Research Project – Part 2

The architecture and the implementation of the opinion mining and sentiment analysis application.

Abstract

The rapid expansion of Internet brought with it the rise of forums, blogs, wikis and other social networks. This new media has proved to be a perfect place for more and more people to express their ideas, opinions and beliefs about diversified subjects from movies and small home appliances to religious beliefs and educational system. The reason for this is the freedom, openness and ease of use that internet provides in opposing to the traditional forms of media like television, radio and newspapers. Also the Internet can protect someone’s true identity, people finding it easier to express themselves better when anonymous, without fear of judgment. Now the Internet has become an enormous source of information, regarding the collective opinion of the public and their subjective thoughts that cannot be ignored anymore.

So more and more turn to Natural Language Processing, Opinion Mining and Sentiment Analysis to gather this information through automatic systems and exploit it for getting a lead in today economical society.

Nowadays, Facebook and Twitter, considered some of the biggest companies in the social web industry, count millions of users, billions of GB of shared data and huge amounts of money. Besides this kind of companies, oriented only in developing a large community, where people can share data, other kind of businesses have been influenced a lot about the rising of the social web. As example, the online stores have developed user opinions and reviews systems for all their products, whose role was to help a person/visitor choosing the right product he needs. The reviews resides both on positive and negative opinions, ratings, recommendations, influencing the selling forecasts and results for a specific product or range of products. There were built too specific online places for writing and reading opinions/reviews, blogs, forums, chats, wikis etc.

Due to the large importance of the two domains, there were published a lot of official papers containing approaches for mining opinions and extracting sentiments. This paper presents the architecture and the implementation of such a system, for mining opinions in an online site specialized in mobile phones reviews and opinions.

Introduction

This paper presents an approach of an opinion mining and sentiment analysis system for the mobile phones industry. It is divided in seven chapters, presenting some general ideas on terms of these two domains, some researches of other persons, the architecture and the implementation of the system and some results.

The first chapter is represented by the introduction to the basic ideas about opinion mining and sentiment analysis. The second and the third ones present few details about what opinion mining and sentiment analysis are, where there are used and which are the main goals of mining and extracting sentiments and opinions.

The forth one presents other research papers with their approaches on mining opinions and extracting sentiments. The fifth one presents the architecture of the system and its modules with their description, some details about the interconnection between modules and about the data flow.

In the sixth chapter there are presented some implementation details, algorithms, data structures.

The final chapter contains the results of the system and few conclusions.

As specified in the first part of our research project, the opinion mining and sentiment analysis concepts are widespread all over the globe, in the hole area of online media, and were debated in a lot of official papers and research reports. The large interest for this two domains show us their importance.

In online media, almost all online shops have implemented in their site reviews and opinion systems, which permit to all users to write some personal, subjective thoughts regarding the products and their features.

As a supplier or producer, everywhere in the world, one of the most important factor in selling a product is the feedback from the client about the product and about the services you had offered to. Because the selling process in many industries was oriented through the online industry, the companies became more and more interested in analyzing the public reviews from their website. Doing this in a manual manner is time consuming and requires more available resources, so the companies have redirected this job to an automated system. So, the need for opinion mining had appeared. Not only the commercial places had appealed to automated opinion mining systems, but also the social communities for determining the types of the relations between users, for analyzing some of their topics and other social stuff.

Opinion Mining

In computer science, the opinion mining is not a simple domain, considering that a machine should determine the attitude of a phrase, the emotions and the sentiments of the writer. It follows the natural language processing and text mining, more than that it resides on natural language processing, text mining and computational linguistics. The opinion mining applications have evolved a lot in the last few years, and more industries have started to use them.

Due to the large range of topics discussed on forums, blogs, social networks, news sites, from places for holiday to the ultimate technology smartphones, the shared data increase a lot. What the opinion mining is and what is it for represents a widespread question among a lot of users.

“The ability to leverage the vast amount of user-generated content on these and similar sites for commercial and social benefit is supported via the development of an emerging technology called opinion mining.”

The development of the opinion mining technology brings three important advantages over traditional polling and focus groups. First, the opinion mining systems are consistent over time on terms of data. The second advantage is that these systems can run almost in real time, updating needed data frequently, and the last one is represented by the multiple languages which an opinion mining system could have, which would transform a such system in a global instrument.

First opinion mining systems were relatively simple, calculating only a binary polarity of a statement, positive or negative. Later, the research departments of some computer science faculties had developed n-ary systems, which classify a statement based on its polarity into a range of values, e.g. [-1; 1], value of -1 means that the opinion is strongly negative and 1 is strongly positive. 0 means neutral opinion. Advanced systems take into calculation more factors than polarity of a text and the most common are:

- subjectivity/objectivity identification,

- feature/aspect-based sentiment analysis, and

- polarity of a given text

Sentiment analysis

Sentiment Analysis refers to an area of interest in Natural Language Processing. It complements the field of Opinion Mining (some saying the line between them is so thin, practically non-existent) and refers to determining the attitude of the writer or reader towards some text or topic. By attitude we refer to the evaluation and judgment (like or dislike), the emotional state of the author or the sentiment communicated to the reader.

Some of the early research and applications in the field of Sentimental Analysis were oriented towards analyzing reviews and comments on different products, movies or restaurants and detecting the emotional polarity: positive, negative or neutral. This classification can be made at the document, sentence level or in-depth regarding a certain feature of the topic such as the color contrast and resolution of a TV.

Other researches in opinion mining and sentiment analysis industry

Professor V.S. Subrahmanian from the Institute for Advanced Computer Studies at the University of Maryland, presents [1] in an article about mining online opinions, which areas use this kind of applications and which their results are. Now that highly accurate opinion information is becoming available commercially, the demand for such services by corporations, governments, nonprofits, and individuals is increasing dramatically.

In his study [8] Jeonghee Yi, et al., describes an application on sentiment classification based on reviews and opinions that people express about a topic. When designing the sentiment miner, the author had the following challenge in mind, not only is the overall opinion about a topic, but also the sentiment about individual aspects of the topic essential information of interest. The reason for this is that the document level classification fails to detect sentiment about individual aspects of the topic.

University of Maryland released their application for opinion mining, named Oasys, which has been extensively tested against humans by both its authors and third parties. They demonstrated that the results of any opinion-mining system, using any algorithm, cannot agree every person or even a small group of 10 persons. Their goal was to obtain an average score given by the people in a system test.

Few persons of academic staff of Curtin University of Technology, Australia, wrote an article based on some researches and studies about opinion mining industry [2]. Their report is based on closely related fields of opinion mining, on analyzing an opinion mining architecture and on presenting the application domains.

Vivek Sehgal, et. al., at the University of Maryland came up with a new method [9] to predict stock market using sentiment analysis on opinions and commentaries from financial web pages.

Qingliang Miao, Qiudan Li and Ruwei Dai, from the Laboratory of Complex Systems and Intelligence Science, Institute of Automation Chinese Academy of Science, present in an official report an unified framework of opinion retrieval [3]. They have developed an ORS (Opinion Retrieval System) to retrieve opinions in customer review domain. Their system is based on the ideas below:

- opinions have temporal dimensions, the opinions about an object may change over time;

- reading one opinion is not enough to extract a correct result;

- opinions are varying in quality, authority, reliability and popularity.

Keke Cai, et. al., [10] form IBM China Research Lab, Beijing, China provided us with an example for a sentiment miner oriented towards extracting the reason hidden behind each sentiment, the background onto which the sentiments are projecting. Continue to follow is the system they proposed.

Freimut Bodendorf and Carolin Kaiser, from the Department of Information Systems of University of Erlangen-Nuremberg, have studied the opinion mining domain too and especially, the mining customer opinions on the web, in the automotive industry. Their approach belongs to the feature-based opinion mining and it is based on extracting and analyzing customer opinions on products in online forums.

Most sentiment analysis has looked on reviews of a single product, in isolation. But many decisions are made after a comparison between several products and choosing the right one. Ronen Feldman, et. al., [12] have provided a method to automatically analyze such form of reviews and determine which products are compared, the characteristics they are compared on and which does the user prefer on each dimension. The information extracted can be presented in product comparison tables and graphs, which can be used by both companies and customers.

Mining a global social network, on the other hand, has the potential to uncover content and interactions aimed to the radicalization of those with no prior interest terrorism. That was the idea of Adam Birmingham, et. al., from Doublin City University [13], who proposed a system that crawled YouTube database and examined the topics discussed and the sentiment polarity of the comments. YouTube was chosen because it provides text- and video-based information and interactivity through users commenting on videos, writing on each other’s profiles, selecting friend, joining groups.

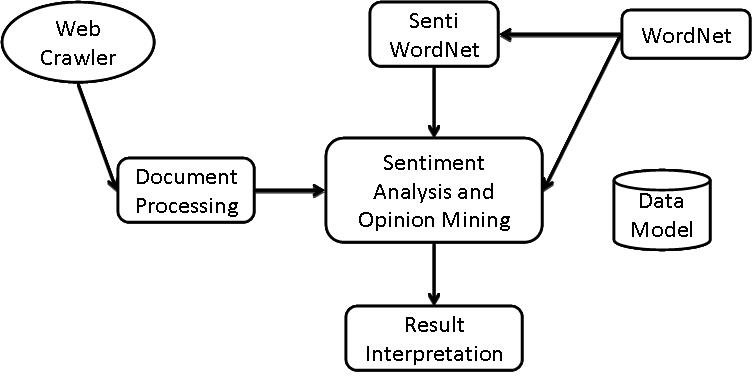
Design of the system

The architecture contains six interconnected modules for data acquisition, data processing, natural language processing, opinion mining, sentiment analysis and results interpretation.

This type of architecture based on components is both robust and flexible when it comes to changes. New components can be added or existing ones can be replaced without affecting the integrity of the architecture. Also new workflows for other sentiment and opinion analysis techniques and algorithms can be supported by adding new features to the current components. Any changes made will not impact other parts of the system besides the components where they originated from.

Every component communicates with the data model that facilitates communication and information sharing.

Fig. 1 The architecture of the system



First of all, a **web crawler** parses the specified site for building the xml source files, containing all the opinions and user reviews of a lot of mobile phone versions and models. The crawler reads the DOM tree of every webpage and extracts the name of the phone, its features, the name of the author who posted the review, the date and the text of the review. All this information help in building raw data files with xml structure that containing tags for author, date and opinion text inherited from the “opinion” parent tag. The root tag of the xml file is for opinions, containing all the “opinion” tags. Each file contains opinion data for only one device (e.g. IPhone, etc.) making thi

So, the input data for the web crawler module is the web page of the reviews site and the output is represented by the xml files.

The **“Document Processing” module** receive as input data the xml files, parses them and builds some data structures for every opinion and review. It can also extract snippets or match certain patterns in the opinion text. It’s main role is to prepare data for the sentiment analyzer. So, the output of this module is an array list of opinion-type objects.

After parsing all files and generating data structures for reviews and opinions, the **sentiment analysis and opinion mining module** gets its job. This module contains some natural language processing techniques and algorithms for extracting opinions and sentiments. The approach is based on computing the polarity for every opinion and feature using SentiWordNet and WordNet and on detecting the negation of different terms and features. The polarity of the opinion is calculated by adding the polarity of every word in that text. The polarity of a word is retrieved from SentiWordNet and if this is not found, we turn over to WordNet and compute the arithmetic average of all its synonyms and refer to it as ne word’s new polarity score. After computing the polarity, the application calls the negation detection algorithm for further processing.

The sentiments towards features are computed much in the same way. The text is divided in sentences and every word in a sentence is checked to see if it is a synonym for one of the features. If so the polarity for the entire sentence is calculated and added to the total polarity of that feature. To find out the synonym set of a word we relay once more on WordNet.

The input data for this model are the Opinion and Feature data structures, maybe other types of snippets and patterns, and the output is completing the polarity attributes for hose structures. More details will be presented during presenting the implementation module’s specifications.

The final module, **results interpretation**, presents to end-user the final results, some charts containing opinions and sentiments of the overall product and of every one of its feature, specified by the user. The input data for this module is represented by the final score of each opinion and by the final score of each snippet containing a feature and the output of this module consists in the interface with the user (arrays of data, charts, trends, etc). The output consists of revealing the polarity score of every opinion, the total polarity score regarding the topic discussed, the score for each feature plus the frequency and number of documents where feature related words are found.

The design of the application is modular, so it can be easily extended, by adding other modules. The data flow starts in the web crawler module and ends in the results interpretation module, after it suffers some processing operations.

Implementation module

This module contains all the classes defined for this project, which form the architecture’s modules. For the Document Parsing Module it is defined the InputDocumentParser class, which receives as parameter of the constructor the file path of the xml file built by the web crawler. First of all, the DOM representation of the xml file is obtained and after that all the opinions are stored in a data structure of ArrayList containing Opinion type objects. Every Opinion object waits for three parameters in the constructor: the date, the author and the opinion text.

For implementing the opinion mining and sentiment analysis module, there are defined the following classes: WordClass, SentiWN, SentimentAnalizer and Feature.

The WordClass class implements all the methods to compute the positive coefficient, the negative and the objective ones and the final score for a word, considering all synsets where it belongs to. The methods getPositive(), getNegative() and getObjective() return an arithmetic average of the polarity scores corresponding to all the meanings (definitions) of that word. The method addSynset() computes the total values of the positive, negative and objective scores, by adding for each meaning in each set synset its corresponding value and counting the number of synsets this wor belongs to.

The same word can have multiple meanings, depending on the context in which it is used, and each meaning belongs to a certain synset of synonyms that are characterized by specific polarity values (e.g. the word “castle” can have the following meanings “castle#n#1” – a noun signifying large building formerly occupied by a ruler and fortified against attack; “castle#n#1” – a noun signifying the piece chess piece “rook” that can move any number of unoccupied squares in a direction parallel to the sides of the chessboard; “castle#v#1 – a verb signifying interchanging the positions of the king and a rook in chess). The letter between the two “#” symbols determines the part of speech that meaning refers to: “n” is for nouns, “v” is for verbs, “a” is for adjectives and “r” is for adverbs. Also the number represents the frequency that meaning is found in free speech, the lower the number the frequent that meaning is found in texts and documents.

In the SentiWN class the SentiWordNet (SentiWordNet\_1.0.1.txt) file is parsed and a data structure (hash map) is built with all the words of the SentiWordNet (the word is the key and it has attached an WordClass object). The SentiWOrdNet file used in this sentiment analyzing system is version 1.1, unfortunately we were not able to procure the latest version 3.0. This text file contains on each row the POS (part of speech) value for a synset, the positive and negative polarity scores and the words that compose the synset (e.g. a 1010301 0.0 0.0 sweet-smelling#a#1 perfumed#a#2 sweet#a#7 scented#a#3 odorous#a#3 sweet-scented#a#1 odoriferous#a#4 ).

The method getWord() retrieves the WordClass object considering the word and its positive and negative values, associated to the word received as a parameter. The methods getPositive(), getNegative(), getObjective() and getScore() returns the polarity scores for the specified word.

When parsing the opinion the words are looked up in SentiWordNet to retrieve their polarity score and if a certain word is not found in the SentiWordNet, the polarity of this word is computed considering its synonyms. This is done in the findNewWord() method. For every synonym of the word we add their positive and negative polarities and divide it by the number of synonyms. This process is not recursive, implying that when determining the polarity of one word and reach a synonym of its that isn’t found in SensiWordNet either, the system does not compute the polarity of the new word and just skips it over. Through practice was shown that even a recursion of depth one can involve large running times. The new word and it’s polarity are added to the SentiWordNet hashmap for future use and reference. SentiWN class has a flag attribute addNewWord to specify if the new word is looked for and added.

The synonyms are obtained using WordNet implementation in Java, RitaWN framework that is a wrapper over Jawbone/JWNL functionality.

SentmentAnalizer class is the core of the system, where the actual sentiment and opinion analysis algorithms and techniques are implemented. The method analizeSentancePolarity() computes the polarity score of a sentence by adding the polarity scores of all its words. These polarities are searched in SensiWordNet as exemplified earlier. Before adding the polarity score of any word a check is made to eliminate “stop words”, using isStopWord() method. By “stop words” we can understand every word that is not relevant to the sentiment analysis or which can badly affect the efficiency and accuracy of the analyzer. In this implementation we only reject negation words like “not”, “no”, “don’t”, but it can easily remove other words like prepositions and word categories which we know don’t have any polarity score, in order to eliminate the look-up time or can be configured to accept only adjectives for a more specific and focused analysis. The reason negation words are eliminated is because the system looks for negation patterns in the sentence. If a word is found to be negated its polarity is negated (1) and then added to the overall polarity score.

*(1)* ,where

p – the initial polarity

– the negated polarity

(Negation algorithm)

analizeSentenceFeatures() method is used to compute the polarity score for features. This method receives a sentence and searches if any word belongs to the synonym set of any feature. If a match was found then that feature is a discussion subject for the sentence and its polarity should be added to the polarity score of the feature. The sentiment score for sentence is computed using the same methods stated earlier.

The methods analizeOpinion() and analizeOpinions() are actually the ones called from outside the sentimentAnalizer object to perdorm sentiment/opinion analysis.

All the methods mentioned above read and modify directly the data structes form the data model.

Result Interpretation module is represented by DataReport class. This class accesses directly the data model to retrieve sentimental information and present it in different ways. For the overall sentiment and opinion analysis the author, date along with the first 6 words and total polarity score of every opinion extracted by the web crawler are printed on the screen. The overall polarity score, for the entire document, is calculated and printed. This functionality is due to printReportOpinions() method. Also for each feature the total polarity score, frequency of appearance and number the of opinions in which it is referenced are printed by calling the method printReportFeatures().

The data structure includes Feature and Opinion classes to model the entities for the problem’s universe. For Feature we have the name, polarity, total frequency and opinion frequency attributes. It also holds an ArrayList with references to all opinions where the feature was discussed and an ArrayList with all its the synonyms. The Opinion class has author, date, text and polarity score attributes for an opinion extracted from the website.

Results

Future work