特別研究報告

題目

文書構造を利用した電子マニュアル評価メトリクス計測ツールの試作

指導教官

井上 克郎 教授

報告者

川口 真司

平成 13 年 2 月 22 日

大阪大学 基礎工学部 情報科学科

平成 12 年度 特別研究報告

文書構造を利用した電子マニュアル評価メトリクス計測ツールの試作

川口 真司

内容梗概

CALS 等に代表される文書の電子化では，情報の共有，文書の再利用性を向上させることを目的に文書を構造化して記述することが一般的である．また，ソフトウェアが幅広い分野で利用されるようになり，高品質のマニュアルを供給することが重要となってきている．しかし，マニュアルを評価する際には，文書の持つ構造が評価対象となることは少なく，再利用性や検索性等に着目した評価を行うことは困難である．

この問題を解決するために電子マニュアルに対する文書構造評価メトリクスが提案されている [15]．本メトリクスでは，統計的手法に基づいて文書構造評価メトリクスを定義している．このため，大量のサンプルデータに対して統計的分析を行う必要があった．

本研究では，この統計的分析を行なうことを目的に，HTML マニュアルを対象にした評価メトリクス計測ツールを試作した．このツールを利用することで，HTML マニュアルの構造から算出可能な計測値を容易に計算できる．

また，ツールを利用して実際の HTML マニュアルから算出される計測値に対する分析も行った．その結果，各計測値について異常な値を示した文書は構造的な欠陥を持っていることが確認できた．

今後の課題としては, ツールを HTML 以外の構造化文書記述言語に対応させることが挙げられる．また, ツールのさらなる改良も重要である．

主な用語

構造化文書
HTML
メトリクス

目 次

1 まえがき **4**

2 文書構造評価メトリクス **6**

2.1 構造化文書 6

2.1.1 概要 6

2.1.2 構成要素 6

2.2 HTML で表現された構造化文書 7

2.2.1 モジュール 8

2.2.2 情報ブロック 8

2.2.3 階層 8

2.2.4 参照 9

2.3 8 つの評価基準 9

2.4 計測項目 9

3 電子マニュアル評価メトリクス計測ツール **11**

3.1 構成 11

3.2 木構造解析部 12

3.2.1 概要 12

3.2.2 外部仕様 12

3.2.3 内部仕様 13

3.3 メトリクス計測部 14

3.3.1 概要 14

3.3.2 外部仕様 14

3.3.3 内部仕様 15

4 評価結果とその分析 **18**

4.1 評価対象 18

4.2 分析方法 18

4.2.1 項目数の絞りこみ 18

4.2.2 異常値をしめしたファイルの抽出 20

4.3 結果 22

4.4 考察 22

5 まとめ **24**

謝辞 25

参考文献 26

付録 28

1　まえがき

近年，大量のソフトウェアが分野を問わず，開発・利用されている．ソフトウェアの開発・利用を容易に行うためにはマニュアルが必要とされるが，マニュアルは説明すべき事柄に関する知識をほとんど持たない人を対象に記述される文書であるために，その品質にはある一定以上の水準が求められている [13]．

一方，効率のよい情報の保存と配布を行うために，様々な局面で，文書を電子化する動きが急速に広がっている．このような文書の電子化においては、情報の共有や，文書の再利用性，検索性等を向上させることを目的に，文書を構造化して記述することが一般的である．

電子化された文書を利用している具体例としては，CALS[12] やグループウェア [19] などがあげられる．CALS は米国防総省が資材調達の支援システムとして開発した規格をベースとした，開発者と顧客の間で製品やサービスに関する情報を共有し，設計，製造，調達，決済をコンピュータネットワーク上で行うための標準規格である．グループウェアはコンピュータネットワークを利用して，情報の蓄積・共有化を行い，作業の効率化を推進することを目的としたシステムである．

ソフトウェアのマニュアルは，元来紙媒体で提供されていたが，CD-ROM に代表される大容量記憶媒体や，インターネットなどの普及によって，最近では電子媒体で提供されることが一般的である．よって，マニュアルの品質を評価するためには，その内容についてだけではなく，その構造が文書の電子化の目的に即したものであるかも評価する必要がある．

これまでに, 文献 [15] において, 電子マニュアルの構造の良さを定量的に評価することを目的としたメトリクスが提案されている. しかし, 本メトリクスでは，統計的手法に基づいて文書構造評価メトリクスを定義しているため，大量のサンプルデータに対して統計的分析を行う必要がある．

そこで, 本研究では，統計的分析やその後のメトリクス算出を行なうために，HTML マニュアルを対象にした評価メトリクス計測ツールを試作した．このツールを利用することで，HTML マニュアルの構造から算出可能な計測値の導出が極めて容易になる．

また，ツールを利用して実際の HTML マニュアルから算出される計測値に対する分析も行った．具体的には, まず主成分分析等の手法の結果を参考に計測項目の集約を行なった. それから, それぞれの計測項目ごとに異常な値を示した文書を抽出し, それらの特色を調査した. なお, 本研究では異常値を各項目の値の平均から標準偏差の 2 倍以上離れている値と定義した.

その結果，各計測項目について, 異常な値を示した文書は構造的な欠陥を持っていることが確認できた．しかし, 一部の項目では, その項目で異常値を示した文書を見ても構造的欠陥が見受けられないケースも存在した.

本論文では, まず 2 章で準備として提案されているメトリクスの概要について述べる. 次に 3 章で試作したツールについて述べる. 4 章では実際にツールを使用して収集したメトリクスデータを元に当該メトリクスの評価を行なう. 最後に 5 章で本論文のまとめと今後の展望について述べる.

2　文書構造評価メトリクス

本章では試作するツールが対象とするメトリクスについてその概要を説明する.

2.1　構造化文書

構造化文書とは，学術論文，マニュアルのように，意識して文書構造を作成し，それを明示した文書のことである．構造化文書を記述するためには，TeX や SGML[2] に代表される文書構造化言語を用いるのが一般的であり，それらを利用することによって構造が明確で整合性を持つ文書を記述することが容易となる．

構造化文書を用いる目的としては，文書作成のスケジュールやコストの管理が容易になる，文書作成を複数人で分担して行える，文書の品質を一定に保てるといったようなことがあげられる．

実際に構造化文書が用いられている一例としては CALS[12] がある．CALS は米国防総省が資材調達の支援システムとして開発した規格をベースとした，開発者と顧客の間で製品やサービスに関する情報を共有し，設計，製造，調達，決済をすべてコンピュータネットワーク上で行うための標準規格である．CALS においては SGML で記述された構造化文書を用いることにより企業間の情報の共有を目指している．

2.1.1　概要

参考文献 [17] における構造化文書とは，文書内容がモジュール単位で記述された文書を意味する．その模式図を図 1 に示している．

2.1.2　構成要素

モジュールとは，ユーザに対して一度に提供することが可能な情報量を表す単位であり，一つの機能，一続きの操作，または一つのテーマなどが説明できる程度の情報量である．

各モジュール内は情報ブロックと呼ばれるさらに細かな情報量の単位で構造化される．情報ブロックは，意味が伝達可能な情報量を表す単位であり，その大きさはおおよそ小見出し一つ分程度となる．

これらモジュール間の上下関係を示すものが，階層である．一般的に階層は，段階的に層をなすもの全体，またはその各層を意味する言葉であるが，文書においては，章，節，そして，項などの上下関係や，章，節，項そのものを示す．また，文書には階層関係以外のモジュール間の関係を示すものとして，参照が存在する．これは，現在読んでいる文書と関連のある文書中のある箇所を指し示すためのものである．

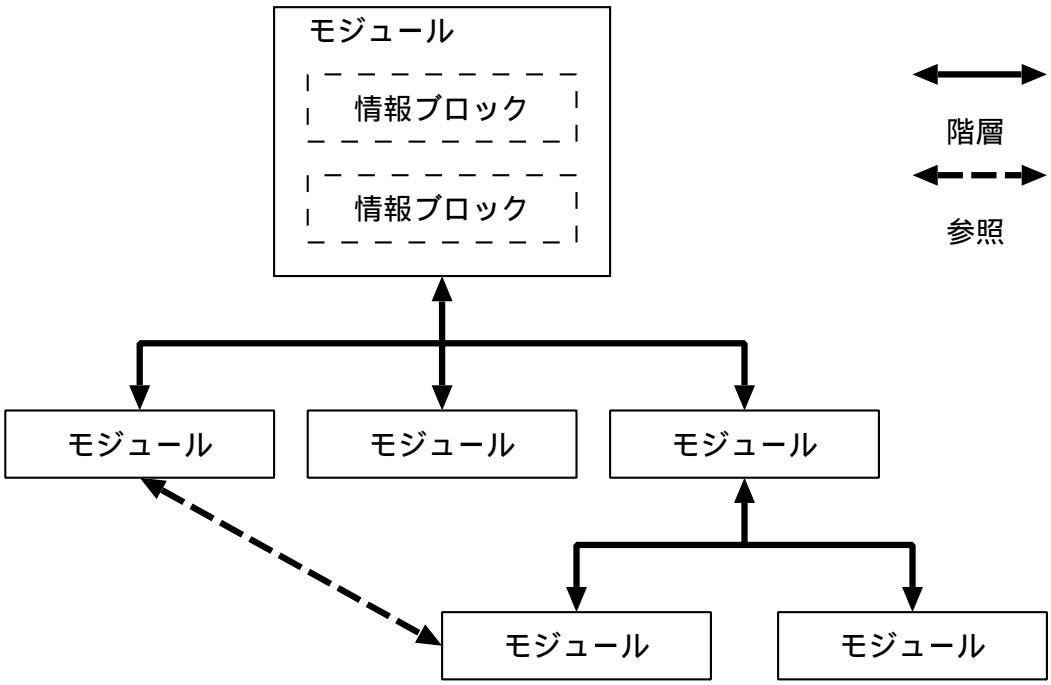

図 1: 構造化文書の模式図

2.2 HTML で表現された構造化文書

本研究では，電子マニュアルの一例として HTML で記述されたマニュアルを対象とした．HTML 文書には紙媒体の文書にはない，ファイルとリンクという概念が存在する．HTML で記述された構造化文書の模式図を図 2 に示す．

HTML で記述された構造化文書では，単一，あるいは複数のファイルから構成され，そのファイル内に各モジュールが記述される．階層は，同一ファイル内でのモジュール間，あるいは，異なるファイル間のリンクによって構成される．この階層を構成するリンクを本研究では構造リンクと呼ぶ．また，HTML 文書においては参照をリンクによって表現することが可能である．本研究では，その参照先が同一ファイル内であるかそうでないかによって，ページ内参照リンク，ページ外参照リンクと区別して考えることにした．

文書構造の評価基準を HTML 文書に対して適用するためには，文書構造が HTML 文書に対してどのようにマッピングされるかを決定する必要がある．以下，モジュール，情報ブロック，階層，参照の HTML 文書への適用について述べる．

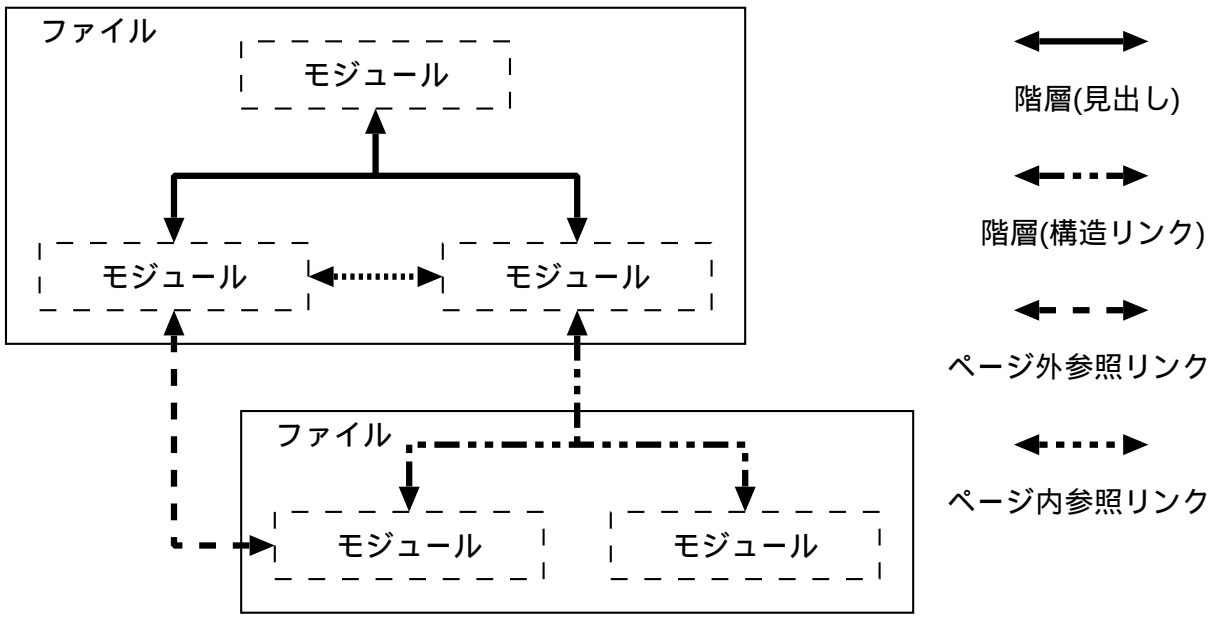

図 2: HTML で記述された構造化文書の模式図

2.2.1 モジュール

モジュールは，文書中で見出しによって分割可能な一連の情報と定義した．HTML タグには，見出しタグ ($\langle$H1$\rangle$,..., $\langle$H6$\rangle$) が存在するので，HTML 文書におけるモジュールを，文書中で見出しタグによって分割可能な一連の情報と定義する．

2.2.2 情報ブロック

モジュールの場合と同様に，HTML には段落タグ ($\langle$P$\rangle$) が用意されているため，HTML 文書における情報ブロックを，各モジュール内に含まれる段落タグにより記述された段落として定義する．

2.2.3 階層

HTML 文書における階層を次の二つのパターンで定義する．まず，一つ目は同一ファイル内に複数のモジュールが存在する場合である．この場合，階層は見出しタグ ($\langle$Hn$\rangle$:n は 1 から 6 の整数) の大小 (整数 n の大小) と定義する．一方，モジュールが複数のファイルに存在する場合は，ファイル間のリンク，即ち，前述の構造リンクを階層として定義する．

2.2.4 参照

既に述べたように，HTML 文書においては文書中に明記した参照以外にもリンクによる表現が可能である．しかし，文書中で明記された参照にもリンクを用いることが一般的であるため，HTML における参照を前述のページ内参照リンク，ページ外参照リンクとして定義する．

2.3 8 つの評価基準

文献 [15] では構造を評価する基準として次の 8 つの基準を提案している．(以下，この番号によって各基準を表す．)

1. モジュールのサイズは 1 ウィンドウ程度
2. 各モジュールのサイズは均等
3. モジュールは複数の情報ブロックから構成
4. モジュールが構成する階層は基本的に 3 階層にする
5. 各モジュールの子供は適切な数にする
6. 1 ファイルに記述されるのは 1 モジュール
7. 1 モジュールにつき 1 つのページ内参照リンク
8. 関連のあるモジュール間でのページ外参照リンク

2.4 計測項目

まず，文書構造から算出可能であると考えられる計測値を導出し，それらと各評価基準との関連を考察した．その変数を以下に示す．() 内は，その計測値が関連すると考えられる評価基準を示す．

- ファイル数 (6)
- 階層の深さ (4)
- 構造リンク数 (4，5)
- ページ内参照リンク数 (7)
- ページ外参照リンク数 (8)

- 外れファイル数:構造リンク (4，5)
- 外れファイル数:ページ内参照リンク (7)
- 外れファイル数:ページ外参照リンク (8)
- 文字数/モジュール (1，2)
- 情報ブロック数/モジュール (1，2，3)
- 子供の数/モジュール (5)
- 文字数/ファイル (7，8)
- モジュール数/ファイルの (6)

各リンク数は 1000 文字単位でその種類のリンクがいくつあるかを示している．また，外れファイル数は，各要素の平均から標準偏差が 2 倍以上離れているファイルの総数である．文字数/モジュール，情報ブロック数/モジュール，子供の数/モジュール，文字数/ファイル (7，8)，モジュール数/ファイルの (6) については，最大値，最小値，平均，分散，標準偏差，尖度，変動係数についてそれぞれ計測する．

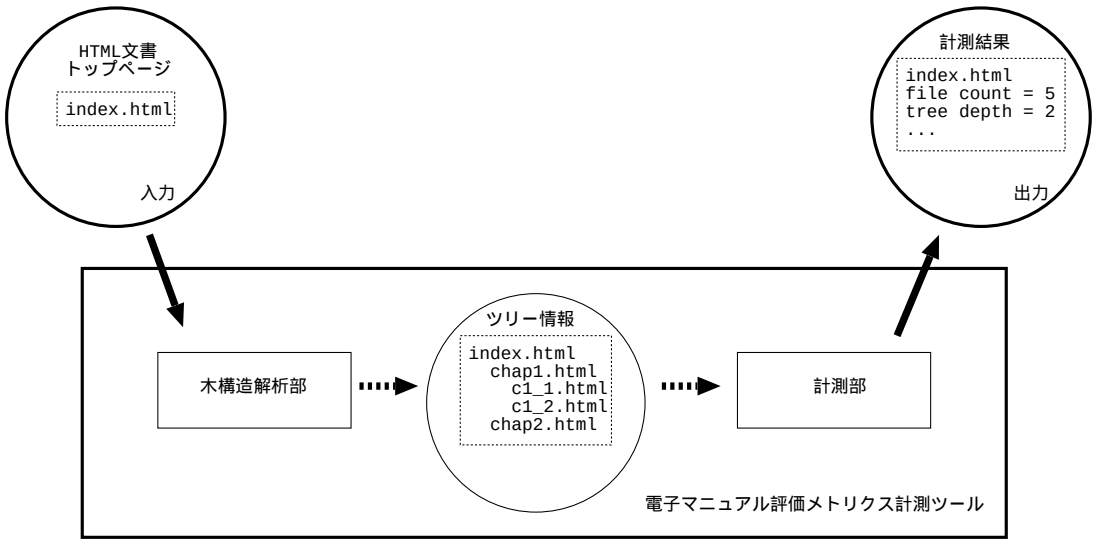

図 3: ツール構成

3 電子マニュアル評価メトリクス計測ツール

本章では, 試作したメトリクス計測ツールについて説明する. 本ツールは Perl [11] で記述されているため Perl 処理系さえ存在すれば, どのようなプラットフォームであっても動作可能である. これまでに, FreeBSD, Linux, Windows で動作することを確認している.

3.1 構成

図 3 に本ツールの構成を示す. この図からわかる通り, 本ツールは木構造解析部と計測部の二つから構成される. 木構造解析部は指定された HTML 文書の構造を解析を行ない, 計測部は抽出された木構造の結果を元にメトリクスを出力する部分である.

木構造解析部ではモジュール間の階層関係ではなく, HTML ファイル単位での階層関係の解析を行なう. なぜファイル単位の階層関係を調べる必要があるのか. それは, 構造リンクとページ外参照リンクの判定のためである.

<A> タグで表されるリンクは, 三つに分類される. すなわち 1. 構造リンク 2. ページ外参照リンク 3. ページ内参照リンク の三つである. このうち 3. ページ内参照リンクは単純に「同一ファイル内へのリンク」であるので判別は用意である. しかし, 構造リンクとページ外参照リンクは単純に区別することができない. この区別を行なうには, リンク先のファイルがリンク元の子ファイルであるかどうかの情報が必要であるが, その情報は <A> タグのなかには存在しない[1].

[1] 実際には <A> タグに REL 属性などのリンク元とリンク先の関係を記述する属性が用意されているので,

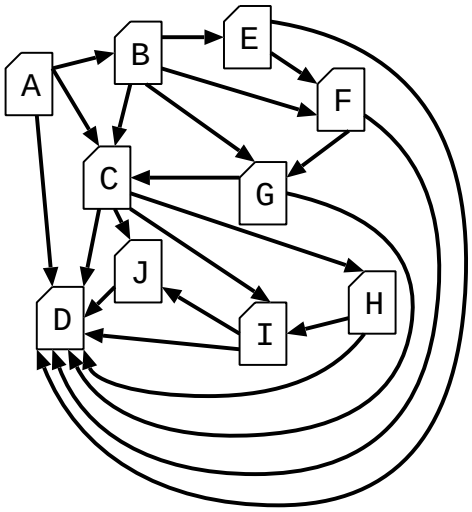

図 4: HTML 文書の例

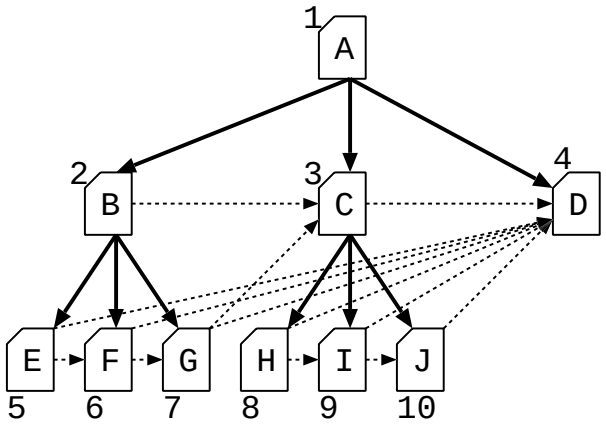

図 5: 解析結果

そこで, まず木構造解析部で HTML マニュアルを構成しているファイルの関係を解析し, HTML ファイルをノードとする木構造を出力する. この情報を使えば, リンクが構造リンクであるのか, ページ外参照リンクであるのかの区別をつけることが可能である.

3.2 木構造解析部

3.2.1 概要

木構造解析部では, 入力として HTML 文書のトップページの URL を受けとる. そして, URL で指定されたファイルをトップページとする HTML マニュアルの取得と, その木構造の解析を行なう. そして, HTML マニュアルに含まれるすべてのファイル名をそのツリー構造とともに出力する.

本ツールでは, トップページで指定される一つの HTML マニュアルを次のように定義した.

以下の条件を満たす全てのファイル (ファイルの種類は問わない)

1. トップページからリンクによってたどれること. かつ,
2. トップページと同じ, もしくは下位のパスをもっていること.

3.2.2 外部仕様

起動方法

この記述は正確ではない. しかし, 実際にはこの属性は全くといっていいほど使われていない.

```
%genHTMLtree filename
```

filename には HTML マニュアルのトップページを指定する.

出力

filename を根とする HTML マニュアルに含まれる全てのファイルを, その木構造と共に出力する. 例として, 図 5 のような HTML 文書が与えられたときの出力は次のようになる.

```
~/testdata/a.html
        ~/testdata/b.html
                ~/testdata/e.html
                ~/testdata/f.html
                ~/testdata/g.html
        ~/testdata/c.html
                ~/testdata/h.html
                ~/testdata/i.html
                ~/testdata/j.html
        ~/testdata/d.html
```

3.2.3 内部仕様

木構造の解析方法

ここで本ツールが採用している木構造解析アルゴリズムを説明する.

本ツールは, まず始めにトップページから全てのリンクを拾いだす. そして, 次の要領でリンクの種別を判定する.

1. リンクが同一ファイル内へのリンクなら, ページ内参照リンク
2. リンク先ファイルが
   - すでにどこか別のファイルの子であることが確定している
   - 走査中のファイルよりも上位ディレクトリや外部のサイトに存在するファイル

   このどちらかにあてはまる場合は, ページ外参照リンク
3. そうでなければ, リンクを構造リンクとみなし, リンク先を子ファイルとする. リンク先を “すでにどこかの子であることが確定しているファイルリスト” に保存する.

そして, 新しく見つけた子ファイルについても同様の手続きを再帰的に適用していく. こうして, トップページからリンクによってたどることが可能でかつ, 同じもしくは下のディレクトリに存在するファイル全てを得ることができる.

3.3 メトリクス計測部

3.3.1 概要

メトリクス計測部は前述の木構造解析部の出力を受けとり, HTML 文書の各メトリクスを計測する. ここでの仕事は主に各ファイルのメトリクスを計測することにある.

3.3.2 外部仕様

起動方法

```
%genMetric [-ot csvfile] [treefile]
```

`-ot csvfile` をつけると, 標準出力への結果出力とは別に csv 形式での出力も行なう. csv 形式での出力は csvfile で指定されたファイルに出力される.

入力は `treefile` で指定する. treefile が省略された場合は標準入力から読み込みを行なう. 入力の形式は, 先ほどの木構造解析ツールが出力する形式と同様のものを期待する. それ以外の文字列が入力された場合の動作は保証しない.

出力

計測部は, トップページのファイル名と各種計測項目の結果を出力する. それに続いて, ファイル単位での解析結果を出力する. ファイル単位での解析結果には, リンクの数やモジュールごとの情報などが含まれる.

計測ツールの出力例を以下に示す.

```
~/tmp/www.iph.pref.hyogo.jp/users/info/dokugeki/kensahou.htm
the number of file = 11
tree height = 3
structure link = 1.33067198935462
inside link = 0.332667997338656
reference link = 3.2435129740519
odd file of structure link = 1
odd file of inside link = 0
odd file of reference link = 1
                       max     min     avg.    sg.     sd.     rsg.    kw      cv
module size            8.7e+02 4       1.9e+02 3.7e+04 1.9e+02 0.049   7.9e+18 1
module block           8       0       0.76    2.3     1.5     0.036   3e+02   2
module child count     22      1       5       50      7       0.1     2.5e+07 1.4
file size              2.2e+03 5.3e+02 1.1e+03 2.6e+05 5.1e+02 0.055   1.3e+22 0.46
module count per file  14      1       5.7     15      3.9     0.078   1.4e+05 0.68

~/tmp/www.iph.pref.hyogo.jp/users/info/dokugeki/kensahou.htm
Link = 5, 1, 1. Tree height = 2. Chars of Code = 3362. True Chars of Code = 1020.
H 3 Height 3 DivNum 0 Size 293 BlockNum 2 ChildNum 19 Depth 0
H 3 Height 0 DivNum 0 Size 433 BlockNum 0 ChildNum 0 Depth 0
H 3 Height 1 DivNum 0 Size 227 BlockNum 0 ChildNum 2 Depth 0
H 5 Height 0 DivNum 0 Size 5 BlockNum 0 ChildNum 0 Depth 1
H 5 Height 0 DivNum 0 Size 62 BlockNum 0 ChildNum 0 Depth 1

        ~/tmp/www.iph.pref.hyogo.jp/users/info/dokugeki/index.htm
        Link = 11, 1, 15. Tree height = 1. Chars of Code = 5650. True Chars of Code = 724.
        H 3 Height 2 DivNum 0 Size 657 BlockNum 8 ChildNum 22 Depth 1
        H 5 Height 0 DivNum 0 Size 5 BlockNum 0 ChildNum 0 Depth 2
        H 5 Height 0 DivNum 0 Size 62 BlockNum 0 ChildNum 0 Depth 2

...(後略)
```

3.3.3 内部仕様

メトリクス計測部の大まかな動作は以下の通り.

1. 入力を解析し HTML ファイル間の木構造を復元する.
2. トップページを起点として, 以下の動作を行なう.
   (a) モジュールを切り分けつつ, リンクをたどりモジュール単位での木構造の把握を行なう.
   (b) 上の結果を元に各種計測項目の算出を行なう.
3. 蓄積された情報を整理して出力する.

モジュール

モジュールとは, <H1>,<H2>,<H3>,<H4>,<H5>,<H6> タグ (以下, <Hn> タグと呼称) によって区切られる, HTML 文書の一部である. ファイル中に <Hn> タグが存在しない場合はファイル全体を一つのモジュールと見做す.

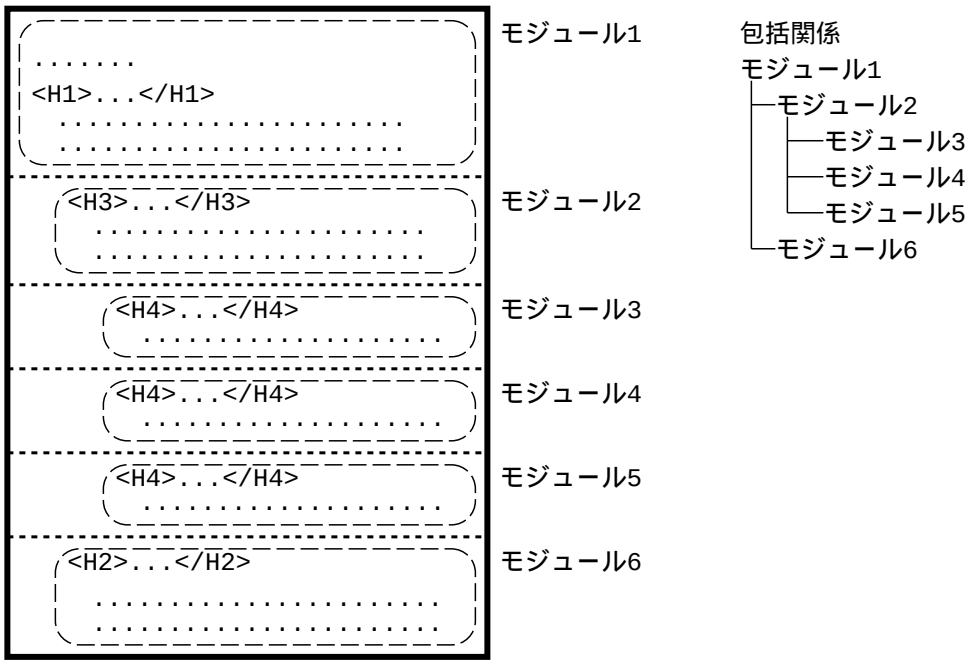

図 6: モジュールの切り分け方

図 6 のように同一ファイルに複数の <Hn> タグがある場合, それらの間には上下関係が存在する. 例えば, 図 6 では <H1> タグで始まるモジュール (以後 <H1> モジュールと略記) の次に <H3> モジュールが続いている. このとき, 後者は前者の子モジュールである. また, 同じ数字のタグで始まるモジュールが並んでいる場合 (モジュール 3, 4, 5) , それらは兄弟であるとして, 同じ親を持つものとする.

その次に, <H2> モジュールが続いている. これをモジュール 2 の子としてしまうと, <H3> モジュールの下に <H2> モジュールがつくことになり, 不自然である. これは <H1> モジュールの子とするのが当然であろう.

このような上下関係を判定するアルゴリズムは次の通り.

1. スタックを用意.
2. スタックトップのモジュールが現在のモジュールよりも上位ならば, 現モジュールは上のモジュールの子となる.

3. そうでなければ, スタックを一つ登る. 以後くりかえし. スタックが空になってしまったら, そのモジュールはファイルのトップレベルのモジュールである.

情報ブロック

<P> タグで表される情報ブロックをカウントする際には, 単純に “<P>” という文字列を検索して, その出現回数をカウントしている. 終了タグとの対応等は一切考慮していない. これは <P> タグは終了タグが省略可能な仕様になっているからである.

4　評価結果とその分析

本章では実際にツールを用いたデータ計測とその分析について述べる.

4.1　評価対象

評価対象となるデータは, 無作為に収集した 142 個の HTML マニュアルである. これらのマニュアルは筆者らが検索サイトを用いて適当に集めたものである. 付録にこれらの文書の一覧を掲載する.

4.2　分析方法

これらのデータに対して 3 章で述べたツールを適用し, csv 形式の出力をつなぎ合わせて一つの表に編成した. このままでは項目数が多すぎて分析が困難なので, まず項目数の絞りこみを行なった.

4.2.1　項目数の絞りこみ

まず最初に尖度, 分散の 2 つの統計量を排除した. 尖度は分散が 0 の場合に無限大になってしまうという問題点があった. 分散は, 単位が元データの単位の 2 乗になっており散らばり具合を判定する基準として問題がある.

次に, 主成分分析 [8] の結果を参考にして無意味な項目を削除した. 主成分分析の結果を表 1 に示す.

主成分 1 では文字数/モジュールの最大, 最小, 平均, ブロック/モジュールの最大, 最小, 平均が高い係数を持っている. また, 主成分 2 ではファイル数, 外れファイル数:構造リンク, 外れファイル数:ページ内参照リンク, 外れファイル数: ページ外参照リンクが高い係数を持っている.

1 の結果から, 全ての項目に置いて最大値, 最小値をとらず平均値のみを指標としても問題ないことがわかる. また, 外れファイル数もファイル数との相関が極めて高かったため, 分析項目から除外する.

本来は, 項目を除外したのち再び主成分分析を行ない, なんらかの意味のある主成分をもとに分析を進めていくべきである. しかし, 無意味な項目を除外してもう一度主成分分析を行なってみても, 相関の低い項目が無理に結び付いていたり, おなじ統計量どうしが一つの項目にまとまる (標準偏差が別の項目の標準偏差とまとまったりする) ばかりで有意な主成分を得られなかった.

そこで, 抽出された主成分は使用せずに, 省略する項目を選び出すための参考として主成分分析の結果を利用するにとどめた.

表 1: 主成分分析の回転後の結果

	主成分 1	主成分 2	主成分 3	主成分 4	主成分 5	主成分 6	主成分 7	主成分 8	主成分 9
ファイル数	1.631E-02	-2.742E-02	.955	.158	-3.599E-02	5.494E-02	6.179E-02	3.720E-02	3.453E-02
高さ	-6.520E-02	-1.491E-02	.126	.287	.312	.552	-.309	.104	7.087E-02
構造リンク	-5.867E-02	-.112	.114	.211	-4.272E-02	2.016E-02	-4.432E-02	.823	-.166
ページ内参照リンク	7.964E-02	.291	-8.967E-02	-.161	-2.815E-02	-7.582E-02	4.922E-02	-.319	-.134
ページ外参照リンク	-4.028E-02	-.116	.304	.638	-5.777E-02	-7.320E-02	.124	.394	-8.483E-02
外れ:構造リンク	8.236E-03	-2.883E-02	.979	2.382E-02	-2.436E-02	8.921E-02	-2.931E-02	4.587E-02	1.611E-02
外れ:ページ内参照リンク	8.279E-03	-4.867E-02	.808	.135	5.942E-02	8.373E-02	9.079E-02	.265	7.955E-02
外れ:ページ外参照リンク	5.183E-03	-2.696E-02	.914	9.577E-02	-2.146E-02	.127	-5.116E-02	-7.717E-02	-6.853E-02
文字数/モジュール:Max	.857	.240	.165	3.699E-02	2.811E-02	4.751E-02	4.631E-02	-3.939E-02	.372
文字数/モジュール:Min	.978	4.262E-02	-2.916E-02	-3.713E-02	-2.746E-02	-.110	-4.592E-02	-5.578E-02	-5.490E-02
文字数/モジュール:Avg	.974	4.768E-02	-2.730E-02	-4.129E-02	-3.471E-02	-.114	-4.113E-02	-7.065E-02	6.043E-02
文字数/モジュール:SD	5.135E-03	8.155E-02	1.646E-02	-1.705E-02	-2.127E-02	7.950E-02	6.088E-02	-7.772E-02	.898
文字数/モジュール:cv	-8.787E-02	7.331E-02	6.283E-02	.170	.162	.806	7.842E-04	6.192E-02	.227
ブロック/モジュール:Max	.908	4.552E-02	4.734E-02	-2.156E-02	-7.006E-02	.204	.200	.150	-5.888E-02
ブロック/モジュール:Min	.981	4.239E-02	-2.937E-02	-2.911E-02	-2.157E-02	-8.471E-02	-4.068E-02	-5.140E-02	-6.781E-02
ブロック/モジュール:Avg	.983	4.380E-02	-3.051E-02	-3.102E-02	-2.618E-02	-7.083E-02	-2.430E-02	-3.474E-02	-6.436E-02
ブロック/モジュール:SD	6.938E-02	2.872E-03	-4.108E-02	-8.318E-02	-.119	.562	.449	.419	3.840E-02
ブロック/モジュール:cv	-5.789E-02	-1.020E-02	.225	3.637E-02	.103	.797	1.653E-02	-.141	-.175
子/モジュール:Max	-1.265E-02	9.847E-03	.150	.882	7.774E-02	.131	.345	2.063E-02	4.071E-02
子/モジュール:Min	8.153E-03	-3.531E-02	1.468E-02	-1.780E-02	5.653E-03	-5.616E-03	.906	-3.152E-02	5.410E-02
子/モジュール:Avg	-5.380E-03	-5.848E-03	6.993E-02	.386	-5.344E-02	-2.220E-02	.895	-3.584E-04	1.843E-02
子/モジュール:SD	-1.290E-02	1.367E-02	8.009E-02	.880	-7.356E-02	2.792E-02	.116	5.381E-02	-1.725E-02
子/モジュール:cv	-8.038E-02	2.213E-02	5.600E-02	.684	.237	.383	-.301	.168	5.067E-02
文字数/File:Max	.217	.925	3.798E-02	4.991E-03	.141	1.723E-02	2.128E-03	3.592E-02	.182
文字数/File:Min	.219	.944	-2.241E-02	-1.960E-02	-8.707E-02	-3.994E-02	-2.104E-02	1.922E-02	.106
文字数/File:Avg	.217	.947	-2.415E-02	-2.239E-02	-3.935E-02	-3.502E-02	-2.063E-02	1.678E-02	.110
文字数/File:SD	-3.037E-02	-2.351E-02	-2.293E-02	-5.138E-02	.914	.145	3.538E-02	-7.854E-03	.149
文字数/File:cv	-5.377E-03	-.168	.199	.240	.468	.482	.102	.394	.275
モジュール/File:Max	-9.764E-02	.689	9.144E-03	.210	.568	.112	-7.127E-02	-.132	-.144
モジュール/File:Min	-9.591E-02	.945	-5.337E-02	-3.452E-02	-.154	2.085E-02	4.384E-03	-.135	-5.763E-02
モジュール/File:Avg	-.109	.946	-7.070E-02	-5.293E-02	3.004E-02	3.981E-02	-5.687E-03	-.157	-9.784E-02
モジュール/File:SD	-5.056E-02	1.596E-02	-6.275E-02	4.224E-02	.919	8.338E-02	-5.251E-02	-5.704E-02	-.162
モジュール/File:cv	-3.283E-02	-9.447E-02	4.608E-02	.584	.531	.230	-.196	8.483E-02	-8.363E-03
固有値	6.642	5.436	4.787	3.127	2.521	1.941	1.542	1.178	1.009
寄与率 [%]	20.126	16.474	14.505	9.475	7.640	5.880	4.673	3.569	3.057
累積寄与率 [%]	20.126	36.600	51.105	60.579	68.220	74.100	78.773	82.342	85.399

最終的には, 以下の項目について分析を行なった.

1. ファイル数
2. 階層の深さ
3. 構造リンク数
4. ページ内参照リンク数
5. ページ外参照リンク数
6. 文字数/モジュール
7. 情報ブロックの数/モジュール
8. 子供の数/モジュール
9. 文字数/ファイル:平均
10. モジュール数/ファイル:平均

文字数/モジュール，情報ブロック数/モジュール，子供の数/モジュールについては，平均，標準偏差，変動係数についてそれぞれ計測する．

4.2.2 異常値をしめしたファイルの抽出

次に, 各計測項目ごとに, 異常な値を示した文書を抽出しその性質を調べた. 異常な値とみなす基準として, 本研究では

$$x < \text{平均値} - 2 * \text{標準偏差}\ or\ \text{平均値} + 2 * \text{標準偏差} < x$$

という基準を用いた. これは, それなりの品質の文書ならば全体の平均からそれほど離れた値を示すことはないだろう, という仮定に基づいている. ファイル数など一部の項目を除いた計測項目に関してはこの仮定は妥当なものと考える.

範囲を標準偏差の二倍としたのは, 一般にさまざまな検定をおこなうときに有意水準を 5% としている場合, 分散の 1.96 倍離れた点を閾値とするからである.

表 2: 検出されたマニュアルの個数

項目名	検出数	平均	標準偏差
ファイル数	3	55.52816901	191.2525601
木の深さ	5	3.218309859	2.312826392
構造リンク数	4	0.795467616	0.937877775
ページ内参照リンク数	9	1.546763111	2.583110566
ページ外参照リンク数	11	4.539997172	5.865490262
文字数/モジュール-平均	2	3049.499294	17241.46901
文字数/モジュール-標準偏差	3	1189.184811	2584.57701
文字数/モジュール-変動係数	3	1.13093818	0.737258368
ブロック数/モジュール-平均	2	27.71743483	209.9704552
ブロック数/モジュール-標準偏差	7	7.413351076	16.4234133
ブロック数/モジュール-変動係数	4	1.311888054	1.228550273
子供の数/モジュール-平均	4	9.980392604	20.86888555
子供の数/モジュール-標準偏差	7	8.382458725	15.30438072
子供の数/モジュール-変動係数	6	0.879385607	0.714265478
文字数/ファイル-平均	3	13508.47483	62830.35277
モジュール数/ファイル-平均	3	12.56327214	29.2199919

4.3 結果

各項目の平均, 標準偏差, そして異常な値をとったマニュアルの数を表 2 に示す.

このように様々なマニュアルが異常な値を示すものとして検出された. 検出されたマニュアルの累計は 76 個となる. が, このなかには重複がかなりふくまれており, 重複を取り除いた個数 48 である.

それでは, 上で述べた計測項目において異常値を示したマニュアルのなかでも特徴的なものについて, いくつか説明を行なう.

group.lin.go.jp/souchi/honmokuji.htm

表 3: group.lin.go.jp/souchi/honmokuji.htm の計測結果

ファイル数	86	ブロック数/モジュール-平均	0.758928571
木の深さ	15	ブロック数/モジュール-標準偏差	2.455238092
構造リンク数	0.469186988	ブロック数/モジュール-変動係数	3.235137251
ページ内参照リンク数	0	子供の数/モジュール-平均	10.46875
ページ外参照リンク数	0.502919386	子供の数/モジュール-標準偏差	28.75215685
文字数/モジュール-平均	970.5327381	子供の数/モジュール-変動係数	2.746474684
文字数/モジュール-標準偏差	2952.220467	文字数/ファイル-平均	3791.813953
文字数/モジュール-変動係数	3.041855623	モジュール数/ファイル-平均	3.906976744

表 3 にこのマニュアルの計測結果を示す. この HTML マニュアルは「高さ」において異常値を示していた. このマニュアルを詳細に調べてみたところ, リンク構造が木構造ではなく直線的につながっている部分を発見した.

www.sra.co.jp/wingnut/gcc/gcc-j.html

表 4 がこのマニュアルの計測結果である. このマニュアルは. ファイル数は一つだったが, サイズが 1.5MB もあった. これは明らかに複数のファイルに分割するべきであると思われる.

4.4 考察

「階層の深さ」や「モジュールサイズ:平均」などほとんどの計測項目において, 構造に問題のあるマニュアルが異常な値を示していた.

これらのことから, 上記計測項目が構造に欠陥のあるマニュアルの検出に一定の効力を持つものと考えられる.

表 4: www.sra.co.jp/wingnut/gcc/gcc-j.html の計測結果

ファイル数	1	ブロック数/モジュール-平均	10.65909091
木の深さ	3	ブロック数/モジュール-標準偏差	11.00022696
構造リンク数	0.004275526	ブロック数/モジュール-変動係数	1.032004235
ページ内参照リンク数	5.07790009	子供の数/モジュール-平均	8.888888889
ページ外参照リンク数	0.001425175	子供の数/モジュール-標準偏差	8.912468725
文字数/モジュール-平均	2657.833333	子供の数/モジュール-変動係数	1.002652732
文字数/モジュール-標準偏差	4680.888467	文字数/ファイル-平均	701668
文字数/モジュール-変動係数	1.761167041	モジュール数/ファイル-平均	264

しかし,「ファイル数」「ページ外参照リンク」では必ずしも問題があるとはいえないマニュアルが異常値をしめした. これらの項目については最初の仮定 (平均よりかけはなれた値を持つマニュアルは構成上問題がある) が誤っていたと判断せざるを得ない.

5 まとめ

試作したツールにより各種計測値をほぼ自動的に算出することが可能になった. そして, 実際に多くの数の HTML マニュアルに対して本ツールを適用し計測された値を元に考察を行なった. その結果, 一部項目を除いて, 各計測値において平均から極端に離れた値を示した文書はなんらかの構造的欠陥をかかえていることを確認した.

今後の課題としては, 木構造の解析をより正確におこない, 可能ならば計測を完全に自動化できるようにすることがある.

また, HTML 以外の構造化文書フォーマットへの対応も望まれる. 特に, XML では書き手が自由にタグを定義できるため, XSLT などの変換技術を用いれば, 書き手が想定した構成を忠実に再現する形でタグを変換できる. したがって, 正確な文書要素の把握が可能になることが期待できる.

謝辞

本論文を作成するにあたり，常に適切な御指導を賜わりました大阪大学大学院基礎工学研究科情報数理系専攻 井上 克郎 教授に心より深く感謝致します．

本論文の作成において，適切な御指導および御助言を頂きました大阪大学大学院基礎工学研究科情報数理系専攻 楠本 真二 助教授に深く感謝致します．

本論文の作成において，適切な御指導および御助言を頂きました大阪大学大学院基礎工学研究科情報数理系専攻 松下 誠 助手に深く感謝致します．

本論文の作成において，多大な御指導および御助言を頂きました大阪大学大学院基礎工学研究科情報数理系専攻 谷口 真也 氏に深く感謝致します．

最後に，その他様々な御指導，御助言等を頂いた大阪大学大学院基礎工学研究科情報数理系専攻井上研究室の皆様に深く感謝致します．

参考文献

[1] Jacques Andre, Richard Furuta, and Vincent Quint, editors : “Structured Documnts”, Cambridge University Press, (1989).

[2] Martin Bryan: “SGML 入門”, 山崎 俊一監訳, アスキー, (1993).

[3] Fred Cole and Heather Brown : “Editing structured documents–problems and solutions”, Electronic Publishing: Origination, Dissemination, and Design, 5(4):209-216, (1992).

[4] Richard Furuta: “Defining and Using Structure in Digital Documents” , Digital Library ’94, (1994).

[5] Robert E. Horn: “ハイパーテキスト情報整理学実践編”, 松原光治監訳, 日経 BP 社, (1995).

[6] Robert E. Horn: “ハイパーテキスト情報整理学”, 松原光治監訳, 日経 BP 社, (1991).

[7] Robert E. Horn: “Mapping Hypertext: Analysis, Linkage, and Display of Knowledge for the Next Generation ofOn-Line Text and Graphics”, The Lexington Institute, (1989)

[8] Richard A. Johnson, Dean W. Wichern: “多変量解析の徹底研究”, 西田俊夫訳, 現代数学社, (1992)

[9] 小林敦: “マニュアル作成の構造化手法”, 日経マグロウヒル社, (1988)

[10] iller, G.A.: “The Magical Number Seven Plus or Minus Two: Some Limits on Our Capacity for Processing Infomation”, Psychological Review, vol.63 pp.81-97, (1956)

[11] Larry Wall, Tom Christiansen, Randal L. Schwartz: “プログラミング Perl 改訂版”, 近藤嘉雪訳, オライリー・ジャパン, (1997)

[12] 水田 浩: “CALS の実践”, 共立出版株式会社, (1997).

[13] ニューメディア開発協会: “電子マニュアル評価ガイドラインの適正標準化に関する調査研究”, http://www.jtca.org/empg/report98/index.html, (1999)

[14] 高橋善文， 牛島和夫: “計算機マニュアルのわかりやすさの定量的評価方法”，情報処理学会論文誌, vol.32, NO.4, pp.460-469, (1992)

[15] 谷口 真也: “電子マニュアルの構造を利用した文書評価メトリクス”, 大阪大学大学院基礎工学研究科修士学位論文,2001

[16] 谷口真也，川口真司，松下誠，井上克郎: “電子マニュアルの文書構造に対する評価メトリクス”, 電子情報通信学会ソフトウェアサイエンス研究会 (2001)．

[17] 横河電気 CyberDoc プロジェクト: “デジタル時代のドキュメント企画と設計”，日本理工出版会, (2000).

[18] XML/SGML サロン: “標準 XML 完全解説”, 技術評論社, (1998).

[19] “特集 グループウェアの実現に向けて”, 情報処理, Vol. 34, No. 8 (1993).

[20] “Extensible Markup Language(XML)”, http://www.w3c.org/XML/

付録

分析のさいに用いた HTML マニュアル

www.libweb.tdc.ac.jp/pubmed/index.html
www.ballet.or.jp/%7Ekbhospis/oyakudati/ganmokuji.html
www.iph.pref.hyogo.jp/users/info/dokugeki/kensahou.htm
www.telesa.or.jp/html/990426.htm
member.nifty.ne.jp/tsato/xfig/html/japanese/tutorial.html
www.kprolog.com/doc/ja/index.html
www.cis1.c.dendai.ac.jp/Mathematica/index.html
www.med.nagoya-u.ac.jp/anesth/assorted/menu.htm
japache.infoscience.co.jp/japanese_1_3/manual/index.html
cclub-flying.dsl.gr.jp/products/ecoact/manual/000.htm
www.hitachi.co.jp/Prod/comp/soft1/pfxml/html/manual/hpf_0001.htm
www.hanelab.k.dendai.ac.jp/tech/tex_begin/html-index.html
www.miyazaki-med.ac.jp/PG-manual/PG00.html
www.linux.or.jp/JF/JFdocs/gnuplot.html
makioka.y-min.or.jp/InfCtlManual.html
www.scl.kyoto-u.ac.jp/scl/appli/appli_manual/rasmol_manual/rasmol/rasmol.html
www.annie.ne.jp/%7Ehachi/frontier/default.html
www.mid.comp.nec.co.jp/SUPPORT/48MANU/guide/guide.html
kappalab.on.pascal.co.jp/Clientele/PlatoGrande/imageviewer.html
www.kashodo.co.jp/products/tabloid/help/index.html
www.dtc.co.jp/Ascend/P50_Setup.html
www.maff.go.jp/work/combime/INDEX.html
www.tru64unix.compaq.com/faqs/publications/i18n/japan/V40E/DOCUMENTATION/HTML/IMLIB/pu8tc.htm
javacenter.sun.co.jp/pnuts/doc/ja/guide.html
www.ae.wakwak.com/%7Esysportcore/sysport/kb/webapp/Html401Manual/index.htm
www.annie.ne.jp/%7Ehachi/frontier/default.html
www11.freeweb.ne.jp/computer/manula/users_manual.html
www2.biglobe.ne.jp/%7Enir/npc/index.html
www.topstudio.co.jp/%7Ekmuto/html/index.html
www5a.biglobe.ne.jp/%7Eaihana/docs/gconf/index.html
www.ceres.dti.ne.jp/%7Ey-endo/slink/i386/index.ja.html
www.big.or.jp/%7Emio/ga/ht/htman/ht_man.htm
www.ueda.info.waseda.ac.jp/%7Esatoshi/padsty/guide-j.html
www.dna.affrc.go.jp/htdocs/Eudora/index.html
www.iizuka.edu.isc.kyutech.ac.jp/html1/index.html
osa.cup.com/doc/perl/SWIX/doc/html/swix.html
www.interq.or.jp/earth/inachi/openldap/admin/index-ja.html
www.vinelinux.org/sparc/install-vine-sparc-2.1.html
www.amulet.co.jp/caldera/colman/index.html
www.mpt.go.jp/policyreports/japanese/misc/NetWork-Manual/index.html
caml.inria.fr/ocaml/htmlman/index.html
dc.to/RbJ/LRJ/index.html
lettuce.ms.u-tokyo.ac.jp/%7Ekato/NAMAZU/manual.html
www02.so-net.ne.jp/%7Ekomuro/Mutsumi/DylanInfo/index.html
dog.intcul.tohoku.ac.jp/w3m/MANUAL.html
www.jca.ax.apc.org/fujimi-net/ktbusr.html
www.post.yusei.go.jp/newnumber/manual/htm_manu/index.htm
www.ballet.or.jp/%7Ekbhospis/oyakudati/volunteer.html
home3.highway.ne.jp/nitop/yk/naiki.htm

home3.highway.ne.jp/nitop/yk/huku0.htm
www.ops.dti.ne.jp/%7Eglass/tousisintaku.htm
www.chest-disease.com/accident.html
www.ncc.go.jp/ibaraki/general/renkmanu.html
www.infoscience.co.jp/technical/proftpd/reference.html
www02.so-net.ne.jp/%7Etakizawa/index-j.html
www.samba.gr.jp/project/samba-ja/install-guide.html
www.samba.gr.jp/project/samba-ja/user-guide.html
www.astro-hs.net/luna2000/jp/man/manual-w.html
www03.u-page.so-net.ne.jp/wc4/murano/manual5.htm
www.nedo.go.jp/itd/fellow/manual/kenkyu/k-mokuzi.html
www03.u-page.so-net.ne.jp/wc4/murano/ope.htm
www.lib.chs.nihon-u.ac.jp/opac/search/help/manual_v3.html
home.ntt.com/docs/html-jman/ncsa-j.html
www.nihs.go.jp/pdw/manual/pellej.htm
www.lamen.or.jp/lib/kansen/saisei/ssi99jp.html
www.saga-ed.go.jp/materials/edq01449/ncalc/ncman.htm
www.microstaff.co.jp/shop/netpack/manual/syssaver_manual.html
tknw.shiga-med.ac.jp/club/yacht/safe.html
www.ipc.chiba-u.ac.jp/docs/ssh.html
member.nifty.ne.jp/usr/hash/lint/use_lint.html
www.prime262.com/otap/manual/manual_otaseek.html
www.soc.titech.ac.jp/titsoc/FAQ/mailman.html
www.sapid.org/html2/GENERAL/INSTALL-0.html
wwwinfo.ncc.go.jp/NCC-CIS/pro/info/other/0sj/kokuchi.html
w11.dion.ne.jp/lan_service/service/hosting/server_rental/guide/guide_index.html
www.htokai.ac.jp/DA/tkhr/lecture/dw/tutorial-j.html
ssjp.net/m7419852/manjs-1.5.html
www.ask-corp.co.jp/UserManual_mac/homepage.html
www.furukawa.co.jp/network/mucho/HOWTO/mucho-stps/html/refmanual.html
www.med.kobe-u.ac.jp/anes/Manual/latex.html
square.umin.ac.jp/2000/doc/990610/sogoukiki.htm
ultra1.misasa.okayama-u.ac.jp/misasa/manu/manu.html
www2f.biglobe.ne.jp/%7Ehito/html40.html
plaza3.mbn.or.jp/%7Eh_ishida/EBStudio/EBStudio.html
timos.str.archi.tohoku.ac.jp/homepage/option/imagemap/ReadMe.html
www.ishikawa-nct.ac.jp/lab/I/iken/public_html/procon/1998/some/manual/index.html
www.chem.nagoya-u.ac.jp/%7Ebar/mdview/MANUAL.html
www.scripps.edu/%7Eyohei/zlog/zlogman.html
panda.ecs.cst.nihon-u.ac.jp/%7Ehiroshi/ac3d/ac3dmanj.html
www.biochem2.m.u-tokyo.ac.jp/web/manualindex_j.html
www.libweb.tdc.ac.jp/pubmed/mesh.html
ww3.tiki.ne.jp/%7Ey-takaya/manual.htm
ss.niaes.affrc.go.jp/pub/hseino/Amedas/manual.html
jt.mozilla.gr.jp/projects/xpcom/nsCOMPtr.html
biotech.nikkeibp.co.jp/POWERCONTENTS/0157/0157MAN.html
www.mahoroba.ne.jp/%7Efurutani/easy_man.html
www2.edd.osaka-sandai.ac.jp/%7Ey-fukui/doc/php3manual.html
lettuce.ms.u-tokyo.ac.jp/%7Ekato/NAMAZU/manual.html
www.apionet.or.jp/%7Eyoishizu/A_Manual/AllergyManual.html
www.ongs.net/quickcd/manual/index.html
www.naklab.dnj.ynu.ac.jp/%7Ehidefumi/manual/main.html
heasarc.gsfc.nasa.gov/docs/software/fitsio/c/f_user/fitsio.html
www.asahi-net.or.jp/%7Ewg5k-ickw/html/online/gdb-4.18/gdb_toc.html
www.asahi-net.or.jp/%7Ewg5k-ickw/html/online/flex-2.3.7/flex_toc.html

www-vox.dj.kit.ac.jp/nishi/cvs/cvs-manual/cvs-jp_toc.html
www-nagao.kuee.kyoto-u.ac.jp/member/tsuchiya/perl/info/perl-ja_toc.html
www.kt.rim.or.jp/%7Ekbk/gawk-30/gawk_toc.html
www2s.biglobe.ne.jp/%7Evtgf3mpr/eb_os2.html
www.imasy.or.jp/%7Ekazz/Gnus/html/gnus-j_toc.html
lagendra.s.kanazawa-u.ac.jp/ogurisu/manuals/mule.new/skk-9.6/skk_toc.html
lagendra.s.kanazawa-u.ac.jp/ogurisu/manuals/mule.new/cmail/cmail_toc.html
www.softagency.co.jp/mysql/Manual/mysql-3.23.18/manual.ja_toc.html
www.hlla.is.tsukuba.ac.jp/%7Eyas/coins/compsys-2000/gdb-4.17-doc/gdb_toc.html
lagendra.s.kanazawa-u.ac.jp/ogurisu/manuals/mule.new/yatex1.66/yatexj_toc.html
nasca.aist-nara.ac.jp/PHPDOC/manual.html
www.math.sci.kobe-u.ac.jp/OpenXM/1.1.3/doc/asir2000/html-jp/man_toc.html
www.mew.org/info/mew-j_toc.html
flex.ee.uec.ac.jp/texi/tgif-plus/tgif-plus.html
sawfish.gnome.gr.jp/doc/sawfish-ja.html
sapporo.cool.ne.jp/cybercafe/jpmanual.html
www.newtonlabs.com/ic/ic_toc.html
www.forest.dnj.ynu.ac.jp/sicstus3/sicstus_toc.html
www.gohome.org/wl/doc/wl-euc_toc.html
www.st.rim.or.jp/%7Ehyuki/man.htm
snaps.php.net/manual/ja/html/index.html
www.isle.kuis.kyoto-u.ac.jp/le4-sp/r4rs-ja/index.html
www.kt.rim.or.jp/%7Ekbk/regex/rx_toc.html
shakosv.sk.tsukuba.ac.jp/ShakoDoc/Editors/NEmacs/Gnus5-j/html/index.html
www.sra.co.jp/wingnut/gcc/cpp-j.html
www.sra.co.jp/wingnut/gcc/gcc-j.html
www.melt.kyutech.ac.jp/%7Emelt/Seminars/SimplyScheme/scm.html
web.kyoto-inet.or.jp/people/jeanne/word/ttfmanual.html
village.infoweb.ne.jp/%7Ekeizou/kouza/kouza.htm
member.nifty.ne.jp/zasiki/h_chair/manual.html
www.omedeto.com/junbiManual/index.html
www.d5.dion.ne.jp/%7Eyoshing/Taishoku.htm
www1.ocn.ne.jp/%7Eshopgirl/manualtoppage.html
members.tripod.com/%7Enatsu/diablo1.htm
www.segausers.gr.jp/sug/rc5/dnetc.html
group.lin.go.jp/souchi/honmokuji.htm
www.yutopia.or.jp/%7Etuchi/contents0.htm
www2.wbs.ne.jp/%7Ektw/manual.htm

以上, 142 個