Content Based Email Classification System by applying Conceptual Maps

S.Baskaran
Associate Professor, Department of Computer Science
Tamil university
Thanjavur – 613 010, Tamil Nadu, India
sbaskarantj@yahoo.com

***Abstract*-This paper discusses the possibility of adopting the concept of Knowledge Based Systems [KBS], in general, and conceptual maps, in particular, in email classification system. Needless to say that email has the potential to improve efficiency and reduce costs involved in communication. Even after the advent of newer technologies such as instant messaging and VoIP, email remains the top most application of the Internet, intranets and extranets. Though the email system is popular, powerful, cost-effective and efficient, it has some shortcomings viz., [a] it remains an unmanaged medium and [b] it is a vast store of unstructured information. Hence, email database will effectively be used by the knowledge workers only if: [a] it enables classification within stringent legitimate frame work [b] it can be searched/sorted and mined in ways that make it useful. Email classification, a technique for adding metadata and visual labels to email, offers an effective strategy for managing and controlling email. Most of the existing systems are based on term-based classification schemes. To achieve better accuracy, semantic information must be made use of. This proposed system could be an add-on application of existing computational as well as term-based methods.**

Keywords: conceptual maps; E-Mail data base; knowledge based systems [KBS]; lexicon; normalization; stemming.

I. Introduction

A number of e-mail classification software like Titus email classification, Sophia 2.1 and spam-fighting tool are available to automatically filter e-mail into any number of categories. In addition they provide an easy to use method for users to select a security schemes that ensure the confidential information does not get leaked [1]. Our proposed system that would be based on domain specific lexicon and conceptual maps could help to classify emails according to business, personal and other relationships. The conceptual maps would provide a means for arranging and displaying classified email messages in a hierarchical structure. This paper discusses a new type of lexicon and conceptual maps and the usages of these techniques for email classification.

II. Components of E-Mail

E-mail consists of 4 components: 1. an Envelop [Message Transfer Agents – MTA Uses] 2. Headers [Mail program or User Agent uses] 3. Body [the recipient uses] and 4. Signature. The Headers are pieces of information that provides a lot of information not only to the user but also to the email system. The most common headers are: Return-Path, Date, From, To, Subject, cc, Comment, In-Reply-To, X-Special-Action and Message-ID etc. MIME extends five more new headers viz., MIME version, Content-type, Content-Transfer-Encoding, Content-ID and Content-Description. Type of the content may be any one of: Text (Plain, Rich text, Enriched), Multipart (Mixed, Parallel), Message (RFC822, Partial, External-Body), Application (Octer-Stream, Postscript), Image (JPEG, GIF), Audio (Basic) and Video (MPEG). The body of the message is the primary information to the receiver. The signature is a sequence of lines that give some information about the sender such as full name, mailing address, phone number and fax number etc. Signature part is optional. The signature details can be stored in a separate file. In general, the name of the file may be signature, .sig, .signature [2]. This file can be appended in an e-mail message. In addition to sending and receiving messages, email systems provide the following features for disposing and processing received email: deleting, downloading, filing, forwarding and replying.

III. Email Database and Classification

Email database that stores a large number of email messages can be considered as a semi-structured database consisting of text data. This text data base should be structured so as to make enable to facilitate multidimensional search by sender, by receiver, by subject, by time and by content [3]. Email classification is essential for storing messages in user-specific folders and information extraction. A sound email classification strategy, deployed in combination with a content management system enables the organization to capture, analyze and retrieve intelligence such as market insights, product or service ideas and more.

IV. Uses of Classification

A. Protect from information leakage

Email classification can protect the leakage of client information by preventing sensitive information from being shared inappropriately.

B. Retention of information

Email classifications can also be used to help with proper retention of information for the desired period. This ensures sensitive emails are not inadvertently deleted.

978-1-4244-4711-4/09/$25.00 ©2009 IEEE

C. *Specify security*

Multi-level email labels created by email classification system are used to specify security levels, releasable markings and legitimate recipients for emails.

D. *Manage*

The classification labels/metadata added to message headers yield benefits to the systems like archiving, content and document management systems.

E. *Control*

Email Classification System block transmission of the message and prompt the sender to remove unauthorized addresses before permitting distribution.

V. Consideration in the classification

The major consideration in the classification is under what features the classification should be done. Manco, et al. defined three types of features to consider in email: unstructured text, categorical text, and numeric data. Unstructured text in email consists of fields like the subject and body. Categorical text includes fields such as “to" and “from". These fields are used in classification using a bag-of-words approach. Numeric data in email includes such features as the message size, number of recipients and counts of particular characters.

VI. Related work [Various Approaches of Email Classification]

The Athena system of Agarwal et *al.* uses a clustering algorithm called C-Evolve and creates folders (classes) for different topics and route e-mail messages automatically to the appropriate folders. CELI's system for email routing and answering integrates statistic techniques (mostly based on keyword identification) with information extraction techniques. Sophia 2.1 is based on full or partial comprehension of message's text. Lewis had used the “in-reply-to" headers in email messages as the truth about thread membership to test his thread detection algorithm. SVM classification system is to classify the folder of each email based on a particular field of data like *From*, *Subject*, *Body*, *To* and *CC*. Titus Message Classification: **[a] c**lassifies and labels Outlook messages [b] applies email markings that clearly identify the existence of confidential and private information [c] encourages proper handling of sensitive information for compliance [d] reduces the likelihood of inadvertent information leakage and [e] makes it easy to determine proper retention period for email.

VII. Proposed Knowledge based System

Our goal is to explore how to classify messages as organized by a human. In this proposed system, it is examined the actual content of body and then perform some knowledge discovery procedure that emphasize meaning. The meaning of a sentence could be found by using lexicon and conceptual graphs [4]. A lexicon relates the words of a language to their grammatical categories and their underlying concepts. In this lexicon, each entry may have the following format: WORD “.” [CATEGORY “|” { TYPE | “no concept”}]. For example, telephone. count noun; TELEPHONE. Transitive verb; PHONE. Conceptual graphs are the logical forms that state relationships between persons, things, attributes, and events. It represents the meaning of a sentence. Let us consider the conceptual graph: COMMAND < MESSAGE. This means is that a command is a message given to a person who is being ordered to do something. [command]<-[OBJ]<-[ORDER]->[RCPT]->[PERSON] [5]. The author has analyzed the suitability of conceptual maps with respect to email folder prediction.

VIII. Major stages of processing

The textual elements of email viz., the subject and body would have been used in classification. The text in these fields is processed to perform: [a] word splitting [b] word normalization [c] detect abbreviation [d] remove stop words [e] word indexing [f] identify noun-phrases by NLP techniques and [g] conversion of phrases into concepts. The author has experimented with the content of some emails. After having carried out the processes of normalization [stemming] and stop word removal, it was found that an email had at most 14% of its total words as distinct words.. The Unified Language System [6] that helps context sensitive classification would be created for several domains with the most frequent concepts used in those domains. A lexicon with the format described in section VII can be created that defines a set of domain-specific terms and the relationships between them like broader term, narrower term, synnonym and related terms. The categoriser part of the calssification could then determine the subject of the email text according to the frequency of the domain specific terms with the help of unified language system and conceptual maps. Thus our proposed system will have content analyser and classifier. Metadata that is created during the classification processes can be stored and viewed.

IX. Conclusions

In general, classification is relied on the belief that the knowledge worker is the most appropriate person to classify email for security, retention and project bases. Our scheme in combination with a content management system and a knowledge based system can automatically classify email messages into user-specific folders.

References

[1] Using Classification to Manage Email Policy for the Enterprises, Whitepaper, Titus Labs, May 2007

[2] Ernest AcherMann, Learning to Use the Internet, BPB Publications, 1996

[3] Haralampos Karanikas and Babis Theodoulidis. Knowledge Discovery in Text and Text Mining Software. http://www.crim.co.umist.ac.uk.

[4] Eugene Charniak, Drew McDermott, Introduction to Artificial Intelligence, Addition-Wesley Publishing Company, 1985.

[5] J.F.Sowa, Conceptual Structures: Information Processing in Mind and Machine, Addition-Wesley Publishing Company, 1984.

[6] Qinghus Zou, Wesley W.Chu, Craig Morioka, Gregory H.Leazer and Houshang Kangarloo. IndexFinder: A method of Extracting Key Concepts from Clinical Texts for Indexing.