Mining Feedbacks and Opinions in Educational Environments

Rajkumar Kannan
Department of Computer Science
Bishop Heber College
Trichy 620017, Tamil Nadu, INDIA

Maria Bielikova
Faculty of Informatics and IT
Slovak University of Technology
Bratislava, SLOVAKIA

ABSTRACT

People, such as students, employees and public, are talking about the institution and its business everyday positively or negatively by means of feedbacks, opinions, comments etc through various social platforms. Their feedbacks and opinions are valuable resources for the institution if listened properly. Since feedbacks are by and large unstructured in nature, understanding and extracting the meaningful information from massive data collections becomes a real challenge. This paper outlines the various tasks that are to be carried out during the knowledge discovery process from the learning environments setting.

Keywords:

Feedback and Opinions, Knowledge Discovery, Learning Environments

I. INTRODUCTION

People are talking about an educational institution and its business everyday [1]. For instance, students are talking about the institution directly face to face and behind its back. Students are saying how much they like the institution and how much they dislike the institution. They often express what they wish the institution could do for them. Students write a feedback to the institution every year or every semester, post blogs about the institution, discuss about the institution endlessly in public forums and in emails.

The employees of an institution are also talking about it. Employees produce greater ideas that are languishing for lack of right context to apply them. They are looking for the right support from their management to help them innovate things, reveal new ways to improve the internal activities and processes and even change the vision of the institution. Employees' talk can be an inexhaustible resource for innovative ideas and quality improvements of an institution.

Similarly, parents and public who are related to the institution are also passing serious and meaningful comments on the institution positively and negatively [2]. They talk about the institution over telephone, public forums, chat rooms and emails that will enable the institution to become the leader among its peers, if listened carefully.

This paper explores how to listen to the talk of the students, employees, parents and public and to convert the talks into valuable resources for the educational institution. To be precise:

- Collect feedbacks, opinions and comments as unstructured text from different information sources such as feedback online forms, emails, blogs, public forums, voice transcripts, chat rooms text, newspapers and televisions.
- Perform ETL (Extraction, Transformation and Loading) preprocessing to remove noise from the information sources.
- Cluster the preprocessed feedback data into meaningful categories by applying K-Means and Intuitive clustering algorithms to create taxonomy.
- Edit the taxonomy by performing rename, merge and split clusters operations to obtain refined set of clusters by applying *cohesion* and *distinctness* metrics.
- Visualize the categories (i.e. clusters in a high dimensional space to understand the feedback documents of a category and the relationships among other categories.
- Discover patterns and relationships that are inherent in the data through correlation and classification techniques.

The rest of the paper is organized as follows: Section 2 illustrates the type of feedback data with an example feedback. Section 3 defines the feedback and opinion mining problem in the context of massive data. The feedback domain and the institution ecosystem are described in section 4. The phases of knowledge discovery from feedbacks and other sources are discussed in section 5 with a special discussion on feedback preprocessing. Finally, section 6 concludes the paper by specifying further research.

II. Unstructured nature of Feedback Data

Students provide valuable feedback every year as they go out of the institution after their graduation or progress themselves to next year of their academic study. The information they supply about their course, facilities and others are highly unstructured in their own language. But, these unstructured data can be much useful for the institute to shape up the curriculum, teaching methods, faculty improvements, infrastructure, its vision statement, students' facilities and so on. However, the unstructured nature of the feedback is relatively complex and large volume of feedback data requires automated analysis [3,6]

Besides unstructured information, feedback from students may also include structured information [4]. The structured information includes details about a student, his course of study, contact, his teaching faculty, facilities etc. The unstructured information is the free text by which students can express anything about everything care freely which we call as comment or opinion. Here is what a typical comment of a student looks like:

> I admire this university and this department especially for such a lovely infrastructure in terms of buildings and computing labs.

> However, the syllabi for the Masters programme need substantial revamp in tune with the recent trends in IT.

Similarly, employees of the institution provide their feedbacks and opinions to the management of the institution about what they feel as faculties and what they expect from them for better teaching and learning processes. One such opinion can be:

> Excellence and innovation are the real drivers for any institution. In fact, excellence through innovation is the key for success. Innovative teaching, learning and management make the institution to attain excellence. We are lucky enough to work for an institution that makes an impact among the rest.

III. The Problem

Imagine 10,000 or more of these meaningful comments and opinions in databases. In the database, they can be indexed and sorted based on year. But this large collection of comments cannot answer even this simple query, 'what are the curriculum related problems reported by the students in this year'. If the data could be leveraged to do this analysis, then attention or focus can be given to those courses that require intense revision, thus significantly improve the quality of the curriculum.

So why was it so hard to answer this question with the data? The reason is that the data is unstructured [5]. There is no set of vocabulary or language of fixed terms used to describe each opinion. Instead, the students and employees describe their feedbacks in ordinary every day language as they would describe to a peer in the institution. As in the normal conversation, there is no consistency of word choice or sentence structure or grammar or punctuation or spelling in describing their comments [7].

This kind of unstructured information is a free text, by which humans have been communicating with each other over thousands of years. Potentially, it is the most valuable and if the hidden pieces are aggregated and summarized, can communicate intelligence about how the institute is running, how its students and employees perceive it, what is going right and what is going wrong, and perhaps solutions to the important problems the institution faces. These comments are the examples of the sources of information for possible mining of the hidden knowledge [8].

IV. Feedback Domain and Institution Ecosystem

Once the problem of discovering interesting and meaningful knowledge from the feedbacks, opinions and comments from the different information sources such as blogs, emails sent to the institutions, suggestion boxes etc has been clearly identified and defined, the next process in the feedback mining is to perform the different phases of the discovery steps. Before describing the different phases of feedback mining, we identify the characteristics of feedback information and what we can learn from feedbacks.

A. Characteristics of Feedback Information

There are common characteristics that are frequently recorded for all student interactions, which typically contain both structured and unstructured components [9,10,11]. The structured information can be:

- ID – Student ID
- Student Info – Name, Course, Address

Unstructured information captures everything else that happened during their course of study as a free text. This unstructured information helps us achieve excellence, goal and vision objectives of the institution.

A good information source is one that should address the vision and processes of the institution and should have one free text field to provide unstructured information, besides the required structured information. The data additionally can also have time stamps for predicting possible trends.

B. What Can We Learn from Feedbacks?

Unstructured information related to students, employees, parents and public could teach us many things about our views and excellence towards them. For example,

- What are the most common issues that our students have?
- What are the most common issues that our faculty and employees have?
- Where are the areas of dissatisfaction of our students?
- Where are the areas of dissatisfaction of our faculty and employees?
- Who are the faculty doing good job?
- What are the areas where the cost can be reduced?
- What are the expectations of parents of students from the institution?

C. The Institution Ecosystem

An institution is not an isolated entity in a society. The institution exists in a complex network of students, parents, employees, public, and suppliers of things, which we call the institution ecosystem as illustrated in Figure 1.

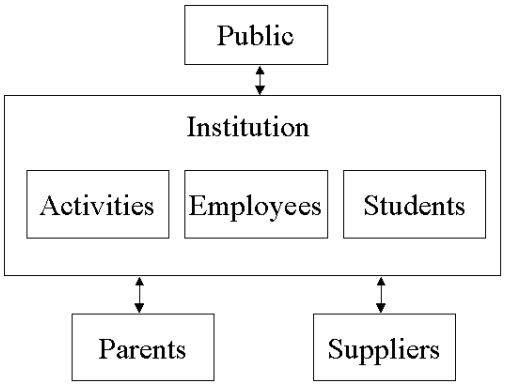

Figure 1. The Institution Ecosystem.

Here, the human knowledge exists in various forms: written, spoken and otherwise. This knowledge is manifested in different media such as newspaper, TV and the Web. With the growth of the Internet and electronic media the rate of growth of the information increases exponentially. However, figuring out what is important in this unstructured information is relatively difficult.

D. The Phases of Feedback Mining

The first step in feedback and opinion mining is the collection of information from the people who are related in the Institution Ecosystem. Then, statistical techniques and algorithms can be used to capture the domain expertise. That is, the underlying

structure inherent in the unstructured information has to be understood and appropriately modeled. Finally, unknown feedback documents can be classified based on the clustered feedback documents. The different phases of the discovery task are detailed in following paragraphs.

E. Feature Selection

Here we need to decide what events we are going to measure and what statistics we will keep. It all depends on what we want to learn and what kind of text data we are dealing with. We can use word and phrase occurrence as the features of the feedback document [12]. However, we will not use every words and phrases. This is because, they are larger in number and not all are meaningful. Therefore, feature space is reduced to a meaningful size by

- eliminating stop words which are frequently occurring and insignificant (Eg. a, an, are, as, be, at)
- stemming which is a process of reducing words to their stems or roots by removing prefixes and suffixes (Eg. the word *Walking* is stemmed to *Walk*)
- removing infrequent features such as digits and hyphens
- indentifying text fields, anchor text and removing HTML tags in web pages
- detecting duplicates of feedbacks on the Web

Several methods can be used to find duplicate information [2]. They are:

- Hash the whole feedback and
- Compute checksum
- *N-gram* method

These hash table and checksum methods find exact duplicates. *n-grams* is an efficient duplicate detection technique. An *n-gram* is a consecutive sequence of words of a fixed window size *n*.

Definition.1. Jaccard Coefficient. Let $N_n(d)$ be the set of distinctive *n-grams* contained in document *d*. Each *n-gram* can be coded with a number. Given the *n-gram* representation of two documents *d1* and *d2*, similarity of two feedbacks can be

$$Sim(d_1, d_2) = \frac{|N_n(d_1) \cap N_n(d_2)|}{|N_n(d_1) \cup N_n(d_2)|}$$

The threshold determines whether d1 and d2 are likely to be duplicate feedbacks.

F. Clustering

Clustering is the process of automatically grouping documents into thematic categories. These meaningful categories constitute taxonomy. Taxonomy provides an overview of what information the feedback document collection contains [1]. We use the variations of K-Means and Intuitive clustering algorithms, because they are fast and give *distinct* clusters reasonably.

G. Taxonomy Editing

Though clustering is a nice technique, it is not sufficient because of the unstructured nature of the feedbacks and opinions and the variations of the language style of different people. Therefore, taxonomies of feedbacks should be supported with facilities to edit them to access the strength and weakness [1]. The following editing steps can be performed.

- Appropriately renaming the category of feedback documents
- Merge the similar feedback categories based on *distinctness*
- Split the feedback categories based on *cohesion*
- Remove one feedback document from one cluster and place in another

H. Visualization

Visualization is an important phase in feedback mining because the visual cortex occupies one third of the surface of the cerebral cortex in humans. Taxonomies of feedback clusters can be visualized as pictures of information to identify possible patterns or relationships. There are several types of visualizations to show the structured and unstructured information of feedbacks and opinions such as scatter plots, trees, bar graphs and pie charts [16,17,18]. Visual representation of text can be done from the information that is represented using Vector Space Model in a high dimensional space.

I. Pattern Discovery

Once the appropriate taxonomy representing the feedback and opinion has been constructed along with the feature set of every document, patterns and relationships inherent in the data can be discovered.

J. Correlation

Taxonomies capture the concepts embedded in unstructured information of feedbacks. Co-occurrence analysis reveals hidden relationships between these concepts and other attributes or categories of different taxonomies. For example, we can look for a relationship between course types and placements to see which courses lack campus recruitments or we can find a correlation between a faculty and the results.

K. Classification

Once we have obtained a good taxonomy of feedback that models the important aspects of the institution, we can apply any classification scheme to classify new unstructured feedback document. We apply the popular classification algorithms C4.5 and Naïve Bayesian Classifier based on the feedback documents in the categories as training data. We can classify feedbacks' sentiments as positive opinions, negative opinions, anger, frustration, enthusiasm, encouragement and so on using Sentiment phrases [13,14,15] on the entire feedback or on the sentence level in a feedback using Semantic Orientation.

CONCLUSION

Feedbacks and opinions play a vital role in any institution or enterprise to achieve the goals of it and to enlarge the horizon of popularity substantially. Every feedback or opinion tells something important in it and that is why we mine it for possible knowledge with the help of domain knowledge. Domain knowledge is necessary because one should not expect the computer to say completely what is relevant. Hence, the talk of the students and employees, if converted into information, will be valuable resources for the institution. Further, we have collected opinions from several subjects and working on evaluation of defined process and techniques.

REFERENCES

[1] Spangler, S and Kreulen, J (2002). Interactive methods for taxonomy editing and validation, *Proc. of ACM CIKM 2002*.

[2] M. Hm and B. Liu (2004). Mining opinion features in customer reviews. *Proc. of AAAI'04*. 755-760.

[3] S. Kim and E. Hovy (2004). Determining the sentiment of opinions. *Proc. of Intl. Conf. On Computational Linguistics (COLING'04)*.

[4] J. Weibe and R. Riloff (2005). Creating subjective and objective sentence classifiers from un-annotated texts. *Proc. of CICLing. 486-497*.

[5] TheresaWilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, JanyceWiebe, Yejin Choi, Claire Cardie, Ellen Riloff, Siddharth Patwardha. OpinionFinder: A system for subjectivity analysis, *Proc. of. HLT/EMNLP 2005*. 34–35.

[6] Yejin Choi, Eric Breck, and Claire Cardie (2006). Joint Extraction of Entities and Relations for Opinion Recognition. *Conf. on Empirical Methods in Natural Language Processiong (EMNLP-2006)*.

[7] Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan (2005). Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. *Proc. of Human Language Technology Conf./Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, Canada.

[8] Ellen Riloff (1996). Automatically Generating Extraction Patterns from Untagged Text. *Proc. of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*. 1044-1049.

[9] Ellen Riloff and Janyce Wiebe (2003). Learning Extraction Patterns for Subjective Expressions. *Conf. on Empirical Methods in Natural Language Processing (EMNLP-03)*. ACL SIGDAT. 105-112.

[10] Ellen Riloff, Janyce Wiebe, and Theresa Wilson (2003). Learning Subjective Nouns Using Extraction Pattern Bootstrapping. *Seventh Conf. on Natural Language Learning (CoNLL-03)*. ACL SIGNLL.

[11] Robert E. Schapire and Yoram Singer. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3): 135-168, 2000.

[12] Janyce Wiebe (2002). Instructions for Annotating Opinions in Newspaper Articles. Department of Computer Science Tech. Report TR-02-101, University of Pittsburgh, Pittsburgh, PA.

[13] Janyce Wiebe and Ellen Riloff (2005). Creating subjective and objective sentence classifiers from unannotated texts. *Sixth Int. Conf. on Intelligent Text Processing and Computational Linguistics (CICLing-2005)*.

[14] Janyce Wiebe, Theresa Wilson, and Claire Cardie (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, volume 39, issue 2-3, pp. 165-210.

[15] Theresa Wilson, Janyce Wiebe and Paul Hoffmann (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Proc. of Human Language Technologies Conf./Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, Canada.

[16] Yejin Choi, and Claire Cardie (2007). Identifying Expressions of Opinion in Context. Eric Breck,. *Twentieth Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 2683-2688 .

[17] Veselin Stoyanov and Claire Cardie (2006). Partially Supervised Coreference Resolution for Opinion Summarization through Structured Rule Learning.. *Proc. of Empirical Methods in Natural Language Processing (EMNLP)*, 2006. 336-344.

[18] Veselin Stoyanov and Claire Cardie (2006). Toward Opinion Summarization: Linking the Sources. *COLING-ACL'06 Workshop on Sentiment and Subjectivity in Text*, 2006. 9-14.