Available online at www.sciencedirect.com
ScienceDirect
Decision Support Systems 42 (2006) 1350–1365
Decision Support Systems
www.elsevier.com/locate/dss

A multi-attribute, multi-weight clustering approach to managing "e-mail overload"

David Schuff [a,*], Ozgur Turetken [a,1], John D'Arcy [b,2]

[a] Fox School of Business and Management, Temple University, 1810 N. 13th Street, Philadelphia, PA 19122, USA
[b] Department of Computer and Information Sciences, Towson University, 8000 York Road, Towson, MD 21252, USA

Received 4 April 2005; received in revised form 2 November 2005; accepted 7 November 2005
Available online 20 December 2005

Abstract

The increasing volume of electronic mail communication threatens to cause a state of "e-mail overload" where the volume of messages exceeds individuals' capacity to process them. To address this problem, this study extends the application of hierarchical clustering to the domain of e-mail. We report on the design and development of a system that applies a multi-weight, multi-attribute clustering approach to a collection of messages. We found strong evidence that clustering messages improves users' ability to locate messages compared to an ordered list, and promising (though weaker) evidence of even greater improvement when given the ability to adjust attribute weights.

Keywords: Electronic mail; Information overload; Clustering; Empirical study

1. Introduction and motivation

Electronic mail (e-mail) has become one of the most powerful communication tools in the workplace. The sheer volume, however, of electronic messages can be a burden; experienced e-mail users receive between 100 and 200 messages per day [14]. This number is expected to increase—a recent study by IDC Market Research predicts the number of worldwide e-mail messages to nearly double from 31 billion to 60 billion by 2006 [22]. Considering the massive number of messages a typical e-mail user receives, many messages are left in the mailbox to be processed later [25,49]. The increasing volume of these messages can become particularly overwhelming since messages are often not sequential and span multiple topics. While the number of e-mail messages continues to grow, the limits of individuals' processing capacity remain the same [29]. Hence, the growth of e-mail communication threatens to cause "information overload"—a state in which the amount of information that merits attention exceeds an individual's ability to process it [24,40].

It has long been recognized that individuals tend to filter and omit information as the primary ways of coping with high rates of information overload [29]. Although omission may reduce mental workload, the process of filtering, in requiring the user to at least partially understand the content of a message, can increase mental workload. Moreover, people tend to omit important information when they are overloaded, thereby jeopardizing decision-making performance

* Corresponding author. Tel.: +1 215 204 3078.
E-mail addresses: David.Schuff@temple.edu (D. Schuff), Ozgur.Turetken@temple.edu (O. Turetken), jdarcy@towson.edu (J. D'Arcy).
[1] Tel.: +1 215 204 3058.
[2] Tel.: +1 410 704 4769.

0167-9236/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2005.11.003

[40]. Omitting evaluation of the information effectively removes it from the decision-maker's information set, leading to both a reduced ability to make fine distinctions and to unambiguously worse decisions [6]. Therefore, both coping strategies of filtering and omission negatively impact performance [21].

A recent survey [4] highlights a growing awareness of the "e-mail overload" problem, indicating that practitioners are actively searching for strategies to reduce e-mail overload (earlier identified by Cranor and LaMacchia [10] in the context of eliminating unwanted "spam" messages). We believe that a system that automatically classifies e-mail into groups is a potential solution for e-mail overload. Such a system can be designed in a way to limit the cognitive load on users even as the volume of e-mail messages (and the information contained within them) continues to increase. This requires the e-mail management system to present messages in a form that is consistent with the way people process and store information. Semantic network theory [9] contends that people encode information into a structured network of nodes. Organizing information into logical groups of related entities should thus assist users in comprehending and retrieving the information contained within those groups [35].

In the case of e-mail, the current methods of creating such a group structure involves assigning messages into manually created folders based on the content of the incoming message and the content of the other folders [26,49,30]. However, the user must still absorb the initial cognitive burden of creating the folder structure. We propose a system that automatically creates the folder structure based solely on the content of the messages in a user's inbox. Although this folder structure is a static snapshot of a user's collection of e-mails at a particular point in time, the technique can be extended to regenerate groups "on-demand" as the collection changes.

The techniques for clustering text are well-developed [39]. In our adoption of these techniques to a collection of e-mails, we observe that several distinct elements of information comprise an e-mail message. In addition to the content of the message body itself, attributes such as date, sender, and recipient may be relevant when forming message groups. Therefore, accounting for multiple attributes when clustering messages should result in clusters that are more useful than those created based on a single attribute. Furthermore, this approach can be made more flexible by allowing users to vary the level of emphasis on each attribute.

In this work, we test the effectiveness of this idea by applying a multi-weight, multi-attribute clustering approach to a collection of e-mail messages. To that end, we developed a Java-based tool to serve as a prototype of this system, which we call the Automatic Clustering E-Mail Management System (ACEMS). Through a controlled laboratory experiment, we tested two versions of the clustering tool—one with and one without adjustable attribute weights.

The rest of the paper is organized as follows. In Section 2, we present the theoretical foundation for our approach through a discussion of semantic network theory. We then review prior research on e-mail management and document clustering. Section 3 presents the details of our multi-weight, multi-attribute clustering method, including an overview of the system (ACEMS) that employs this approach. In Section 4 we present our research hypotheses and in Section 5 we describe the experimental design we used to test these hypotheses. Section 6 presents the analysis of the data and discussion of our findings. Section 7 presents implications and potential limitations of our study. The paper concludes with a discussion of the potential contributions of this work and suggested directions for future research.

2. Background

2.1. Theoretical foundations from cognitive psychology

Well-established theory in cognitive psychology [29] contends that humans organize items into logical groups (or "chunks") as a way of dealing with large amounts of information. Miller found that more information than the generally accepted "seven plus or minus two" limitation of human memory can be retrieved when the information is supplemented with additional attributes such as a relationship to a larger group. Mandler [27] first showed the link between these chunks and human performance on cognitive tasks. Later, Ashcraft [2] found that subjects group lists of meaningful words into categories and retrieve the items by those categories rather than in their original random order, presumably because the words have more semantic meaning due to their association with a chunk.

The relationship between chunks also serves as an aid to memory and comprehension. The central tenet of semantic network theory [9] argues that information is stored in human memory as a network of linked nodes. Presenting information consistent with such a network pattern can convey meaning regarding the information itself [27]. These networks can be arranged in a hierarchical structure to more clearly convey relationships between concepts [34]. Semantic networks have been