Algorithm for efficient Opinion Summarization and categorization (AEOSC): A web Based approach

Ms. Anjali Dadhich[1], Dr. Blessy Thankachan[2]
Research scholar, Assistant Professor
School of Computer & Systems Science
Jaipur National University.

Abstract:
On recent years, special attention has been giving also to the amount of produced user-generated content. The e-commerce sector is one of the most affected by the amount of data produced by customers, which increased dramatically during the phase known as Web 2.0. Customer's opinions represent a valuable unique type of information which should not be mistreated or ignored by the research community. Thus, this work emphasizes the need of special mechanisms that aims to provide the community better ways to take full advantage from this data. From the customer perspective, considering others opinions before purchasing a product is a common behavior long before the existence of Internet. In the era of the digital world, the difference is that a customer has access to thousands of opinions, which greatly improves decision making.
The proposed framework will combine several techniques to extract valuable information out of natural language text (user-generated content), in order to provide enrichment of the experience of users by taking advantage of the available content in a more intelligent and organized way.
Keywords: Opinion Mining, NLP, Machine learning, Social media

1. INTRODUCTION:

In the era of the digital world, the difference is that a customer has access to thousands of opinions, which greatly improves decision making. as a consequence of this new order, providing a public space for community discussions became almost a pattern for new web applications. To meet the new demand, most existing web sites like social networks, web magazines, newspapers and e-commerce had to change their systems to comply with new design standards (introduced by Web 2.0), which among other things provides a common space allowing interaction of users through the exchange of opinions and experiences. Since then, a huge amount of data has been produced by users, a valuable content that can be extremely useful both as a complementary, but also in many cases as a primary and unique source of information. Because of the nature of this content, which is unstructured data in form of free text, the recovery and extraction of meaningful information depends on specialized techniques. This work explores the use of such techniques focused on the analysis of user-generated content in the e-commerce context. Hence it will specifically analyze consumer's opinions or customer's opinions, also called users reviews. Even though this work narrows the subject domain, nearly the same concepts can be applied and extended to any other domain which is likely to admit opinions.

Basically, customers want to find the best for the lowest price. In other words, they search for products that best fulfill their needs inside a price range that they are willing to pay. It is common to and products with thousands of opinions, thus it could be a hard task for a customer to analyze all of them. Also, it could be a very tiresome work to find opinions about just some features from a product, usually a requirement for an experienced customer. This work presents ways for locating, extracting, classifying and summarizing opinions or reviews on the Internet. The proposed framework will combine several techniques to extract valuable information out of natural language text (user-generated content), in order to provide enrichment of the experience of users by taking advantage of the available content in a more intelligent and organized way. As a consequence of the employed techniques, data can be structured; this will also provide a necessary bridge for many applications to be able to fully interact with others in a Web 3.0 context.

2. BASICS

Web Mining stays at the crossroads of Information Retrieval, Information Extraction and Data Mining. Both Information Retrieval and Information Extraction play important roles to locate and extract valuable information out of unstructured data, before it is suitable to be processed by data mining applications. Exploring better these techniques is extremely necessary to cope with the amount of available data in the Information Overload Era. Also, with the Web being increasingly oriented towards the importance of semantics and integration of information, these areas of study become

very important to address the new future trends of the Web.

2.1.1 Web search engine

Information Retrieval (IR) is a field of study concerned with the retrieval of documents from a collection of other documents (relevant and non-relevant), usually based on keyword searches. With the Internet expansion, Information Retrieval got a very special focus, as search engines

a) Web Crawlers

Web crawlers have two important issues to address: First is to use a good crawling strategy (which includes the algorithm strategy for traversing new web pages) and intelligent mechanisms to optimize the process of re-crawling. Second, because this task is computational intensive, the system must be able to deal with many different scenarios under different circumstances (hardware failure, server problems, errors while parsing documents) while still maximizing the work to ensure that the maximum advantage is taken out of the available resources.

b) Stopword Removal

To optimize the search process and to maximize storage capacity, recent crawled web pages are preprocessed before indexes can be built and the pages safely stored. Stem-ming is a process that reduces words by removing suffixes, thereby mapping them to the same root stem.

c) Inverted Indexes

A search engine system might have to search for billions of documents. Searching all of them for specific terms (from a given user query), would take a great amount of time. To help search engines performing the search in an acceptable amount of time, the retrieval system uses data structured called indices.

2.1.2. Information Extraction

The goal of IE is to identify useful parts out of raw data (unstructured data) and extract them to finally build more valuable information through semantic classification. The result may be suitable for other information processing tasks, such as IR and Data Mining.

2.1.2.1 Natural Language Processing (NLP)

NLP also known as computational linguistics is a field of computer science that studies interactions of human languages with computers. The main goal of NLP is to enable effective human-machine communication, which could be either as spoken or written form. Here, only the written form will be addressed. For many applications, is desirable to automatically process texts written in natural language. Computers can parse and automatically generate natural language texts, extract semantics from them and identify real world objects.

3. Related Work

Many of the researches in Opinion Mining have been placing efforts on product features identification and finding opinion sentiment/orientation. In this works done by Mining and Summarizing Customer Reviews" and A Holistic Lexicon-Based Approach to Opinion Mining" are going to be exposed with more details than others, with more attention to the last one. The reason why these works were chosen among others is their solution for identifying features automatically and sentiment analysis at an optimal granularity level. They also deal with some cases of sentence context, and the efficiency of their methods could be verified on their evaluative results. Also, both define problems that resemble in many ways the one explored by this work, especially for coping with opinions in an e-commerce context. Finally, an important argument favors the study of with more detail. In sentiment is analyzed at the sentence level, while this approach works reasonably well can hide many important details. In this problem is solved through a fine-grained sentiment analysis done at the feature level.

3.1 Defining Opinion Components in a Opinion Mining Context

The main goal of an opinion is to highlight possible strengths and weaknesses about objects under discussion (OuD). Objects can represent a variety of things in the real world, such as products, organizations and persons.
An OuD is defined as a tree and uses the part-of relationship to decompose an object in different components (which in turn can be decomposed by sub-components). An object is associated with the pair O:(T,A), where T is a taxonomy of components (or parts of an object) and possibly sub-components, and A is a set of attributes of O. Like in a tree hierarchy, the components can also have their own set organized. For example, a camera represents the root node and opinions can highlight aspects about a camera attribute as well as attributes of parts of the camera (components).The sentence "This camera has a great design" as an example. Here, design is an attribute of camera (the root node). On the other hand the sentence "The battery life is too short" talks about battery, which is a component of camera and life which is an attribute of battery (battery life). An opinion must not necessarily highlight just attributes of objects or components; they can also refer to the object itself.

Figure 1: illustrates how opinions reference objects.

Example 1:
I. "The battery life of this camera is too short"
II. "This camera is too large"
In the first sentence, battery life is an explicit feature, while in the second one, size is an implicit feature. Size is not mentioned in this sentence, but it is easy to realize that large indicates a negative feature of the size attribute.

Example 2:
I. "The picture quality of this camera is amazing."
II. "This earphone broke in two days."
In the first sentence is clear and explicit that the opinion about picture quality is positive. In the second case however, the opinion about the earphone is not explicit, but one can assume that it is negative, based on the sentence context. Identification of implicit features were not addressed on the mentioned works.

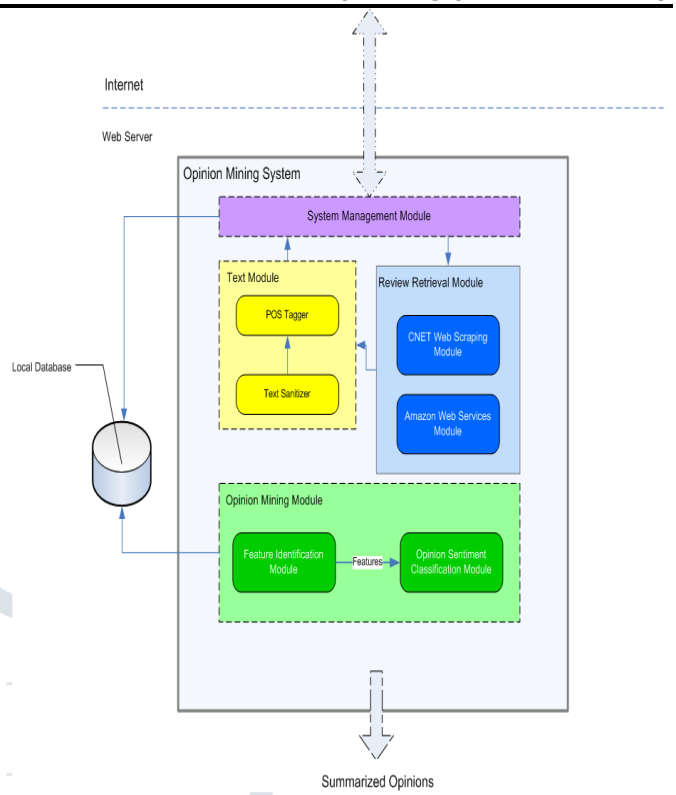

Figure 2: System architecture

The advantage of this architecture is to let the modules perform their jobs asynchronously, which also decouples the long running tasks from the HTTP request/response cycle. Therefore the SMM can serve the user and provide answers in an acceptable amount of time, while still scheduling long running jobs for parallel processment on the background or on later opportunities.

4. Technologies Overview

Different technologies were used in the development of AEOSC. The following considerations concerning the decision about which technologies to use were taken into account:

1. The system must provide strong non-functional requirements guarantees (e.g performance). The system needs guarantees that the non-functional requirements (e.g performance) will be adhered to, so that the whole system can work properly. Without such guarantees the system may become infeasible. As an example, the SMM discussed on the last chapter, has to deal with performance and fault tolerance requirements, and building them from scratch can take a considerable amount of time in developing the prototype. It is wise to consider technologies oriented to agile development, which provides as many as possible ready-to-use" components. Thus, more time can be devoted to the development of functional
requirements, delegating the responsibility of building the whole infrastructure to COTS (Components o_-the-shelf). Also many frameworks using the agile development methodology supports the use of tools that

promote code debugging and verification, ease of installation of new modules and automation of common tasks. Altogether, these features contribute largely to leave the programmer only concerned about domain specific problems.

2. An existing platform called Fedseeko used the same core technologies used by AEOSC, and AEOSC was also intended to be used as a plugin to seamlessly integrate this existing system.

5. Conclusion

During the evaluation of POECS it was possible to see that it is feasible and reliable to build system capable of classifying and organizing opinions through the so called feature-based summary, which resumes the most relevant information for users. However, it is undeniable that a great number of opinions are difficult to classify due to the complexity of the human language.

6. References

[1]. Bing Liu, Sentiment Analysis, Mining opinions, Sentiments, and Emotions, Book, June (2015).

[2]. M.R. Saleh, M.T. Martín Valdivia, A. MontejoRáez, and L.A. Urena Lopez, Experiments with SVM to classify opinions in different domains, Expert Syst. Appl. 38, pp. 14799-14804 (2011).

[3]. W. Medhat et al., Sentiment analysis algorithms and applications: a survey, Ain Shams Eng. J. (2014).

[4]. A. Balahur, Methods and Resources for Sentiment Analysis in Multilingual Documents of Different Text Types, PhD Thesis, University of Alicante, Spain, 273, (2011).

[5]. A. Montoyo, P. Martínez-Barco, A. Balahur, Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments, Decis. Support Syst. 53 (2012) 675–679.

[6]. Y.M. Li, T.-Y. Li, Deriving market intelligence from microblogs, Decision Support Syst. 55 (2013) 206–217.

[7]. D. Kang, Y. Park, Review-based measurement of customer satisfaction in mobile service: sentiment analysis and VIKOR approach, Expert Syst. Appl. (2013).

[8]. H. Rui, Y. Liu, A. Whinston, Whose and what chatter matters? The effect of tweets on movie sales, Decis. Support Syst. 55 (2013) 863–870.

plore tourism

[9]. J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, J. Comput. Sci. 2 (2012) 1-8.

[10]. B. Pang. L. Lee, Opinion mining and sentiment analysis in Multilingual Retrieval 2 (2008) 1-135.

[11]. T. Nasukawa, Sentiment analysis: Capturing favorability using natural language processing definition of sentiment expressions (2003) pp. 70-77.

[12]. X. Ding. S. M. Street, B. Liu, and P. S. Yu, A Holistic Lexicon based approach to opinion mining (2008) pp. 231-239.

[13]. H. Tang. S. Tan. X. Cheng, A survey on sentiment detection of reviews, Expert system application 36 (2009) 10760-10773.

[14]. A. Montoyo, P. Martinez-Barco. A. Balahur, Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments. Decision support system 53 (2012) 675-679.

[15]. D.E. O'Leary, Blog mining -review and extensions: From each according to his opinion. Decision support system 51 (2011) 821-217.

[16]. M. Tsytsarau, T. Palpanas, Survay on mining subjectivity data on the web, Data Mining Knowledge Disc 24 (2012) 478-514.

[17]. E. Marrese-Taylor. J. D. Velasquez, and F. Bravo-Marquez, Opinion zoom: A modular tool to ex