3H6-4

メーリングリストのメタ情報の RSS/iCalendar による配布

Publication of Meta-Information of Mailing-List Contents in RSS/iCalendar Formats

神嶌 敏弘*1
Toshihiro Kamishima

産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

The Japanese Society for AI manages the "jsai-ann" mailing-list for the announcement of AI related meetings and other AI topics. Many articles are actively delivered in this list, so some method to check published information more effectively. To this aim, we developed a method to deliver meta-information's of this mailing lists in RSS/iCalendar formats.

1. はじめに

人工知能学会では,「人工知能学会からのお知らせメーリングリスト」[4] (以後，AI 学会 ML) にて，人工知能や関連領域のイベント情報や話題などをアナウンスしている．この AI 学会 ML は，1998 年 6 月 1 日より正式運用を開始した．2006 年 6 月 1 日までの，購読者数と，各年の年間配布記事数を図 1 に示す．購読者者数は実線，年間記事数は点線で示した．最近は，購読者数は約 2,600〜2,700 人の間で推移しているが，配布される記事数は着実に増加している．

こうした記事数の増加は，購読者の情報の管理の負担を増やしてしまう．この問題を緩和するため，メーリングリストの情報を，二種類のメタデータに変換して配布する「人工知能学会 RSS/iCalendar ファイル」サービスを開始した [3, 5]．配布フォーマットは RSS と iCalendar の二種類で行っている．RSS はブログや Web サイト更新の要約情報を配信するためのフォーマットで，RSS リーダなどを利用して一元的に情報を管理できる．一方の iCalendar は，スケジュール情報の交換フォーマットであり，このフォーマットに対応したスケジュール管理ソフトで利用できる．本稿では，このメタデータの設計指針，データの生成方法，配布システムについて述べる．

2. 節では AI 学会 ML について，3. 節では配布システムの構成について述べる．4. 節と 5. 節では，それぞれ RSS と iCalendar ファイルの設計指針や生成方法について述べる．最後はシステムの実行例とまとめである．

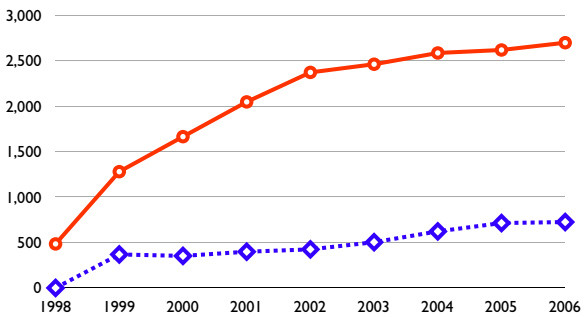


図 1: メーリングリスト購読者数と年間記事数

連絡先: 神嶌敏弘，http://www.kamishima.net/

2. AI 学会 ML

AI 学会 ML では，人工知能に関連したイベント情報や話題を配信している．管理者が許可した記事のみを配信する moderated な体制によって，スパムメールの問題を回避している．また，メールはアーカイブされ，WWW からも閲覧できるようになっている．だが，アーカイブの容量は 2MB の制限があり，通常は 200 件 ほどしか蓄積できない．そのため，約 2ヶ月ほどでアーカイブから削除される．

moderated な運用体制をとったことで，管理者による投稿時に，幾つかのメタ情報を手作業によって記事に付加できる．さらに，メーリングリストの管理ソフトやメール転送エージェントなどが加えるメタ情報も利用できる．これらの情報について順に述べる．

管理者が人手によって付加するメタデータの例を図 2 に示す．メタデータは Subject 行に付加する．最初の `[jsai-ann 4738]` はメーリングリストの管理ソフトが付加するもので，メーリングリスト名 "jsai-ann" と記事番号 "4738" が示されている．それ以後の部分は管理者が付加したもので，次の四つの部分で構成される*1．

{ 記事の種類 {; 付属情報 }:} 略タイトル {(詳細タイトル)}

「記事の種類」は分類を示し，CFP（原稿募集）や MEETING（参加募集）の場合のみ利用している．記事の種類には,「付加情報」をつける場合があり，CFP や MEETING では，関連イベントの開催日や締切日の情報を付加している．多くのイベントでは略記名があるため，略タイトル部に略記名を，詳細タイトル部にフルネームを記す．略記名が無い場合は，詳細タイトル部はなく，略タイトル部にフルネームを記す．図 2 の (1) の例では，記事の種類は "CFP ''で「原稿募集」であることを示し，付属情報として "h=070623" と "d=070417" があり，それぞれ，開催日と締切日を表す．詳細タイトルは「第 8 回 AI 若手の集い」であり，その略記名として「MYCOM2007」が記されている．

メール転送エージェントなどが付加する情報としては Date で示された，メールの発信日時を利用した．Content-Type に付加される charset を，本文の言語を判定に利用することを検討した．しかし，AI 学会 ML の記事は，管理者に届くまでに，各国を転送を転送される場合が多く，本文を反映しない指定になっている場合が多い．また，英語の本文に日本語で少しだけ紹介が加えられる場合などの対処といった問題もある．これらの理由により，charset の情報は利用しなかった．

*1 詳細な書式は [4] にある jsai-ann-info.txt を参照

```
(1) Subject: [jsai-ann 4738] CFP;h=070623;d=070417: MYCOM2007(第 8 回 AI 若手の集い)
(2) Subject: [jsai-ann 4623] MEETING;h=070227: AI 学会 DMSM(データマイニングと統計数理研究会)
(3) Subject: [jsai-ann 4731] OFFICE: 2007 年全国大会「参加申し込み」のお願い
(4) Subject: [jsai-ann 4713] J-STAGE 新着論文のお知らせ [人工知能学会論文誌]
```

図 2: 管理者が付加するメタデータの例

3. システム構成

常駐の強力な管理体制を準備できない事情から，システム構成は簡素にした．配信は次の三つの段階で構成される．

1. 配信されたメールの取得
2. RSS/iCalendar ファイルへの変換
3. サーバへの転送

配信されたメールを fetchmail[*2]によりメールを POP サーバより取り出し，procmail[*3]によりテキストとして蓄積する．これらのテキストから，RSS/iCalendar ファイルへ perl スクリプトにより変換する．その後，1 時間に一度，配信用の WWW サーバへ，生成したファイルを転送している．

4. RSS

RSS は，ブログや Web の更新情報を，利用者からの要求に応じて pull 型で配信するためのメタ情報のフォーマットである．歴史的経緯により幾つかのフォーマットが並立しているが，ここではセマンティック Web[6] の RDF(Resource Description Framework)[1] フォーマットに準じており，日本国内で普及している RSS1.0 (RDF Site Summary) を採用した．人工知能学会では，AI 学会 ML の最近の記事の jsai-ann.rdf，開催日情報の event.rdf，および締切日情報の cfp.rdf の 3 種類のファイルを配信している．図 3 は jsai-ann.rdf の例である．RSS ファイルは，全体の情報を含み，一度だけ現れる channel 部 (7〜24 行) と各記事ごとに繰り返し現れる item 部 (25〜32 行) で構成される．それぞれに含めた情報の詳細を以下に述べる．

本 RSS ファイルには，RSS 自体の語彙の他に以下の語彙を利用している

- `dc`: Dublin Core
  Dublin Core Metadata Initiative が定義した作成者などの語彙
- `sy`: RDF Site Summary 1.0 Modules: Syndication
  更新周期などの同期情報を記述する RSS の拡張語彙
- `content`: RDF Site Summary 1.0 Modules: Content
  Web サイトの一部などの記事を HTML 形式で記述する RSS の拡張語彙
- `ev`: RDF Site Summary 1.0 Modules: Event
  イベント情報を含めるための RSS の拡張語彙

RDF ファイル自体の URI は，ファイルの配布 URL を利用した．channel 部では，RSS の語彙から title，link，description，Dublin Core の語彙からは dc:publisher，dc:creeator，dc:ritghts，dc:language をその意味に応じて適宜設定した．更新日時を示す dc:date には，スクリプトによる RSS ファイルの生成時刻を設定した．実際の更新は 1 時間に一度だが，本 RSS ファイルは通常より大きく通信量が多いので，サーバへの負荷を考慮し，Syndication の同期情報には 2 時間に一度に設定した．

item 部では，各記事を表す URI として，AI 学会 ML の該当記事のアーカイブを参照する URL を用いた．RDF 語彙の title には管理者が付加した略タイトルを抽出して設定，link にはアーカイブの記事を参照する URL を設定した．ここで，dc:date と ev:startdate の内容は，この item 部が，記事を表すのか，記事が言及しているイベントや締切を示すのか異なるという問題が生じる[*4]．前者であれば dc:date には記事の配信

```
 1 <?xml version="1.0" encoding="utf-8" ?>
 2 <?xml-stylesheet type="text/xsl" href="http://www
…
 3 <rdf:RDF
 4   xmlns="http://purl.org/rss/1.0/"
 5   《ファイル中で利用する語彙に関する宣言》
 6   xml:lang="ja">
 7 <channel rdf:about="http://www.ai-gakkai.or.jp/rs
…
 8 <dc:language>ja</dc:language>
 9 <title>人工知能学会 ML(200)</title>
10 <link>http://www.ai-gakkai.or.jp/jsai/ml/</link>
11 <description>《このファイルの案内文》</description>
12 <dc:date>2007-04-11T16:02d:02d+09:00</dc:date>
13 <dc:rights>Copyright(c), The Japanese Society for
…
14 <dc:publisher>人工知能学会</dc:publisher>
15 <dc:creator>人工知能学会</dc:creator>
16 <sy:updatePeriod>hourly</sy:updatePeriod>
17 <sy:updateFrequency>2</sy:updateFrequency>
18 <items>
19 <rdf:Seq>
20 <rdf:li rdf:resource="http://jsai-ann:ai-gakkai@…
  《…中略…》
21 <rdf:li rdf:resource="http://jsai-ann:ai-gakkai@…
22 </rdf:Seq>
23 </items>
24 </channel>
25 <item rdf:about="http://jsai-ann:ai-gakkai@mlwww…
26 <title>C&amp;O:RR-2007</title>
27 <link>http://jsai-ann:ai-gakkai@mlwww.iijnet.or.…
28 <description>《記事のテキストによる説明》</description>
29 <content:encoded>《記事の HTML による説明》</content:…
30 <dc:date>2007-04-11T08:10:32+09:00</dc:date>
31 <ev:startdate>2007-08-21</ev:startdate>
32 </item>
33 <item rdf:about="http://jsai-ann:ai-gakkai@mlwww…
34 <title>IPC-07</title>
  《…中略…》
35 <ev:startdate>2007-06-18</ev:startdate>
36 </item>
37 </rdf:RDF>
```

図 3: RSS ファイルの例

*2 http://fetchmail.berlios.de/
*3 http://www.procmail.org/

*4 神崎正英との個人的議論より

日時を，後者であれば開催日や締切日を設定すべきである．ここでは，記事のアーカイブを示す URL を URI として用いているので記事を示すとみなす方が妥当であり，また，dc:date は記事の配信日時と解釈する RSS リーダがほとんどという実用的な観点から，dc:date には記事の配信日時を設定した．この解釈では，ev:startdate はこの記事の開始日ではなく，この記事が指し示すイベントの開始日としなくてはならない．よって本来は以下のような記述が妥当であろう．

```
<item rdf:about="http://...">
<dc:date>2005-07-26T17:49:20+09:00</dc:date>
<foaf:topic ev:startdate="2006-02-13"/>
</item>
```

しかし，Event モジュールのドラフト[*5]では次のような例が挙げられている．

```
<item rdf:about="…">
<ev:startdate>2001-09-18</ev:startdate>
</item>
```

そのため，Event モジュールをサポートするアプリケーションは，こうした記述を前提とすると予想される．そこで，厳密な意味づけには不都合を生じるが，図 3 の 31 行のようにドラフトの記述に従った．この場合，ev:startdate のセマンティクスは「記事の扱うイベントの開始日」と解釈することになる．なお，ev:startdate を解釈しない RSS リーダも多いため，meeting.rdf と cfp.rdf ではタイトルの先頭に開催日と締切日をそれぞれ挿入し，これらの日付で記事を整列できるように配慮した．

最後に，description と content:encoded にはそれぞれ，記事の概要をテキストと HTML によって記述した．記事の概要は四つの部分で構成される：(1) メタ情報，(2) 記事タイトル，(3) 関連 WWW，(4) 記事の抜粋．(1) は記事番号に加え，管理者が加えた開催日や締切日の情報である．(2) には，管理者が Subject に記した略タイトルと詳細タイトルの両方を合わせて記した．(3) では，記事の本文中から URL の文字列を抽出し記載している．これにより，アーカイブから削除された古い記事でも，会議の WWW ページを直接参照して，会議の内容を知ることができる場合が多くなった．(4) としては，本来は文書要約技術などを用いて要約を作成すべきである．だが，ここでは簡単に先頭の 20 行だけを取り出して用いてる．このヒューリスティックで多くの場合問題は生じないが，先頭に日本語の紹介文がある場合には本文が含まれなかったり，ほとんど改行のない記事などは取り出される部分が長くなりすぎるといった問題を生じる場合もある．

5. iCalendar

iCalendar はスケジュール情報を相互運用するためのフォーマットであり，RFC2445 として標準化されている．人工知能学会では event.ics と cfp.ics の 2 種類のファイルを配布している．それぞれ，開催日や締切日の情報を管理者より付与されているものを AI 学会 ML の記事中から抽出して，このフォーマットに変換したものである．スケジュール管理ソフトでこのファイルを購読することで，日程表の形で一覧し，スケジュール管理に役立てることができる．図 4 に event.ics の例を示す．iCalendar 形式は全体が VCALENDAR 部 (1〜30 行) になっている．この中には 1 度だけ現れる TIMEZONE 部 (9〜18 行) と，各記事ごとに繰り返し現れる VEVENT 部 (19〜29 行) とがある．それぞれの詳細を以下に述べる．

```
 1 BEGIN:VCALENDAR
 2 VERSION:2.0
 3 PRODID:-//www.ai-gakkai.or.jp/ics/event 1.0//EN
 4 CALSCALE:GREGORIAN
 5 X-WR-CALNAME:人工知能関連イベント
 6 X-WR-CALDESC:《このファイルの案内文》
 7 METHOD:PUBLISH
 8 X-WR-TIMEZONE:Asia/Tokyo
 9 BEGIN:VTIMEZONE
10 TZID:Asia/Japan
11 LAST-MODIFIED;VALUE=DATE:20070411
12 BEGIN:STANDARD
13 DTSTART:19510908T020000
14 TZOFFSETTO:+0900
15 TZOFFSETFROM:+1000
16 TZNAME:JST
17 END:STANDARD
18 END:VTIMEZONE
19 BEGIN:VEVENT
20 DTSTAMP;TZID=Asia/Japan:20070410T174017
21 DTSTART;VALUE=DATE:20070623
22 DTEND;VALUE=DATE:20070624
23 SUMMARY:MYCOM2007
24 UID:jsai-ann-4754@ai-gakkai.or.jp
25 URL;VALUE=URI:http://jsai-ann:ai-gakkai@mlwww.ii…
26 DESCRIPTION:[jsai-ann 4754 開催 2007/06/23 締切 200…
27  I 若手の集い）\n 概要:\n\n 第 8 回 AI 若手の集い MYCOM…
 《…中略…》
28  /mllist/jsai-ann/index.cgi/html:4738\n
29 END:VEVENT
 《…中略…》
30 END:VCALENDAR
```

図 4: iCalendar ファイルの例

VCALENDAR 部には，ファイルの ID(PRODID)，配信形式 (METHOD)，更新日時 (LAST-MODIFIED) などの項目がある．VERSION 項目の 2.0 の記述により，iCalendar 形式の前身である vCalendar 形式と区別される．X-WR-CALNAME と X-WR-CALDESC は RFC2445 にはない拡張項目だが，多くのソフトでサポートされている．それぞれ，ファイル自体のタイトルと説明文を表す．

TIMEZONE 部は時差を処理するために必要になる．本来は，記事の開催地を解析して時差を考慮すべきところである．しかし，高度な固有表現抽出などが必要になるなどの問題があるので，海外のイベントでも日本の日時で表示する．このため，タイムゾーンを厳密に解釈するソフトを，日本国外で閲覧すると問題を生じる．

VEVENT には個々のイベントの情報が含まれる．管理者が付与した略タイトルから締切延長など特定の語を除去したものを SUMMARY に設定した．これは，カレンダー中でイベント名などとして表示される．DTSART と DTEND はイベントの開始時刻と終了時刻である．多くのソフトでは，0 時に開始し翌日の 0 時に終了する場合は，その日の時間を特定しない終日イベントと解釈される．管理者が付与した開催日 (event.ics) や締切日 (cfp.ics) の終日イベントとして指定した．DESCRIPTION には，RSS の description と同様の内容を設定してある．他に，識別子を示す UID や，記事へのリンクを示した URL などの項目がある．

∗5 http://web.resource.org/rss/1.0/modules/event/

5.1 レコードリンケージ

スケジュール管理ソフトではイベントの情報をカレンダー上にまとめて表示する．そのため，リストとして表示する RSS リーダより，よりコンパクトにまとめるて表示する必要がある．AI 学会 ML では，同じイベントのアナウンスが複数回配信されることがある．これらの情報は冗長なので，これらをできるだけまとめて，コンパクトな表示に役立てる工夫を行った．このように，同じ実体を参照すると考えられるものを識別する問題は，データベースの分野では record linkage（[2] など）と呼ばれ，1950 年代から研究が行われている．また，機械学習や自然言語処理分野でもこうした問題は identity uncertainty や reference matching などと呼ばれ，研究がなされている．だが，完全に識別するのは困難で，計算量も多いのが現状である．さらに，手法も問題のドメインに深く依存していて，ここで利用するにはかなりの修正を要する．よって，次のような簡単なヒューリスティックを用いた正規化を，略タイトルに行い，重複する記事を検出した．

- '07 などの 2 桁年号を，開催日の情報を勘案して 4 桁の年号に変換
- "SIG-" の除去や，略称名の統一 (人工知能学会⇒ AI 学会) など，この分野に依存した正規化
- 空白や記述記号などの除去と英字の大文字への変換

これらの正規化により略タイトルが同一になり，かつ，開催日が一致すれば同じイベントの記事とみなした．これらの記事の中で，最も新しいものを iCalendar ファイルに含める．だが，偶然に一致してしまう場合や，論文募集とワークショップ提案募集を同一視してしまった場合など，違う内容の記事でも過剰にまとめてしまう問題を生じることがある．そこで，DESCRIPTION の最後に，同一と判定された古い記事へのリンクを，関連記事として添付した．これにより，過剰にまとめられた記事が生じても，利用者はそれらの記事を参照することができる．

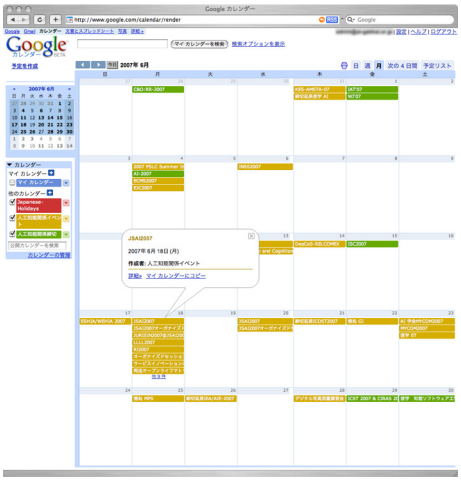
6. 利用例とまとめ

最後に図 5 に，RSS/iCalendar ファイルの利用時の様子を示しておく．

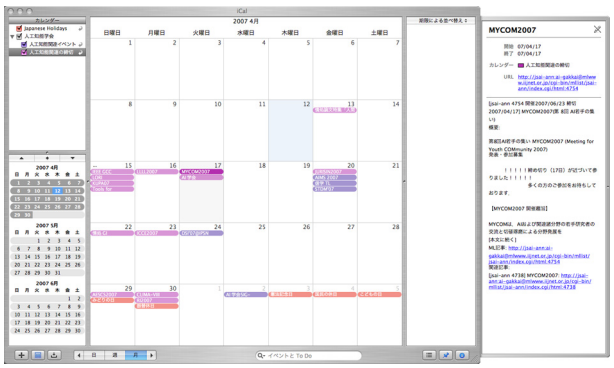
以上,「人工知能学会 RSS/iCalendar ファイル」の設計指針，ファイルの生成過程，および配布システムの構成について述べた．

謝 辞：本サービスの開発にあたり，武田英明先生，浦本直彦様，神崎正英先生，大向一輝先生，ならびに人工知能学会広報・編集・総務委員の方々には貴重なアドバイスをいただき，またテストへ協力いただいた．稲垣良一様には本サービスの稼働のための作業にご尽力いただいた．以上の方々に感謝する．

(a) MacOS X の Web ブラウザ Safari による RSS の表示



(b) Google Calendar による iCalendar ファイルの表示



(c) MacOS X の iCal による iCalendar ファイルの表示

図 5: RSS/iCalendar ファイルの利用例

参考文献

[1] RDF site summary. http://web.resource.org/rss/1.0/.

[2] M. Bilenko and R. J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *Proc. of The 9th Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 39–48, 2003.

[3] 人工知能学会. 人工知能学会 RSS/iCalendar ファイル. http://www.ai-gakkai.or.jp/jsai/event/.

[4] 人工知能学会. 人工知能学会からのお知らせメーリングリスト. http://www.ai-gakkai.or.jp/jsai/ml/.

[5] 神嶌敏弘, 大向一輝. 人工知能学会 RSS/iCalendar ファイルの利用法. 人工知能学会誌, Vol. 21, No. 6, pp. 732–736, 2006.

[6] 神崎正英. セマンティック・ウェブのための RDF/OWL 入門. 森北出版, 2005.