into account a frequency of a term in a specific window of terms (see (Duthil et al., 2011) for a full description of the process as well as the measure that can be used to score the terms) we can extract quite relevant terms: the higher the score, the higher the probability of this term belonging to a class. Nevertheless as the number of documents to be analyzed is limited, some terms may not appear in the corpus. Usually these terms will have more or less the same score both in the class and the anti-class. They are called *candidates* and as we do not know the most appropriate class, a new query on the Web will extract new documents. Here again, a new score can be computed and all the terms with their associated scores can finally be stored in lexicons. Such lexicon can then be used to automatically segment a document for instance.

3 Extracting opinions

A quite similar process may be adapted for extracted terms used to express opinions: adjectives, verbs and even grammatical patterns such as $<$adverb $+$ adjective $>$ in order to automatically learn positive and negative expressions. Then by using the new opinion lexicon extracted we can easily detect the polarity of a document. In the same way by using the segmentation performed in the previous step it is now possible to focus on criteria and then extract the opinion for a specific criterium. Interested reader may refer to (Duthil et al., 2012).

4 Conclusion

In the presentation we will present more in detail the main approach. Conducted experiments that will be presented during the talk will show that such an approach is very efficient when considering Precision and Recall measures. Furthermore some practical aspects will be addressed: how many documents? how many seed terms? the quality of the results for different domains? We will also show that such lexicons could also be very useful for recommending systems. For instance we are able to focus on the criteria that are addressed by newspapers and then recommend the end user only with a list of newspapers he/she could be interested in. In the same way, evaluating how opinions evolve over time on different criteria is of great interest for many different applications. Interested reader may refer to (Duthil, 2012) for different applications that can be defined.

Acknowledgments

The presented work has been done mainly during the Ph.D of Dr. Benjamin Duthil and in collaboration with Gérard Dray, Jacky Montmain, Michel Plantié from the Ecole des Mines d'Alès (France) and Mathieu Roche from the University of Montpellier (France).

References

B. Duthil, F. Trousset, M. Roche, G. Dray, M. Plantié, J. Montmain, and Pascal Poncelet. 2011. Towards an automatic characterization of criteria. In *Proceedings of the 22nd International Conference on Database and Expert Systems Applications (DEXA 2011)*, pages 457–465, Toulouse, France. Springer Verlag.

B. Duthil, F. Trousset, G. Dray, J. Montmain, and Pascal Poncelet. 2012. Opinion extraction applied to criteria. In *Proceedings of the 23rd International Conference on Database and Expert Systems Applications (DEXA 2012)*, pages 489–496, Vienna, Austria. Springer Verlag.

B. Duthil. 2012. *Détection de critères et d'opinion sur le Web (In French)*. Ph.D. thesis, Université Montpellier 2, France.

B. Pang and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trend in Information Retrieval*, 2(1-2):1–135.