

A-doc に基づく WEB リンク集再利用化ツールの試作

市東 隼[†] 遠山 元道^{††}

^{††} 慶應義塾大学理工学部情報工学科 〒 223-8522 神奈川県横浜市港北区日吉 3-14-1

E-mail: [†]sitow@db.ics.keio.ac.jp, ^{††}toyama@ics.keio.ac.jp

あらまし A-doc とは、WEB においてユーザ主体に情報資源の結合を実現する XML 形式のファイルで、複数のキーワードとそれに対応する文書からなるエン트리を持つ。A-doc ファイルには多くのエン트리があると考えられるが、ファイルを 1 から作成する際、作成者に大きな負担がかかる。ファイルを容易に作成するためには、既存の情報の再利用が必要となる。そこで本研究では、WEB 上の従来型の HTML のリンク集に注目し、そこから A-doc ファイルを生成・編集するツールを試作した。本手法を用いることで、ファイル生成にかかる時間を大幅に削減することができると考えられる。

キーワード A-doc、Web、ハイパーリンク

Making of tool that recycles WEB links based on A-doc

Jun SITOW[†] and Motomichi TOYAMA^{††}

^{††}Department of Information and Computer Science, Faculty of Science and Technology,
Keio University

Hiyoshi3-14-1, Kouhoku-ku, Yokohama-shi, Kanagawa, 223-8522 Japan

E-mail: [†]sitow@db.ics.keio.ac.jp, ^{††}toyama@ics.keio.ac.jp

Abstract A-doc is a file of the XML form that achieves uniting the information resources in the user subject on WEB with entry that consists of two or more key words and documents corresponding to it. The large encumbrance hangs to the A-doc manufacturer though it is thought that there are a lot of entries in the A-doc file when the file is made from the beginning. To make the file easily, recycling existing information is needed. In this paper, therefore, we propose, the tool that generates and edits the A-doc file from HTML links of the old model on WEB. As a result, our tool makes half automatic A-doc file generation possible and the efficiency of it will be improved.

Key words A-doc, Web, hyperlink

1. はじめに

Web における利用者主導による情報資源の結合を実現するために、我々は A-doc と呼ぶ情報資源表現形式の提案、開発を行っている。A-doc は見出し語とそれに対応する文書のペアの集合で、辞書形式であり、それを閲覧中の Web 文書に結合すると、閲覧中の文書の見出し語部分が自動的に対応する文書へのハイパーリンクに変換される。

A-doc には利用者と生成者があり、生成者が予めプールしておいた A-doc リストを利用者が任意に使う形となる。その利用目的が辞書的作用を果たすため、A-doc には多数のエン트리があると考えられる。生成者は、その多くの見出し語と対応する文書を記述しなければならない。そこで本論文では、生成者が任意に指定した、WEB 上に存在する情報の集約するページから A-doc を自動で生成することを実現した。

本稿の構成は以下の通りである。まず、2. 章で、A-doc の概要について述べる。3. 章では A-doc の記述形式について、4. 章では提案手法について説明する。5. 章では本システムの動作例を示し、最後に 6. 章で結論を述べる。

2. A-doc

ここでは、本研究の出力結果を用いる A-doc の概要について述べる。A-doc は Web における利用者主導による情報資源の結合を実現するための情報資源表現形式である。

2.1 背景

近年、検索エンジンの普及によって人々は日常的に Web を利用し情報検索を行うようになった。その上で、ユーザは検索エンジンによって必要な情報資源を検索し結果の Web 文書中から必要な情報を得る。Web 上では関連する情報はハイパーリンクで結合される。リンク元の文書とリンク先の文書間の関連

表 1 dictionary.adoc

見出し語	ドキュメント
constitution	1. 名 構成, 組織, 構造 2. 名 体質, 体格, 気質, 性格
dexterity	1. 名 器用さ 2. 名 機敏
mentality	1. 名 精神性 2. 名 考え方
...	

はホームページ作成者の意図した関係のみで提供されている。このため、ユーザが Web 文書中で見つけた単語や名称を元に新たな情報を得たいという要求が生じた場合、ユーザはさらにその単語を検索エンジンなどにかかけなければならない。

2.2 アタッチ

基本的な A-doc は見出し語とそれに対応する文書のペアの集合で、言わば辞書形式といえる。見出し語は関係データモデルのキーに相当し、文書はキーに代表されるテキスト、HTML もしくは XML 文書である。例として英和辞書に関する dictionary.adoc の概念表を表 1 に示す (実際には A-doc は XML で記述する)。

Web 利用者は閲覧中の文書にある山についての詳細な情報を得たい場合、従来では山の名前を一つずつ検索エンジンで検索せねばならない。これに対し A-doc では現在の文書に “dictionary.adoc” を結合することで、閲覧中のページにある山の名前から、それぞれの山の情報へのハイパーリンクが一斉に生成される。これは、関係データベースにおける外部キーと主キーに基く結合と本質的に等しく、この結合を “アタッチ” と呼ぶことにする。A-doc の名称における A はその形態が連想的であることを表す Associative と、その利用においてアタッチを行うことから Attachable の両義を代表している。

同じ単語をキーとした A-doc でも、例えば英和辞書の他にも英仏辞書や英露辞書まで生成者が提供すれば、利用者は必要に応じて自由に A-doc を選択し、自分の閲覧している文書にアタッチすることで状況に応じた情報を簡単に切り替えることができる。

A-doc はアタッチを行う文書に独立に存在するものであるため、誰でも “辞書” を作成・公開することができ、Web ページ作成者の意思に関係なく、ユーザ主体で動的に自由な結合を行うことができ汎用的であると言える。

アタッチには以下の三種類がある。

フルアタッチ

キーワードに合致する文書中の全ての単語に対しリンクを生成する

アンカータグによるアタッチ

Web 文書作成者がアタッチを行いたい単語に対しアンカータグを付与し、指定された単語にのみリンクを生成する

```

1  <?xml version="1.0"
2    encoding="Shift-JIS" standalone="yes" ?>
3  <adoc type="static-standard">
4    <header />
5    <body>
6      <entry>
7        <kw>constitution</kw>
8        <kw>Constitution</kw>
9        <doc>-名 構成, 組織, 構造</doc>
10       <doc>-名 体質, 性質, 気質</doc>
11       <doc>-名 憲法</doc>
12     </entry>
13     <entry>
14       ...
15     </entry>
16   </body>
17 </adoc>

```

図 1 static-standard A-doc

ユーザ指定アタッチ

Web 文書を閲覧するユーザがアタッチを行いたい単語を指定し、指定された単語が A-doc のキーワードと合致すればリンクを生成する

3. 記述方式

A-doc は XML 形式であるが、その記述方式には大きく分けて 2 種類あり、項目が実体として存在する静的 (スタティック) 方式と、項目の一部がデータベースに格納されている動的 (ダイナミック) 方式の二種類がある。更に、静的方式では標準型 (スタンダードタイプ) と分離型 (セパレートタイプ) の二つのタイプを定義する。

3.1 静的 A-doc

Static A-doc は単純に一つ以上のキーワードとそれに対応する文書のペアを一つのエントリとして記述する。キーワードを <kw> タグ、ドキュメントの HTML を <doc> タグで囲いそれら一つとして <entry> タグで囲う。各エントリの集合を <body> タグの子要素とする。

3.1.1 標準型

エントリ内の <doc> に HTML 形式のドキュメントが直接埋め込まれた形式をとる。A-doc のエントリ数は数万を超えるものもあると考えられる。エントリの HTML に共通部分が存在する場合、各エントリにそれぞれ記載するとファイルサイズも大きくなり冗長である。そこで、HTML の始めと終わりの共通部分をそれぞれ <html_preamble> と <html_trailer> に保持する。

また、<adoc> タグの type 属性の値を static-standard とする。実際の記述例を図 1 に示す。

```

1 <?xml version="1.0"
2   encoding="Shift-JIS" standalone="yes" ?>
3 <adoc type="static-standard">
4   <header />
5   <body>
6     <entry>
7       <kw>dexterity</kw>
8       <kw>Dex</kw>
9       <doc>http://www.XXX.com/AAA.html</doc>
10      <doc>http://www.YYY.co.jp/BBB.html</doc>
11    </entry>
12  </body>
13  ...
14 </entry>
15 </body>
16 </adoc>

```

図 2 static-separate A-doc

3.1.2 分離型

スタンダードタイプと違い、エントリ内の `<doc>` タグに、URI を格納する形式をとる。実際の記述例を図 2 に示す。`<adoc>` タグの `type` 属性の値を `static-separate` とする。このタイプの A-doc はある文書から他の文書へのインデックスであると見なすこともできる。本論文では WEB リンク集からの A-doc 生成を目的とするため、この型に焦点をあてる。

3.2 動的 A-doc

Static A-doc はキーワードと文書のペアを一つのエントリとしていたのに対し、Dynamic A-doc はキーワードに対する文書を持たず、キーワードの集合のみを持つ。Dynamic A-doc をアタッチした場合、文書中に A-doc のキーワードと合致する単語を見つけると、静的なページへのハイパーリンクを生成するのではなく、そのキーワードを元にデータベースから必要な情報を得て、レイアウトする動的なページへのリンクを生成する。また、キーワード集合さえも保持せず指定された単語をデータベースから検索し、その情報を動的に生成するタイプの Dynamic A-doc も考えられる。この場合、アタッチはアンカータグによるアタッチと、ユーザ指定アタッチのみが可能である。

Dynamic A-doc は静的に保持することが困難な情報、例えば気象情報や株式情報などの時系列データを提供するのに適している。

4. A-doc ファイルの自動生成

本章では、WEB 上からの A-doc ファイルの生成手法について述べる。

4.1 対象

WEB 上には様々なリンクを持つページがあるが、その中には一定の共通点を持つリンク集を内包するページが存在する。そのようなリンク集があるページの例を図 3 に示す。図 3 における一定の共通点を持つリンク集とは、シアトルマリナーズの選手名である。生成者が他のチームの A-doc を作りたいと思え

Statistics: 2007 Regular Season - Total

Batting | Pitching | Fielding

2007: Spring Training | Reg. Season
2006: Reg. Season
2005: Reg. Season
2004: Reg. Season
2003: Reg. Season
2002: Reg. Season

REGULAR | EXPANDED

Seattle Mariners Team Batting Statistics														
NAME	G	AB	R	H	2B	3B	HR	TB	RBI	BB	SO	SB	CS	BA
Chariton Jimerson	11	2	5	2	0	0	1	5	1	0	0	2	0	1.000
Vladimir Salentien	3	3	1	2	1	0	1	6	4	0	0	0	0	.667
Jeff Clement	9	16	4	6	1	0	2	13	3	3	3	0	0	.375
Jarrod Washburn	2	2	0	1	0	0	0	1	1	1	1	0	0	.500
Mike Morse	9	18	1	8	2	0	0	10	3	1	4	0	0	.444
Raul Ibanez	149	573	80	167	35	5	21	275	105	53	97	0	0	.291
Ichiro Suzuki	161	678	111	238	22	7	6	292	68	49	77	37	8	.351
Jose Guillen	153	593	84	172	28	2	23	273	99	41	118	5	1	.290
Adrian Beltre	149	595	87	164	41	2	26	287	99	38	104	14	2	.276
Jose Vidro	147	548	78	172	26	0	6	216	59	63	57	0	0	.314
Jamie Burke	50	113	19	34	8	0	1	45	12	7	17	0	1	.301
Kenji Johjima	135	485	52	138	29	0	14	218	61	15	41	0	2	.287
Ron Robinson	99	240	27	66	10	0	7	97	29	17	50	2	0	.275
Yunesky Betancourt	155	536	72	155	38	2	9	224	67	15	48	5	4	.289

図 3 Mariners.html

ば、それに対応するページを WEB 上から探し出し、対象ページとする。

```

1 /* http://www.XXX.com/ */
2 ...
3 <a href="AAA.html">ichiro</a>
4 <a href="BBB.html">johjima</a>
5 ...

```



```

1 /* mariners.adoc */
2 ...
3 <entry>
4   <kw>ichiro</kw>
5   <doc>http://www.XXX.com/AAA.html</doc>
6 </entry>
7 <entry>
8   <kw>johjima</kw>
9   <doc>http://www.XXX.com/BBB.html</doc>
10 </entry>
11 ...

```

図 4 HTML から A-doc の出力

4.2 処理

生成者が対象としたページのアンカータグのみを処理する。実際には A-doc 生成・編集ツールで URL を指定し、その HTML ソースを取得する。対象としたページには一定の共通点を持つリンク集があるので、アンカーのある単語にはリンク先にその単語を説明する文書がある可能性が高いと考えられる。そこで `<a>` タグで囲まれた単語をキーワードとし、そのリンク先を対応するドキュメントとする。処理例を図 4 に示す。生成者は対象とするページを実際にあるリンク集によって決定するので、視覚的理解可能なリンクのみを扱う。したがって、コメントアウトや JavaScript 内のリンクは扱わない。

4.3 ノイズ

HTML から A-doc を生成しただけでは、その A-doc は生成者が望むリンクの集合だけにはならないと考えられる。それは対象としたページに、“戻る”や“TOP”といったアンカーやイメージファイルのアンカー多々あるからである。これらの、生成者の意図しないエントリをノイズと定義する。本研究では、ノイズの除去は生成者がエントリを選択して行う。

4.3.1 ノイズ除去支援

対象としたページのリンク集の規模が小さい場合、ノイズの除去は容易である。しかし対象とするページの多くは、意図しないリンクが多数あると考えられる。なぜなら、広告などのリンクや対象ページのホームディレクトリ以下のリンクが多く含まれているからである。図 3 からの生成でも、全エントリ数が 386 であったのに対し、目的としたエントリは 28 であった。

対象としたページの一定の共通点を持つリンクの URI 集合は、ドメインが共通部分を含むことが多い。これは WEB ページ管理者が、そのリンク先を同じディレクトリ以下に置くためと考えられる。目的とするエントリ全てが共通するドメインを持つ場合、そのドメインを含むものだけを指定してノイズを除去することが可能である。しかし、その共通するドメインは HTML ソースを生成者が読まないと得ることはできない。より容易なノイズ除去を支援するために、意味的・構造的に共通点のあるリンク集に注目し、ドメインに共通点のあるエントリ毎に URI 集合をクラスタリングし、それぞれのクラスタごとにエントリを表示した。これにより、容易なノイズ除去が可能となった。

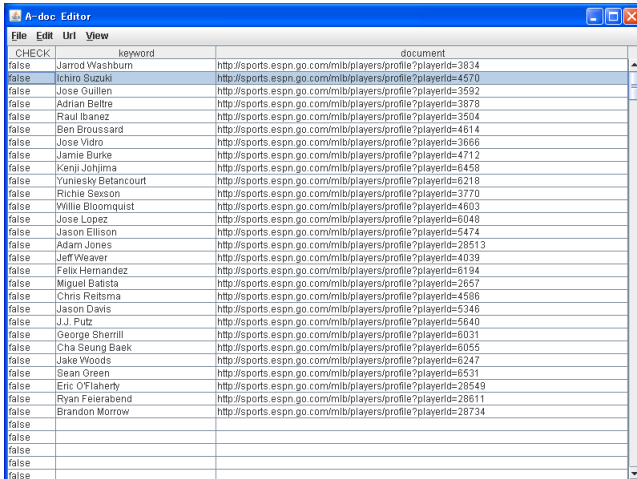


図 5 ノイズ除去

5. 実行例

生成の手順は、以下の通りである。

- (1) 対象となる WEB ページを一つ指定する
- (2) HTML を取得し、A-doc を生成する
- (3) URI に基づくクラスタリングをし、ノイズを自動除去する。
- (4) ノイズを手動で除去する

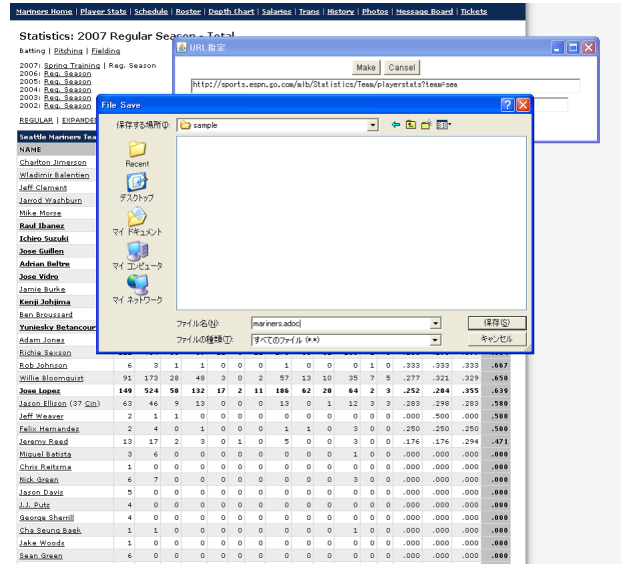


図 6 A-doc 自動生成

表 2 対象ページのノイズ

対象ページ	目的エントリ件数	ノイズ件数	全エントリ件数
ESPN	28	358	386
東急電鉄	98	55	153
価格.com	26	363	389

まず、以下の情報がそれぞれ格納された A-doc を生成することを仮定する。

- (1) シアトルマリナーズの選手
- (2) 東急電鉄の駅
- (3) ディスプレイの価格

対象ページは ESPN [3], 東急電鉄 [4], 価格.com [5] を参照した。次に、対象ページから A-doc を生成する (図 6)。さらに編集ツールを使用し、ノイズ除去をクラスタリング・手動の 2 段階で行う (図 5)。

5.1 評価

対象ページから A-doc を生成したとき、ノイズの割合は非常に大きいものとなる (表 2)。目的とするエントリ以外がノイズであるので、目的エントリ件数とノイズ件数の合計が全エントリ件数となる。

手動で数百のノイズを除去するとなると、生成者には大きな負担となる。クラスタリングにより自動的にノイズを除去した結果が表 3 である。

表 3 クラスタリングによるノイズ除去後の対象ページのノイズ

対象ページ	目的エントリ件数	ノイズ件数	全エントリ件数
ESPN	28	0	28
東急電鉄	98	24	122
価格.com	26	26	52

同じリンク先を参照しているイメージや、図 7 のようにブラウザ上のスクロール先が記述されているエントリが存在するため、クラスタリングのみによるノイズの除去は不可能である。

バス情報 (東急バスサイトへ)	http://www.tokyubus.co.jp
[あ]	http://www.tokyu.co.jp/railway/railway/train/#a
[か]	http://www.tokyu.co.jp/railway/railway/train/#ka
[さ]	http://www.tokyu.co.jp/railway/railway/train/#sa
[た]	http://www.tokyu.co.jp/railway/railway/train/#ta
[な]	http://www.tokyu.co.jp/railway/railway/train/#na
[は]	http://www.tokyu.co.jp/railway/railway/train/#ha
[ま]	http://www.tokyu.co.jp/railway/railway/train/#ma
[や]	http://www.tokyu.co.jp/railway/railway/train/#ya
[わ]	http://www.tokyu.co.jp/railway/railway/train/#wa
青葉台	http://www.tokyu.co.jp/railway/railway/train/top_aobadai.html
赤坂五丁目	http://www.tokyu.co.jp/railway/railway/train/top_akasaka.html

図 7 クラスタリングで除去出来ないノイズの例

ここで残ったノイズを最終的に手動で除去する。表 4 より、手動によるノイズの除去の負担を大幅に軽減することができると言える。

表 4 クラスタリングによるノイズ除去の割合

対象ページ	ノイズ除去
ESPN	100%
東急電鉄	56%
価格.com	92%

5.2 検討と今後の課題

5.2.1 ノイズ除去の自動化

現在の仕様では、ノイズは生成者がエンTRIESをそれぞれ選択して除去する形となっている。より容易な生成をするためには、ノイズの除去を自動化するのが望ましい。

異なる二つ以上の A-doc 間で同じキーワードのエンTRIESを削除している可能性がある。前述したように、“TOP”や“戻る”・“次へ”などは出現頻度が高いハイパーリンクだと言える。このようなキーワードリストを迷惑メールフィルタのように、予め用意して(若しくは登録し保持して)おき、生成した際、リスト上にあるキーワードとマッチするエンTRIESを削除する学習機能を実装する手法も考えられる。また逆に、多用されるキーワードに対応するドキュメントとして有害サイトが記述されている可能性もあるので、ドキュメントリストによるフィルタも必要と言える。

5.2.2 エンTRIESの多重化

A-doc は一つのエンTRIES内に複数のキーワードと、それに対応する複数の文書を格納することが可能である。しかしながら本研究ではキーワードと文書が 1 対 1 のエンTRIESしか扱えない。より柔軟な記述のために、多対多のエンTRIESを扱うことが必要である。

5.2.3 対象ページの検索

WEB ページ 1 つを指定することで、A-doc を自動生成することが可能になった。これにより生成者にかかる負担は軽減されたが、生成者が対象となるページを探さなければならない。ある一定の情報量を持つ、A-doc 化可能な WEB ページのみを出力する検索手法を提案することで、対象となるページを探す負担も軽減されると考えられる。

6. ま と め

本研究では A-doc の自動生成の手法を提案し、そのツールを

試作した。これにより生成者が WEB ページ一つを選択することで容易に A-doc を編集し、プールすることが可能となった。

文 献

- [1] 高橋 健太郎, 遠山 元道, “SuperSQL による A-doc ファイルの生成”, データ工学ワークショップ, DEWS2007, 2007
- [2] 佐藤 裕紀, 遠山 元道, “A-doc ファイルのアタッチの機能を持つ専用ブラウザの試作”, データ工学ワークショップ, DEWS2007, 2007
- [3] ESPN: <http://espn.go.com/>
- [4] 東急電鉄: <http://www.tokyu.co.jp/index.html>
- [5] 価格.com: <http://kakaku.com/>
- [6] Microsoft: Internet Expoloerer