Search Reliability

Searching may not work as expected in all cases primarily due to different character set in the search string and in the mails.

Base64 or “quoted-printable” encoded text content blocks and text attachments in the email are first decoded before the search to avoid false negative results.

Searching of the header fields is performed on the received raw data.

In most cases the received mail data contains plain text and html text blocks. Both encode the same/similar content. If the mail contains plain text blocks, only the plain text blocks will be searched. If the mail doesn’t contain any plain text blocks, the html text blocks will be searched.

False positive results may sometimes happen when searching text/html content blocks. MBox Viewer will attempt to extract text from html but the current text extraction solution is not perfect due to simplistic but the low cpu overhead approach.

As mentioned already, the search string and mail text are not decoded to the same character set such as UTF8, so both false positive and false negative matches may occur. Future releases will need to address this limitation by building separate database, upon user request, containing decoded and normalized mail text. Decoding and normalizing mail text is too expensive to do on the fly for large mail sets and it was decided not to do this currently.

If possible, to minimize false positive and false negative matches, use ASCII characters only to compose the search string. Characters from extended ASCII or ANSI character set can also be used assuming all or most of the emails are encoded in one specific ASCII 8-bit extension otherwise number of false positive matches may increase.