Mining E-Mail Content for a Small Enterprise

Emmanuel Udoh
Department of Computer Science
Indiana University-Purdue University
2101 E. Coliseum Blvd, Fort Wayne, IN 46805 USA

Abstract – **Emails constitute a rich source of a company's information, replacing fax, letters, phone and memos as the dominant form of inter- and intra-business communication. An email system is now a place where task is received, managed and delegated in a company. For small companies, the backbone of modern business, sieving and analyzing tons of business emails consume much business time. There are abundant data mining techniques that can assist email analysis. This paper describes a web-based approach to parse and mine email logs from a POP3 server for content information. The email system affords users a computer-aided tool for decision making.**
can be associated with diseases with improved visualizations.

Keywords - **Business Email, Data Mining, Decision Support System, JSP, Tomcat, Oracle.**

I. INTRODUCTION

Computer-aided decision making tools are essential for modern day business managers in view of the continued proliferation of information in electronic format, especially email messages. Managers spend a big chunk of office hours daily sieving through emails, since emails are now the dominant form of inter- and intra-business communication. Information once transmitted by phone, letters, memos or fax is now sent via email. From a company wide perspective, there is value in mining email content. Content analysis affords firms means to manage such large datasets and efficiently serve customers [3,4]. This paper focuses on mining emails as a business decision making tool. In this vein, the author is working with a small local engineering firm to mine email datasets, such as product information requests, technical support queries and job applications.

Content analysis seeks patterns in textual data [1,5]. Emails are mostly textual messages, but increasingly HTML and XML messages are also available. Email messages are derived from two different applications running on a server machine [1]. The simple mail transfer protocol server (SMTP with listener commonly at port 25) handles all outgoing emails, whereas post office protocol server (POP with listener at port 110) or Internet mail access protocol server (IMAP with listener at port 143) handles incoming mails. For mail mining purposes, POP3 or IMAP servers (provided by JavaMail distribution) are of interest. POP is one of the most popular mechanisms for retrieval of email messages and therefore likely to be of interest to the largest number of readers [1].

Researchers agree that harnessing the large email datasets requires data mining technologies [1,8]. There are research activities in content analysis of email for different purposes, such as automated filing, spam detection, filtering, and personalization services [7]. There are also email behavior analyses for forensic and intelligence purposes [8]. In some of these works, security tends to be the dominant consideration, since emails can be used for legitimate (document distribution) or illegitimate purposes [6,7]. Some misuse cases are virus applications, spambot activity and security policy violations [8]. These misuse examples can also be automatically detected by data mining techniques. But this paper focuses on an email system as a decision support system for a small business, barring anomalous activities by fraudulent Internet users.

An email mining system is a decision support system in the broad sense. It may not have all the paraphernalia associated with online analytical processing (OLAP) systems, but it can help managers compile information from emails for decision making. Thus, the primary objective of this work is to develop an online demo email system that harvests emails from a POP3 server and mines the content with techniques like classification, clustering and ranking for decision making. These techniques are essential because

T. Sobh (ed.), Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, 179–182.

users naturally like to deal with the most important messages first, which can be obtained easily through classification, clustering and ranking.

II. TECHNICAL APPROACH

The developed prototype email system has conceptually two components – the email extraction unit and the data mining component. A database is used as a conduit or plays an intermediation role in facilitating the interaction between both components. Database is essential because archiving strategy affects retrieving efficiency. To actualize this approach, a three-tier architecture consisting of a browser front-end, a Tomcat application server middleware and an Oracle database backend was implemented. This web-based system has JSP providing the glue for the efficient execution of the logic layers as well as the background processes.

The extraction unit is a JSP package deployment running on a Tomcat server that connects to the email POP3 server, parses the email content, and eventually stores the output in an Oracle database server. As part of the JSP package, a mail parser wrapper provides a common interface to the disparate formats of the email protocols, determines the content type (plain text, HTML or XML) and ultimately invokes the correct parser.

Furthermore, the JSP package contains a filtering engine that picks out specific HTML, XML or plain text elements from the emails according to the positions or content of these elements within the structure. The filtered data can be stored or output to another unit. In addition, the extraction unit functionally sorts the emails separately, converts them to strings and stores them also in a vector for easy manipulation. Ultimately, the extracted data can be accessed by the data mining unit for manipulation after authentication.

The data mining unit visualizes the stored data in the Oracle database to the user. Currently, it implements three data mining techniques, namely classification, clustering and ranking. The algorithms first classify the emails into major components, then group them to clusters and finally rank them. Each method extracts different types of information from the email dataset. Clustering makes explicit the relationship between the emails, while classification identifies the key topics of the emails. Interested readers are referred to [1,3,4] for elaborate explanations of these algorithms. Based on the analysis of the email, profiles can be generated by extracting the frequencies of certain terms. Clustering and filtering can be carried out on the basis of both repetitive occurrence and co-occurrence, thereby providing a coherent picture of the functional relationship among large and heterogeneous email dataset.

III. RESULTS

To mine emails from the developed system, a user has to be authenticated by username and password. The system verifies the entry with the Oracle server, then logs into the POP3 server and retrieves all e-mails in the user's inbox/archive. Figure 1 shows the login screen, while Figure 2 shows email messages retrieved from the POP3 server. These e-mails are sorted into individual e-mails, converted into strings and then stored in the database. As one of the system functionalities, a user can extract information from the retrieved data, by selecting a type from a dropdown list (currently implemented are: Personal, Business, Accounting, Production and Engineering). Each type has an associated list of keywords that resides in a table in the database. Upon selecting a type, a result page (Figure 3) appears showing the total number of e-mails available, how many are now classified, a listing of all the keywords assigned to the type, available clusters and the number of hits for each word.

Another functionality provided by the system is the automatic clustering or labeling of the main topics instead of the selection as described above. The system does that by grouping appropriate keywords and detecting the topics through the parsing and extraction of the email content for easy decision making. It also exploits the domain knowledge and indexing for the retrieval of the needed information from the email datasets.

However, this system does not analyze emails for viral or spam activities. It presupposes that the emails are free from virus, spam and other malicious components. There are several approaches to achieve that, using for instance antivirus scanners like McAfee, Norton or Malicious Email Tracking (MET) systems. Many firms regularly use firewalls and antivirus