# Elimination of Redundant Email

**By Lan Jiang** **Project ID: H114050** **Supervisor: Prof. Wong Lim Soon**

## Background

Email has become a very important tool for people to communicate. Usually there are some redundant emails in the mail folders. It wastes time for user to read or manage these emails.

## Objectives

- Eliminate most of the redundant emails
- Minimize the information lost
- Fast

## Concept

### Types of redundant email

- A and B has the same content
- A is repeated continuously in B
- A is repeated discretely in B
- A is repeated discretely in more than one emails

## Methodology

1. Download emails from mail server
2. Cleanse the content of emails
3. Search a cluster for each email
4. Identify the redundant emails
5. Remove the redundant emails

### Longest Common Subsequence (LCS)

A: {w2, w5, w6, w7, w8, w9, w10, w11}
B: {w2, w3, w2, w5, w6, w7, w8, w12, w11}

After smooth the LCS
A: {w2, w5, w6, w7, w8, w9, w10, w11}
B: {w2, w3, w2, w5, w6, w7, w8, w12, w11}

### Cleanse the content of emails

Remove the following fields from an email:

- attachments
- email header
- HTML formatting symbols
- email system specific formatting symbols
- punctuation symbols, empty lines, and so on

**Output:** a sequence of words

### Search a cluster for each email

**Purpose:** for each email, a cluster will be created to store the emails related to it. To check whether the email is redundant or not, it will be much faster compare it with all the emails in its cluster only.

**Methods:**

- use “Reply-To” field.
- a common line
- a common substring

### Identify the redundant emails

let A be an email, B be an email in the A’s cluster
**Loop1:** if A is substring of B, then set A to be redundant
**Loop2:** let LCS(A,B) be the smoothed LCS between A and B
if (LCS(A,B) = A), set A to be redundant
else if
let Merge(A) be the merged LCS of all LCS between A and emails in A’s cluster.
if(Merge(A) = A), set A to be redundant
else, set A to not redundant.

**Smooth LCS:**

- find the LCS between A and B (Diff algorithm)
- present the matched words in matched blocks
- merge continuous blocks
- remove short blocks

**Merge several LCS:**

- remove overlapped sections
- combine all matched sections

**Remarks** : if only some words in A are not matched, then set A to be suspicious. User can remove the emails selectively.

## Implementation

Language: C#
Software : MS Visual Studio 2005, Chilkat.NET

## Testing

The testing dataset contains 736 messages. Most of the messages were collected from SOC mail folders or newsgroups. 335 message among them are redundant.

| (Testing result) | Redundant | Not redundant | Total |
|---|---|---|---|
| Redundant found | 302 | 1 | 303 |
| Suspicious found | 30 | 6 | 36 |
| Regular found | 3 | 394 | 397 |
| Total | 335 | 401 | 736 |