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\}

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.