

2018年12月1日 NISOC会合 勉強会



Mailing List検索システムの構築

新潟インターネット研究会 塩路 和彦

shioji@shioji.org, shioji@nisoc.or.jp

はじめに

- 昨年は勉強会で「勉強会の歴史」をしました
- 今回、最初は「医療とAI」とか「Precision Medicine」「免疫チェックポイント阻害剤とは」といったネタを考えていました
- が、去年の勉強会の資料を読み直すと、勉強会の歴史を調べるのにmailを探しながら読む必要があり、非常に大変だった。次はMLの検索システムの話をしようかなあと書いてありました

最後に

- 今回のまとめがNISOCのWebリニューアルの助けになればいいなと思います
- そろそろメールソフト(mnewsやEdMax)でタイトルからメールを探すのは限界かも
 - 年のせいもあります
 - 最近老眼が...
- 次はML検索システムの構築について話せばなあ...

と言うことで...

- 今回はML検索システム構築のお話です
- とりあえずやったことを後で振り返られるように記録しておきます
 - Mailの保存環境、問題点
 - MailのHTMLへの変換
 - Namazuによる検索システムの構築

やりたいこと

- members ML、niigata-inet ML、(+ α)の全文検索
- cronとかで一定期間(月1回とか)で自動的にインデックス生成とかまではしない
 - 気がついたときにインデックスを更新
 - 最近mailの流量も少ないし...
- 著作権等の問題あり、とりあえず一般公開はしない
 - あくまで勉強会としてやってみましたという形

Mailの保存環境(1)

- niigata-inet、membersには「shioji@shioji.org」で登録されている
- shioji@shioji.org宛てのmailは.forwardで、別のアドレスに転送し、shiojiのスプールにも保存
- 普段はWindowsからThunderbirdでPopしてmailの読み書き、Pop後は削除
- 別アドレスに転送されたmailはMHの自動分配機能で、ML毎のフォルダに1mail、1fileで保存(MH形式)

Mailの保存環境(2)

- 現時点(2018年11月25日22時)で、niigata-inetは7655番、membersは11540番までのmailあり
- MH形式で保存しているPCではDebianがインストールされており、環境変数LANG=ja_JP.UTF-8になっている
- mnewsがインストールされているが、mnewsはEUC-JPの環境で動くため、端末のエンコードをEUC-JPにする必要あり
- 保存されているmailの状態は？

日本語の扱い

- 最近Subject (Header全体?)はUTF-8またはISO-2022-JPでMIME変換されていることが多い?
 - 「=?ISO-2022-JP?B?hogehogehoge?=
• ISO-2022-JPでBase64変換
- RFC的には間違っているらしい?
- 昔はSubjectだけは英語やローマ字で書くこともあったが、最近はほとんど気にしない
- 昔のMailではSubjectにそのままJISというものもあり
 - 「ESC\$B2q...ESC(B)」

本文の日本語の扱い

- HeaderのContent-Type、Content-Transfer-Encodingを確認
- Content-Typeはtext/plainかmultipart/alternative (MIME) かに、Content-Transfer-Encodingは7bitか8bit かBase64かに分かれる

本文の形式

- Typeがtext、charsetがISO-2022-JP、Encodingが7bit
 - 本文がJIS (ISO-2022-JP) のMail
- Typeがtext、charsetがUTF-8、Encodingが8bit
 - 本文がUTF-8のMail
- Typeがtext、charsetがUTF-8、EncodingがBase64
 - 本文がBase64変換されたUTF-8のMail
- TypeがMIMEの場合
 - 本文がBase64変換されたUTF-8のtextが添付(?)

Quoted-Printable

- 時々変換がBase64でなくQuoted-PrintableのMailがあり、うまくデコードできず文字化けしやすい
- 今回MailをHTMLに変換するMHonArcでもうまくいかないことがあった

その他...

- From:に日本語があったり、Subject:に改行が入っていたり
いろいろな処理に悩むmailもあり
- 昔のように本文はJIS、Headerには日本語を使わないという
暗黙の了解はすでになく、今回説明した以外にもHTML
メールだったり、絵文字が入っていたり何でもあり
- 最近のMailerだとそれでも文字化けすることなく読めるので
気にならないのかも知れません
- mnewsはUTF-8に対応していないので読めないmailが増え
ました

SubjectがUTF-8

```
192.168.10.50 - PuTTY
Mini News Reader Ver 1.20 Copyright(C) By A.Takuma [MAIL][MTAP][MSPL]
フォルダ:Nisoc/member 位置:Top[A:m:-]
マーク 番号 日付 行数 差出人 サブジェクト名 [ (1-41) ]
R 1 06/04 139 tkimura@visi [members 11501] =?UTF-8?B?UmU6IFttZW1iZXJzID
R 2 06/11 111 taka@vos.nag [members 11502] Re: 【2018年度会費納入のお願い
R 3 06/27 112 ishimoto@gin [members 11503] 2018年06月定例会合ログ
R 4 06/27 85 ishimoto@gin [members 11504] 2018年07月定例会合のお知らせ
R 5 06/27 80 shinobu@itec [members 11505] =?utf-8?B?UmU6IFttZW1iZXJzID
R 6 07/07 86 moto.mita01@ [members 11506] SIN-NET
R 7 07/09 86 ishimoto@gin [members 11507] Re: SIN-NET
R 8 07/09 156 moto.mita01@ [members 11508] Re: SIN-NET
R 9 07/09 144 ishimoto@gin [members 11509] Re: SIN-NET
R 10 07/11 152 yasushi@kwci [members 11510] =?UTF-8?B?6KiD5aCxIChSZTogU0
R 11 07/11 149 moto.mita01@ [members 11511] =?UTF-8?B?UmU6IFttZW1iZXJzID
R 12 07/11 84 moto.mita01@ [members 11512] =?UTF-8?B?UmU6IFttZW1iZXJzID
R 13 07/11 105 yasushi@kwci [members 11513] =?UTF-8?B?6KiD5aCx44Gu6KiC5q
R 14 07/12 102 yasushi@kwci [members 11514] =?UTF-8?B?6KiD5aCx44Gu6Kkmz57
R 15 07/23 100 ishimoto@gin [members 11515] 2018年07月定例会合ログ
R 16 07/23 86 ishimoto@gin [members 11516] 2018年08月定例会合のお知らせ
R 17 08/06 88 yasuyuki.kan [members 11517] =?utf-8?B?UmU6IFttZW1iZXJzID
R 18 08/28 100 ishimoto@gin [members 11518] 2018年08月定例会合ログ
R 19 08/28 85 ishimoto@gin [members 11519] 2018年09月定例会合のお知らせ
R 20 08/30 129 kaw@on.rim.o [members 11520] Re: 2018年09月定例会合のお知
R 21 08/30 133 ishimoto@gin [members 11521] Re: 2018年09月定例会合のお知
?:ヘルプ j,^N:次行 k,^P:前行 SPACE,i,:参照 o,q:復帰 Q:終了
```

本文がBase64変換されたUTF-8

```
192.168.10.50 - PuTTY
Date: Wed, 27 Jun 2018 19:00:32 JST
From: shinobu honma <shinobu@itech.co.jp>
Reply-To: members@nisoc.or.jp
Subject: [members 11505] =?utf-8?B?UmU6IFttZW1iZXJzIDExNTA0XSAYMDE45bm0MDfmnIjlr
prkvos=?=
=?utf-8?B?5Lya5ZCI44Gu44GK55+l44KJ44Gb?=
To: members@nisoc.or.jp
In-Reply-To: <20180627164819.C7D8.D74F4412@ginzado.co.jp>

5pys6ZaT77yg5paw5r2f44Gn44GZ44CCDQrli4nlvLfkvJrjga7jgYrpoYzjga/jgIHjgIxDTPVj
gYLjgozjgZPjgozjgavjgI3jgZfjgb7jgZnjgIINCuWunOOB1+OBj+OBiumhmOOBhOOB1+OBvuOB
meOAggOKDQo+IDIwMTgvMDYvMjcgMTY6NDjjgIFJc2hpbW90byBLYXRzdWljaGkgPglzaGl1b3Rv
QGdpbnphZG8uY28uanA+44Gu44Oh44O844OrOgOKPiANCj4g44GE44GX44KC44GoQOS6i+WLmeWx
gOOBp+OBmeOAggOKPiANCj4gDQo+IOasoeOBrumAmuOCiuOAgTIwMTj1ubQwN+aciE5JU09D5a6a
5L6L5Lya44KS6ZaL5YKs44GX44G+44GZ44CCDQo+IOS8muWQiOODu+WLieW8t+S8muODu+aHh+im
quS8muOBr+OAgeS8muWToeS7peWklUOBruaWueOBq+OCguOB1OWPguWKoOOBhOOBn+OB0OOBkeOB
vuOBmeOAggOKPiANCj4g5Z+65pys5q+O5pyI56ys5LiA5Zyf5puc5pel44Gu6ZaL5YKs44Go44Gq
44KK44G+44GZ44Gu44Gn44CBTUzjgpLoqq3jgpPjgacNCj4g44GE44KJ44Gj44GX44KD44KL5pa5
44Gv44Gc44Gy44GK6aGU5Ye644GX44GX44Gm44G/44Gm5LiL44GV44GE44CCDQo+IAOKPiAtLS08
IDA35pyI5a6a5L6L5Lya5ZCIID4tLS0NCj4gDQo+IOOAKOaXpeaZguOAKTIwMTj1ubQwN+aciDA3
5pelKOWcnykgMTU6MDAgLSAxNjo1MAOKPiDjgJDkvJrloLTjgJHmlrDmvZ/luILnlJ/mtq/lrabn
v5Ljgrvjg7Pjgr/jg7zjg7vkuK3lpK7lhazmsJHppKggMzAz6Kyb5bqn5a6kDQo+IGH0dHA6Ly93
d3cuY210eS5uaWlnYXRhLmxnLmpwL2tvc29kYXRlL21hbmFiaXNob2dha3Uvc2hpc2V0c3Uva2Fr
dWtvbWlua2FuL2FyZWFFY2h1by9jaHVvL2NodW91Lmh0bWwNCj4gDQo+IAOKPiDjgJDkvJr1kIjj
[5] line:1 byte:1478/2075 71%
```

MIME形式

```
192.168.10.50 - PuTTY
Date: Mon, 4 Jun 2018 02:11:28 JST
From: Takao KIMURA <tkimura@visionware.jp>
Reply-To: members@nisoc.or.jp
Subject: [members 11501] =?UTF-8?B?UmU6IFttZW1iZXJzIDExNDg4XSDJgJAYMDE45bm05bqm5
Lya6LK757SN5YW144Gu44GK?=
=?UTF-8?B?6aGY44GE44CR?=
To: members@nisoc.or.jp
In-Reply-To: <20180508151432.AA7F.D74F4412@ginzado.co.jp>

--00000000000000e1433056dbfe978
Content-Type: text/plain; charset="UTF-8"
Content-Transfer-Encoding: base64

44GU54Sh5rKZ5rGw44GX44Gm44GK44KK44G+44GZ44CB5pyo5p2R5a2d6ZuE44Gn44GZ44CCDQoN
CumBheOCjOOBsOOBm+OBquOBjOOCieWFfioeoi+S8muiyu+OCkuaMr+OCiui+vOOBv+OBvuOB1+OB
n+OAggOK44KI44KN44GX44GP44GK6aGY44GE44GE44Gf44GX44G+44GZ44CCDQoNCjIwMTj1ubQ1
5pyIOOaXpSjngaspIDEIOjEOIElzaGltb3RvIEthdHNlaWNoaSA8aXNoaW1vdG9AZ2luemFkby5j
by5qcD46DQoNCj4gTk1TTOPkvJr1k6H1kITkvYONCj4NCj4g44GE44GX44KC44Go77yg5LqL5YuZ
5bGA44Gn44GZ44CCDQo+DQo+DQo+IOS7peS4i+OBruOBqOOBiuOCiuOAgeW5tOS8muiyu++8iDIw
MTj1ubQ15pyIMeaXpSAAtIDIwMTn1ubQO5pyIMzDml6XvvInjga7ntI3lhaXjgpINCj4g44GK6aGY
44GE6Ie044GX44G+44GZ44CC57eP5Lya5pmC44Gr57SN5YW144GM44GL44Gq44KP44Gq44GL44Gj
44Gf5pa544Gv5L2V5Y2S5pep44KB44Gu57SN5YW1DQo+IOOBq+OB1OWN1OWKms4i+OBleOBhOOA
ggOKPgOKPiA25pyIOeaXpeOBruS8muWToee2mee2mueiuuWumuS9nOalreaZguOBq+OAgQOKPiDj
g7vlp5Tku7vnirbmj5Dlh7rjgYzmnKrmj5Dlh7rjgIHjgYLjgovjgYTjga/pgKPntaHnlo7pgJrj
[1] line:1 byte:1278/5225 24%
```

昔のSubject

```
192.168.10.50 - PuTTY
Mini News Reader Ver 1.20      Copyright(C) By A.Takuma [MAIL][MTAP][MSPL]
フォルダ:MLS/inet1-info-eng    位置:Top[A:m:-]
マーク 番号 日付 行数 差出人      サブジェクト名 [      ] (1-1535)
R      1 08/26 124 ina@i4esl.in [niigata-inet 1] niigata-inet ML available
R      2 08/27 48 amano@gaplus [niigata-inet 2] Self-intro (AMANO Kei)
R      3 08/29 35 kim@nuis.ac. [niigata-inet 3] Kim Tong Ho
R      4 08/29 166 kim@nuis.ac. [niigata-inet 4] Inter-Regional Mailing List
R      5 08/29 43 siva@scc.co. [niigata-inet 5] Self-Introduction
R      6 08/29 50 SGQ01445@nif [niigata-inet 6] Regist Inter-Regional Maili
R      7 08/30 46 kim@nuis.ac. [niigata-inet 7] Re: Regist Inter-Regional M
R      8 08/30 90 SGQ01445@nif [niigata-inet 8] Re^2:Regist Inter-Region ML
R      9 08/30 66 siva@scc.co. [niigata-inet 9] Re: Regist Inter-Regional M
R     10 08/30 34 e01123@nagao [niigata-inet 10] Self-Introduction from Ken
R     11 08/30 59 siva@scc.co. [niigata-inet 11] Re: Re^2:Regist Inter-Regi
R     12 08/30 54 SGQ01445@nif [niigata-inet 12] Echigo Chiiki Network
R     13 08/30 29 PEB04074@nif [niigata-inet 13] HELLO
R     14 08/30 33 Majordomo@ni [niigata-inet 14] Welcome to regional-office
R     15 08/30 70 siva@scc.co. [niigata-inet 15] Re: Regist Inter-Region ML
R     16 08/30 72 siva@scc.co. [niigata-inet 16] Re: Echigo Chiiki Network
R     17 08/30 42 furukawa@ted [niigata-inet 17] Re: Re^2:Regist Inter-Regi
R     18 08/30 38 20241125@peo [niigata-inet 18] Self-Introduction <KIMURA
R     19 08/31 39 ina@i4esl.in [niigata-inet 19] network is unreachable ??
R     20 08/31 43 PEB04074@nif [niigata-inet 20] RE:RE Regist Inter-region
R     21 08/31 35 ina@i4esl.in [niigata-inet 21] network is unreachable ??
?:ヘルプ j,^N:次行 k,^P:前行 SPACE,i,:参照 o,q:復帰 Q:終了
```

HTMLへの変換

- 長々とMailについて説明しましたが、HTMLへの変換には「MHonArc」を使用します
- 最近のMailはいろんな形式がありそうだけどちゃんと変換できるのかなあとMailでの日本語の取扱いを調べ始めたらドツボにはまりました
- 結局多くの場合特に問題なく変換されました
 - MHonArcは優秀です

以前の方法

- 以前はMHonArcでそのまま変換すると文字化けすることがある(?)ので、一旦漢字コードをEUCに変換してから、MHonArcでHTMLに変換していました
- さらに変換されたHTMLをEUCのままだとBrowser側でうまく表示できないことがあるのでShift-JISに変換していました
- 現在はMHonArcの設定ファイルを記載することで、コード変換せずにHTMLに変換できるようです

漢字コードの選択

- rcfileに記載することで、生成されるHTMLファイルの漢字コードを選択することができますが...
- ISO-2022-JPの設定すると結構文字化けします
- UTF-8にすると文字化けが少なくなりますが、次のNamasuがUTF-8に対応していないようです
- ISO-2022-JPやUTF-8で出力したものを強引にEUCやShift-JISにするとHTMLのcharsetと合わないのでBrowserで表示できません

最終的に？

- 文字化けの少ないUFT-8でHTMLファイルに変換することにした
- いろいろ調べるとnamazuもインデックスの作成は問題なく、namazu.cgiがUTF-8を認識できないようです
 - LocaleがUTF-8のコマンドラインで使う時はTipsがいるようですが、Webで使うなら問題なさそう？
- 試しにインデックスを作成し、Browserで検索したらOKでした

Namazuのインストール

- OSがDebianなので
 - `# apt-get install namazu2 namazu2-index-tools libtext-kakasi-perl kakasi`
- 以下がインストールされました
 - kakasi, kakasi-dic, lib-mmagic-perl, libkakasi2, libtext-kakasi-perl, libnmz7, namazu2, namazu2-common namazu2-index-tools
- NKF.pmがない → 以下でOKでした
 - `apt-get install libnkf-perl`

インデックスの作成

- \$ mknmz -k htmlあるディレクトリ -O インデックスを置くディレクトリでインデックスが作成されます
- あとは検索用のHTMLを作れば完成

今回

- 以前作ったインデックス (EUSに変換してmhonarcでHTMLを作成、sjisに変換してmknmzしたもの) と、今回のmhonarcでUTF-8のHTMLを作成し、そのままmknmzしたインデックスを一緒にしましたが、問題なく検索できました
- niigata-inetの7001からと、membersの9001からが今回作ったHTML、インデックス、それ以外は以前のものです

さあ！検索してみよう

- 古いMLを自分の名前で検索するとはずかしくなります
- mailのならばは必ずしも番号順にはなっていません
- 文字化けしているのもありますが、どうしてもというなら、オリジナルのファイルを変換する必要があるかも知れません