RRSS - Rating Reviews Support System purpose built for movies recommendation

Grzegorz Dziczkowski[1,2] and Katarzyna Wegrzyn-Wolska[1]

[1] Ecole Superieur d'Ingenieurs en Informatique et Genie des Telecommunicatiom (ESIGETEL) 1,Rue de Port de Valvins 77-215 Avon-Fontainebleau Cedex
[2] Ecole des Mines de Paris 35, rue Saint-Honore 77305 FONTAINEBLEAU
grzegorz.dziczkowski, katarzyna.wolska@esigetel.fr

Summary. This paper describes the part of a recommendation system designed for the recognition of film reviews (RRSS). Such a system allows the automatic collection, evaluation and rating of reviews and opinions of the movies. First the system searches and retrieves texts supposed to be movie reviews from the Internet. Subsequently the system carries out an evaluation and rating of the movie reviews. Finally, the system automatically associates a digital assessment with each review. The goal of the system is to give the score of reviews associated with the user who wrote them. All of this data is the input to the cognitive engine. Data from our base allows the making of correspondences, which are required for cognitive algorithms to improve, advanced recommending functionalities for e-business and e-purchase websites. In this paper we will describe the different methods on automatically identifying opinions using natural language knowledge and techniques of classification.

1 Introduction and issue

With the growth of the Web, e-commerce has become very popular. A lot of websites offer online-sales. To increase their sales, online shops include the special recommended systems (RS) to suggest products to the clients. While people like to check out the recommendations of other users before creating their own opinion, those predictions become very useful for the customers. RS allow customers to make the choice without any personal knowledge of alternatives. Algorithms for suggestion are based on the experience and the opinion of other users. It is helpful to find recommendations from people who are familiar with the same problem, who have made their choice in the past, whose perspective is valued, or who are recognized experts [12]. RS also provide correspondences between users, who have a similar profile. A new user has to create his own profile. The RS will suggest a new precise choice based on the similar taste of other users. The efficacity of such system depends on data quality and quantity. This is why RS need huge databases of user profiles: the more profiles it gets, the beter the algorithms are. RS proposes the choice to the user, which is based on correspondences between the users' opinions. Our system (RRSS) furnishes the users' profiles, which are necessary for algorithms of cognitive engines. This result cannot depend on commercial reasons, because it could make people distrustful. RCSS consist of two principal modules:

- extraction and filtering opinions from the text, which consists in the identification of quite precise information in natural language and its representation in a structured form [8].
- assigning a mark only to subjective sentences, which express or describe opinions, evaluations, or emotions [10][15].

The relative failure of the generic systems is well-known today. Many researchers try to describe natural languages in the same way as formal languages. Maurice Gross undertook with his team (LADL; French Laboratory for Linguistics and Information Retrieval) the exhaustive examination of simple sentences of French [5], in order to have reliable and quantified data predicted to rigorous scientific treatments. To exploit the linguistic knowledge, the LADL developed a special application named Unitex [9]. This is an enhancement environment used to build formalized descriptions for broad coverage of natural languages and apply them to substantial texts. Unitex processes the texts of several mega-bytes to morpho-syntactic indexation in real time, to search for set phrases or semi-fixed phrases, to produce agreements and statistical evaluation of the results. The linguistic resources used to achieve the information retrieval and extraction are as follows: dictionaries, networks of recursive transitions (local grammar) and lexicon-grammar tables.

Another way to analyse an opinion automatically from the text is to use statistical classifiers. All of the analyzed objects are assigned to the previously prepared classes. Statistical methods suppose that descriptions of the same class respect a specific structure of the class. For classification of huge corpora we often use the special learning methods based on tested instances (examples). Problems consist in constituting a representative corpus of the evaluated field, and finding the rules or constituting an operational model of this corpus. The model created allows the system to predict the behaviour for new candidates. At present, classification of opinions as subjective/objective or positive/negative is a very interesting challenge for research: Turney, Littman [13], Dave, Lawrance [3], Pang, Lee [7]. Classifiers assign the new objects for analysis to correspondend to previously prepared classes. The classifiers performance depends on the model for each base learning class.

2 Marking a review

RRSS has modular architecture. The principle tasks are: collecting the reviews from Internet; checking if the text found is a review; assigning a mark to the review and presentation of the results. This paper focuses only on the review marking module. Generally the mark assignment process distinguishes the linguistic and probabilistic parts. In our approach the linguistic part is responsible for pre-processing of the text, creating a learning base and finding behaviour of identical mark groups [paragraph 2.1]. The probabilistic part classifies reviews to the mark [paragraph 2.3]. Our algorithm follows the next steps:

- learning base creation,
- vector representation,
- classification.

The process of mark assignment to the review consist of two main steps: first estimation of a mark based on the behaviour of the same review mark groups and the final assignment of the mark [fig 1]:

- gathering the reviews according to their mark,
- finding the behaviour of each group of mark,
- for a new review the first estimation of the mark directly from the characteristic of the group behaviour ,
- creation of a learning base for Bayes classifiers,
- assignment of a final mark to the review.

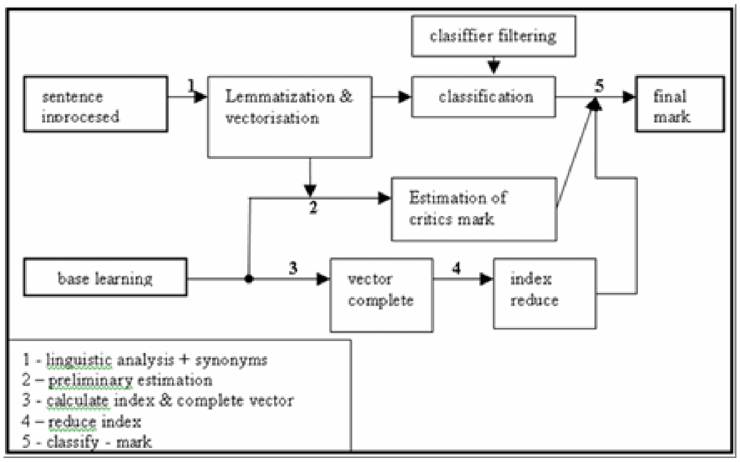
clasiffier filtering
sentence inprocesed
1
Lemmatization & vectorisation
classification
5
final mark
2
Estimation of critics mark
base learning
3
vector complete
4
index reduce
1 - linguistic analysis + synonyms
2 – preliminary estimation
3 - calculate index & complete vector
4 – reduce index
5 - classify - mark

Fig. 1. Mark assignment process

2.1 Learning base

To perform the review assessment we need a group of characteristics already evaluated - a learning base. Different websites publish film reviews with its mark assigned (e.g. IMDB, Amazon). We used this data (reviews, users, marks) to create our learning base. We use the scale of marking from 1 to 5. We regrouped all the reviews according to their mark. This way, we obtained 5 different groups of film reviews: a group according to reviews with score 1, 2 5. Then, we tried to determine the characteristics for each group. We supposed that delimited parameters characterize behaviour of a group. These characteristics are for example: a typical word, typical expression, a size of a sentence, the frequency of characteristic word repetition, the number of punctuation marks (!, ;), ?) and so on. For group categorizing, we used a linguistic analyser Unitex, to lemmatize the words, to assign semantic classes to the words, to add synonyms [4] and to detect negation. For this task we used a linguistic processing, which requires lexicons and specialized grammar.

The development of such resources is a long and tiresome task, which generally requires an expertise in the field and knowledge in data-processing linguistics; techniques of filtering, categorization of documents and extraction of information. The linguistic processing needs a good text comprehension. It means transduction, which transforms a linear structure into a conceptual structure, i.e. text (the linear structure) is transformed into an intermediate logico-conceptual representation, which is then used to make conclusions. The semantic analysis aims at producing a structure representing, as accurately as possible, a unit of the sentence, with its meanings and its complexity [1][11][6]. Semantico-conceptual structures can be more or less broad, rich and complex and more or less ambiguous [4].

To determine the behaviour of a group we parse the large corpus of reviews, which were assigned with the same mark to find the characteristic. Our linguistic resources are the dictionaries and local grammars. The electronic dictionaries describe the simple words and the complex words of a language associating them with a lemma made up of a series of grammatical, semantical and inflexional codes. Grammars are representations of linguistic phenomena by recursive transitions (RTN). Generally a grammar represents sequences of words and produces linguistic information such as for example information on the syntactic structure. The local grammars, represented in the forms of graphs, describe elements which concern the same syntactic or semantic field. On fig 2 we show an example of local grammars used to determine the behaviour of groups. We assigned the semantic classes to our word corpus. To do this we used subjective word dictionary - General Inquirer Dictionary [3]. Then we parsed the corpus using local grammars to obtain statistical results.

Finally, we obtained a series of characteristics, which precisely determine a group. The characteristics are different for all of the study groups and generally they describe the statistical scores of typical words, their synonyms and they take into account negations. The results shoved strong differences between the characteristics of those groups. The creation of the group behaviour allows the determining of to which group a new review belongs. We

[3] http://www.wjh.harvard.edu/ĩnquirer/spreadsheet_guide.htm

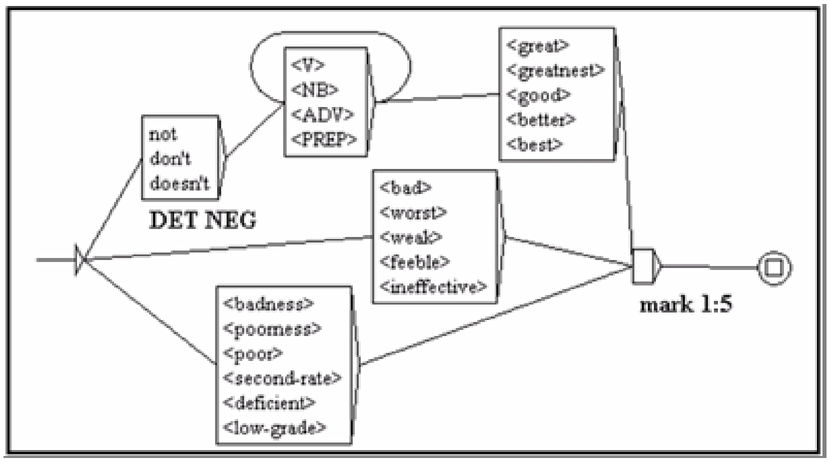

Fig. 2. Example of local grammar

used characteristics of groups for a preliminary estimation of the mark [fig 1; action 2]. This estimation helps us in the selection of classifier, which will process reviews.

2.2 Vector representation

The vector representation of a corpus requires an initial linguistic preprocessing to eliminate all of the ”empty” words not taking an active part in the meaning of the document. A first step is to build the index of the text learning base. Then, each text is represented by its coordinates in this index. Introducing the classes to initial index filters reduces the dimension of the vector representation used by the classifier. A linguistic filter is applied in order to eliminate some types of words considered useless for the categorization. All language subtleties contained in the text in order to analyze are necessary to guarantee a good performance of categorization. We added the synonyms by using the semantic classes. Then, we built a vector representation of all the text in the learning base corpus. The dimensions of the vector correspond to the complete index. The vector components are frequencies of the index terms in the document [fig 1; action 3].

Finally, the dimension of learning base space vector is enough to proceed the classification. Very often, the vector selected from the classifier includes many components with the value equal to zero. Those values do not have any incidence on the classification process. Thus, it is possible to reduce the size of the index to improve the performance of the classifier [fig 1; action 4]. Several methods were proposed to carry out a selection of the words representative of the field [14]. We chose the method of mutual information measurement proposed by Cover [2], which is especially well adapted for application in natural language processing.

Definition 1. *For the group of documents under consideration, the average mutual information* ***I*** *is the difference between the entropy of variable* ***C*** *and its conditional entropy relative with the word* $\boldsymbol{m_t}$*.*

$$I(C, M_t) = \sum_{c \in C} \sum_{m_t \in M_t} P(c|m_t) \times \log \left(\frac{P(c, m_t)}{P(c) \times P(c|m_t)} \right) \tag{1}$$

where : P(c) is the number of documents of the class C divided by the total number of documents

P(m_t)is the number of documents containing word m_t divided by the total number of documents

P(c, m_t) is the number of documents of the class C and containing word m_t divided by the total number of documents

C is the random variable associated with all the classes (c),

M_t is the random variable, representing existence of the word m_t in a document

This technique allows the calculation a reduced index dimension used by the classifiers. This method largely decreases the size of the index of a classifier. We select the words of which the mutual information is higher then a given threshold (sI). The reduced index of each classifier define a new vector space dedicated to the classifier.

2.3 Bayes classifier

The way of carrying out the classification is to find characteristics for each class and to associate a function of belonging. Among the methods using this process we can quote the decision trees, the Bayes classifiers, the method of SVM, etc. For our first approach we have used the Bayes classifier, which is a categorizer of the probabilistic type founded on the theorem of Bayes [13]. In our approach, we have presented five different classifiers, each classifier corresponds to a group of marks. The description of review behaviours, which belongs to different groups of marks, were done manually. The opinions are analysed sentence by sentence. Each classifier gives a mark (from 1 for 5) to the sentence. The classifier privileges the same mark, which was received in preliminary estimation process. For example, a classifier that corresponds to mark 1 will privilege assigning a mark 1[fig 1; action 5]. At the end of our process, we obtain the mark for all the sentences of the reviews processed. A final mark assignment of the reviews is the value of the arithmetic mean of all sentences treated. Our algorithm of rating the opinion is composed of two steps: first, the initial estimation of mark by the behaviour classification of the groups and finally the assignment of a mark by using the appropriated classifier allocated by an initial mark. By using this architecture we hope to improve the F-scores of systems, which directly use classifiers.

3 Conclusions

The objective of our work is to build a system for collecting, evaluating and ranking movies reviews. RRSS the Rating Reviews Support System is the proposal for the system, which carries out a collection and marking of reviews. This paper presents only the evaluating and ranking part of the system. RRSS will be a support to RS. The goal of our work is to automate the whole system particularly to improve the estimation of individual user's reviews. The system allows an automatic assignment of a mark; however to increase the research on other fields it will be necessary to create a linguistic base and a new analyzis of the different elements of the group's behaviour.

References

1. Altai, H., The core language Engine. MIT Press (ACL-MIT Press Series in Natural language Processing), Cambridge, 1992.
2. COVER Cover, Thomas. Elements of Information Theory. John Wiley 1991.
3. Dave, K., Lawrance, S., Pennock, D.M.2003. Mining the Peanut Gallery: Opinion Extraction with HMM Structures Learned by Stochastic Optimization. In AAAI-2000.
4. Dziczkowski, G., Wegrzyn-Wolska, K., Graph based system purpose - built for automatic retrieval and extraction of the electronics data. To appear in proceeding of Euro-IMSA, Mars 2007.
5. Gross, M., The construction of local grammars. In Finite-State Language Processing, Cambridge, MIT Press, pp 329-354, 1997.
6. Kamp, H., Evenements representations discursives et reference temporelle. Langages, nb 64, 1981, pp.34-64.
7. Pang, B., Lee, L., 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In ACL-04.
8. Panzienza, M.T., Information extraction (a multidisciplinary approach to an emerging information technology). Springer Verlag (Lecture Notes in Computer Science), Heidelberg, 1997.
9. Paumier S., De La reconnaissance de formes linquistique a l'analyse syntaxique. These, Marne-la-Valee, 2003.
10. Riloff, E., Wiebe, J., Philips, W., Exploiting Subjectivity Classification to Improve Information Extraction.
11. Sowa, J., Conceptual Structures. Information processing in Mind and Machine. Addi son Wesley Publishing CO., Reading, 1984.
12. Tarveen, L., Hill, W. (2001). Beyond recommender systems: helping people help each other. In HCI in the millennium , J. Caroll, ed., Addison-Wesley, pp 1-21, 2001.
13. Turney, P., Littman, M., 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. ACM Transaction on Information Systems (TOIS) 21(4):315-346.
14. Wang, Y., Hodges, J., Tang, B. Classification of Web Documents using a Naive Bayes Method. IEEE, pp 560-564. 2003.
15. Wiebe, J., Wilson, T., Bruce, R., Bell, M., Martin, M., 2004. Learning Subjective Language. Computational Linguistics 30(3):277-308.