修士学位論文

題目

電子マニュアルの構造を利用した文書評価メトリクス

指導教官

井上 克郎 教授

報告者

谷口 真也

平成 13 年 2 月 14 日

大阪大学 大学院基礎工学研究科

情報数理系専攻 ソフトウェア科学分野

平成 12 年度 修士学位論文

電子マニュアルの構造を利用した文書評価メトリクス

谷口 真也

内容梗概

CALS 等に代表される文書の電子化においては，情報の共有や，文書の再利用性，検索性を向上させることを目的として，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ってきたため，高品質のマニュアルを供給することは重要となってきている．大容量記憶媒体やインターネットの普及によって，マニュアルが電子化文書として提供される機会が増え，高品質のマニュアルを供給するために文書の構造が電子化の目的に即したものであるかを評価する必要がある．しかし，マニュアルを評価する際には，文書内容の評価に重点がおかれ，文書の持つ構造が評価対象となることは少ない．このため，マニュアルの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．

本研究では，HTML で記述された電子マニュアルを対象として，文書構造の良さを定量的に評価するためのメトリクスを，既存の文書に基づいて統計的手法を用いて導出し，提案するメトリクスによって検出された，品質の低い電子マニュアルに対する改善手法を提案し，具体的に修正を行うことによって文書構造を改善できることを示した．

主な用語

構造化文書 (Structured document)
電子マニュアル (Electric manuals)
メトリクス (Metrics)

目 次

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 [11].

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[10] やグループウェア [16] 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．

しかし，現状では，サンプル文書を形態素解析した結果に対する統計解析と，サンプル文書についての読者へのアンケート調査をもとにした，マニュアルのわかりやすさの評価 [12] は行われていても，マニュアルの構造に着目した評価はほとんど行われていない．このため，マニュアルの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った．

本研究では、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に，大量のマニュアルから品質の劣る文書を検出し，そのマニュアルをより品質の高いものにするための修正法を示す手法を提案する．

まず，参考文献 [14] を元に構造化文書とその構造の品質を評価するための基準を定義し，今回対象とする HTML 文書に対してその定義を適用した．次に，文書構造から算出可能な計測値を文書構造の評価基準と対応させ，Web 上から収集した HTML マニュアルからツールを利用して算出した．さらに算出された計測値を主成分解析等の統計的手法を用いることで集約した．最後に，各計測値に関して有意水準 5% で検定したときに異常な値を示すデータの分析を行い，それをもとに文書構造を評価するためのメトリクスを定義するとともに，そのメトリクスによって検出される文書の修正ガイドラインを定めた．また，検出されたデータを実際にガイドラインに従って修正し，構造の品質が高くなることを確認した．

以下，2 章では構造化文書，3 章では構造化文書に対する構造評価，4 章では電子マニュアルへの適用，5 章では構造化文書の評価手法，6 章では考察を述べる．最後に 7 章ではまとめと今後の課題について述べる．

2 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，SGML(Standard Generalized Mrakup Language)[2]，XML(Extensible Markup Language)[15, 17] 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管理やコスト管理をしやすくする，ドキュメントの作成作業を複数の担当者で分担する，ドキュメントの作成作業を効率化する，ドキュメントの品質を一定に保つ，ドキュメントの書式を統一す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[10] 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 文書構造化言語

文書構造化言語とはその名の通り文書構造を記述するための言語のことである．これらの言語では，原稿のファイルの中に，文字列のキャラクタコード以外のさまざまな属性情報（文字のタイプ（イタリックか，ボールドか，など）や組版情報，ハイパーリンク情報）などを，あらかじめ定義されたコマンドとして記述する．これらの言語の処理系は，そのファイルの中に記入されたコマンドを読みとり，そのコマンドで指定された通りに組版または表示を行うことになる．

文書構造化言語とは，文書構造を記述するためのものである．したがって，これらの言語を利用することにより構造が明確で整合性を持つ文書を記述することが可能となる．

このような文書構造化言語の一つの例として，TeX/LaTeX があげらる．これらは本来自然科学系の論文を書くために開発されたものだが，精通すれば非常に多種多様な文書を書くことができる．また，インターネットのホームページを書くときに利用する HTML(HyperText Markup Language) も文書構造化言語の代表的なもので，電子的参照用に単純化されたものである．

3　構造化文書に対する構造評価

前章で述べたように，さまざまな目的に応じて文書の構造化が行われているが，その目的を達成するためには適正な構造化を行う必要がある．文書の構造化については，さまざまな文献でその必要性，及び，その方法論について述べられている [1, 3, 4, 5, 6, 7]．本研究では，各文献の要素が統合されて記述されていた参考文献 [14] において述べられている文書の構造化を基準にすえることとした．

3.1　構造化文書

参考文献 [14] 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．

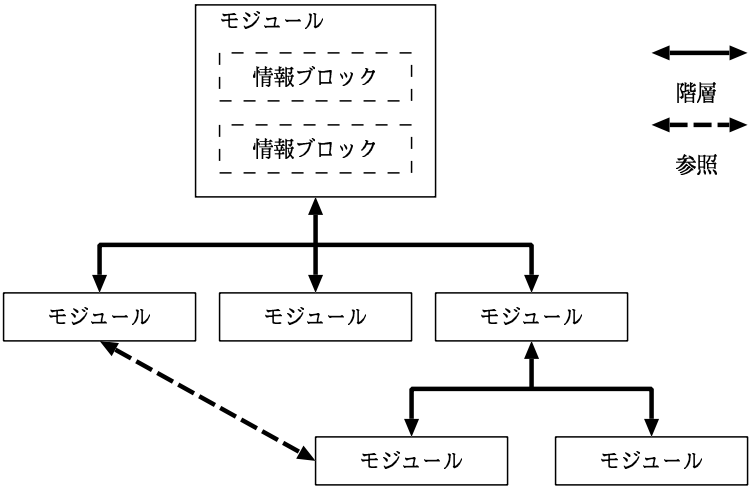

図 1: 構造化文書の模式図

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．

3.2 文書構造の評価基準

構造化文書が持つ長所を損なわないように，文書の構造化を適切に行うための方法論が，参考文献 [14] では述べられている．3.2.1 節でモジュール，3.2.2 節で階層に関する評価基準について述べる．

3.2.1 モジュール

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る．ユーザは文書内容をモジュール単位で理解することになるため，構造化文書作成時における文書内容のモジュールへの分割は適切に行わなければならない．

- モジュールのサイズは 1 ウィンドウ程度
  モジュールの具体的なサイズとしては，紙媒体で記述された文書であれば見開き 2 ページ程度，画面上で閲覧する電子化された文書であれば 1 ウィンドウ分，日本語に換算すると約千二百文字程度がよい．これは，ユーザがページをめくらずに情報を一覧できる情報量であり，これを保つことによってユーザが情報を把握しやすくなり，読みやすさが向上すると考えられる．

- 各モジュールのサイズは均等
  ユーザに対する情報の一覧性を保つために，文書中のモジュールのサイズは可能な限り均等に保つことが必要である．つまり，サイズの小さなモジュールは他の関連するモジュールと組合わせ，サイズの大きなモジュールはより小さく分割して文書を構造化することが望ましい．ただし，一続きの操作手順といったような内容的に複数のモジュールに切り分けるのが難しいものに関してはこの限りではない．

- モジュールは複数の情報ブロックから構成
  内容的な理由から一覧性を保つサイズに切り分けられないモジュールに関しては，モジュール内を適度な大きさの情報ブロックに分けることによって，ユーザに対する一覧性を提供する必要がある．また，そのようなモジュールに限らず，モジュール内を複数の情報ブロックに分割することで、読みやすさを向上させることに加えて，文書作成の柔軟性を高め構造化作業をやりやすくすることが可能である．

3.2.2 階層

モジュール単位に分割された文書を階層化することによって，情報のまとまりや上下関係を明確に表現することが可能となる．ただし，その階層化が適切でないと逆に情報の把握が困難になるという結果を生むことになる．

- モジュールが構成する階層は基本的に 3 階層にする
  文書が階層化され情報を把握しやすいように整理されていた場合でも，実際にユーザが文書を読む際には次ページに進むか，前ページに戻るかという二者択一となる．この結果，どんなに階層が深くなってもユーザは情報の並び順に従って読むことになる．しかし，階層が増えることは，ユーザが現在読んでいるモジュールの階層関係を直感的に認識することを困難にするため，情報が細かく分類された階層の深い文書よりは，情報の並び順が十分に考えられた階層の浅い文書のほうが，ユーザにとっては読みやすいといえる．

- 各モジュールの子供は適切な数にする
  必要以上に長い階層は，ユーザが階層関係を理解を妨げる上に，ユーザの文書を読む意欲を減少させるため，文書を構成するモジュールが持つ子は適当な数 (1 桁以内程度) に抑える必要がある．ただし，リファレンス系の文書では 1 つの階層内の項目数が 1 桁を越える場合がある．

4　電子マニュアルへの適用

本研究では，電子マニュアルに対する文書構造の品質に対する定量的な評価をすることを目的としている．そのためには前章で述べた文書構造についてより明確な定義をした上で，その定義をを電子マニュアルに対して適用する必要がある．

4.1　文書構造の定義

前述のように，文書構造を決定する要素としてはモジュール，情報ブロック，階層，参照の 4 つのものが挙げられる．参考文献 [14] におけるこれらの要素に関する定義には曖昧な部分が含まれているため，そのままでは定量的評価に用いることは難しい．そこで，各要素に関してより厳密に定義を行った．

4.1.1　モジュール

参考文献 [14] の定義では読者に一度に提供するための情報量を表すための単位とあるが，このままでは文書中のどの部分からどの部分が一つのモジュールであるかということを決定する因子に欠ける．そこで，本研究ではモジュールを，文書中で見出しによって分割可能な一連の情報と定義する．つまり，モジュールを決定する因子として見出しを用いることにした．これによって、文書を見出しによって，章，節，項という形でモジュール単位に分割することが可能となる．

4.1.2　情報ブロック

参考文献 [14] の定義では意味を伝達可能な情報量を示す単位とあり，モジュールの場合と同様に，モジュール内を情報ブロックで分割するための決定的な因子が存在しない．よって，モジュールの場合と同様に情報ブロックを分割するための因子として，段落を用いることとした．つまり，情報ブロックを各モジュール内に含まれる段落と定義によりすることによって，例えば，一般的な文書で段落の表現に用いられる 1 文字のインデントを用いることでモジュール内を分割することが可能となる．

4.1.3　階層

階層は，モジュール間の上下関係を示す，という参考文献 [14] の定義そのままで十分である．前述のモジュールの定義により，文書が見出しによって分割可能となるため，その見出しの情報をもとにモジュール間の階層関係は容易に判別することが可能である．

4.1.4　参照

階層の場合と同様に参照もまた，関連のある箇所同士を指し示すという定義で十分機能する．モジュールを見出しによって分割した後に，階層関係を判別してやれば，参照は階層関係以外を示すモジュール間の関係として認識可能である．

4.2　HTML で記述された構造化文書

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．

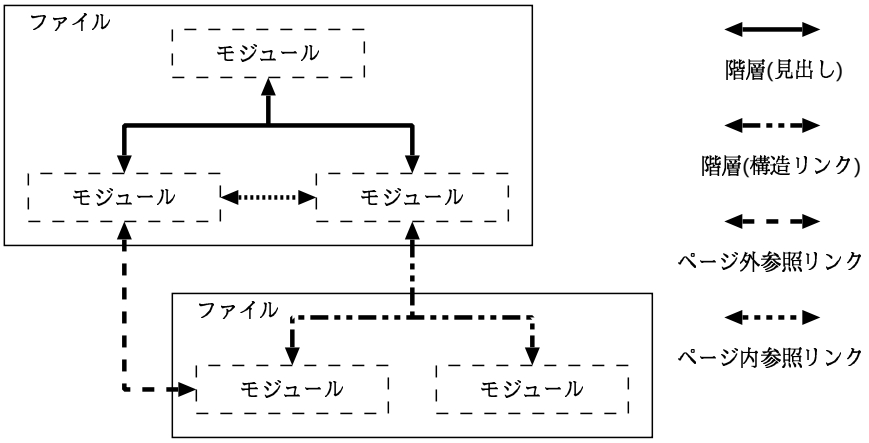

図 2: HTML で記述された構造化文書の模式図

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．

4.3　HTML マニュアルの定義

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．

4.3.1 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 (⟨H1⟩,...，⟨H6⟩) 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．

4.3.2 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 (⟨P⟩) 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．

4.3.3 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 (⟨Hn⟩: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．

4.3.4 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．

4.4 HTML マニュアルの評価基準

4.2 節で，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ことを述べた．そのため，前述の文書構造の評価基準に加えて，それらを HTML 文書に適用したときに生じる構造評価基準についても考えておかなければならない．

- 1 ファイルに記述されるのは 1 モジュール
  HTML 文書においてモジュールをどのようにファイル上に構成するかは筆者の自由である．しかし，構造化文書の特色である，文書作成作業の分担や，再利用性の向上等を考慮すると，1 ファイル上に 1 モジュールがあることが理想的である．

- 1 モジュールにつき 1 つのページ内参照リンク
  一つのファイルに複数のモジュールが含まれるなどファイルに記述されている内容を

ウィンドウで一覧できない場合，ページ内参照リンクを利用することで，モジュール間の移動がやりやすくなる．そのため，少なくともモジュールの数と同じだけのページ内参照リンクは必要である．

- 関連のあるモジュール間でのページ外参照リンク
  ファイル数や木の深さが大きい文書では，ユーザが文書構成を把握しにくい上にモジュール間の移動もやりづらくなる．その問題を解決するために，関連のあるモジュール間にはできるだけページ外参照リンクを用いたほうがよい

5　構造化文書の評価手法

4 章において，3 章で述べた文書構造，及び，その評価基準を定量的な評価が可能となるように HTML マニュアルに対して適用した．その適用の結果得られた文書構造を評価するための基準は以下のようなものである．(以下，この番号によって各基準を表す．)

1. モジュールサイズは 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
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

本章では，これらの評価基準を定量的に評価するためのメトリクスの定義とその結果検出される文書の修正ガイドラインについて述べる．

5.1　文書構造から算出可能な計測値

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，正規化分散，尖度，変動係数についてそれぞれ計測する．

次に，無作為に収集した 142 件 (7885 ファイル) の HTML で記述されたマニュアル文書に関して，計測ツールを利用して，前述の計測値を算出した．その結果から，主成分解析などの統計的手法を用いて相互に関連性の高い評価値を集約した．その結果得られた計測値を以下に示す．() 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．

- 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)
- 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/ファイルの平均 (7，8)
- モジュール数/ファイルの平均 (6)

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．

5.2 データ分析

収集した HTML マニュアルから得られた各計測値に関して，有意水準 5%で両側検定を行ったときに棄却域に属するデータの分析を行った．その結果を検出された文書の特徴を簡単に述べる．

- 階層の深さ
  - 線形に構成された部分を含む
- 構造リンク数
  - 線形に構成された部分を含む
  - サイズに対して章構成が詳細
- ページ内参照リンク数
  - 単一ファイルに全ての内容を含む
- ページ外参照リンク数
  - トップページに全てのノードのリンクがある
  - ナビゲーションリンクが詳細にはってある
- 文字数/モジュール

平均	文書内をモジュールに分割していない
標準偏差	文書中に極端に大きいファイルがいくつかある
変動係数	トップページに文書内容のほぼ全てが含まれる 様々な大きさのモジュールが含まれる

- 情報ブロック数/モジュール

平均	モジュールが大きいため必然的に情報ブロックが大きくなっている 段落タグの使い方に誤りがある
標準偏差	一部分だけに段落タグが使われている 文書の中に巨大なモジュールが存在し，そのモジュール内に多数の情報ブロックが含まれている
変動係数	一部分だけに段落タグが使われている 文書の中に巨大なモジュールが存在し，そのモジュール内に多数の情報ブロックが含まれている

- 子供の数/モジュール

平均	トップページが全てのノードへのリンクを持つ
標準偏差	トップページがほとんどのノードのリンクを持つ
変動係数	トップページがほとんどのノードのリンクを持つ

- 文字数/ファイルの平均
  - 単一ファイルで全てが記述されている
- モジュール数/ファイルの平均
  - 単一ファイル構成で，サイズが大きい

標準偏差と変動係数に関して検出される文書は概ね似た傾向を示すが，結果となってあらわれる文書には違いがみられた．標準偏差の場合は，ある程度文書が大きく，かつ，ちらばりが大きい時に値が大きくなる文書が検出された．一方，変動係数の場合は，全体的に小さい値のなかに一つだけ大きな値がある文書が検出された．

5.3 構造評価メトリクス

分析したデータをもとに，各評価基準に関して品質の低い文書を検出するための評価メトリクスを定義した．

1. モジュールサイズは 1 ウィンドウ程度
   この基準に関連のあるデータとしては，文字数/モジュール，情報ブロック数/モジュールの平均値がある．分析の結果，情報ブロック数/モジュールは予想以上に文字数/モジュールとの関連が深く，検出されるデータは類似していた．そこで，この評価基準を検出するメトリクスとして，文字数/モジュールの平均値を用いる．

2. 各モジュールのサイズは均等
   文字数/モジュール，情報ブロック数/モジュールの標準偏差と変動係数をこの基準を計測するための計測値として用意しておいたが，情報ブロック数/モジュールの標準偏差と変動係数から，検出されている文書は評価基準とはあまり関係のない文書であった．これは，情報ブロックの文字数の散らばりがモジュールの文字数の散らばりとそれほど関連が深くなかったためであると思われる．よって，評価のためのメトリクスには文字数/モジュールの標準偏差と変動係数を利用する．同じ因子に属する標準偏差と変動係数から検出される文書は，似た傾向を示すが，結果となってあらわれる文書には違いがみられるということは既に述べたとおりである．よって，評価メトリクスとしては，各計測値において外れ値となった文書の和集合を利用する．

3. モジュールは複数の情報ブロックから構成
   この基準を評価するための計測値は，情報ブロック数/モジュールのみであり，これをそのままメトリクスとして利用する．

4. モジュールが構成する階層は基本的に 3 階層
   この基準は，深さと構造リンクが関連している．構造リンクは文書構造の階層を構成する要素であるが，対象となる HTML マニュアルでは構造リンク以外にもモジュールの論理構造によって，階層が構成される．本研究では階層を構成する二種類の方法が同じ確率で生じると考え，この基準の評価メトリクスについて，深さは構造リンクの 2 倍の重みづけをして扱うことにした．よって，深さと構造リンクの偏差値を 2:1 の割合で加算し，平均をとったものを評価基準として用いる．

5. 各モジュールの子は適切な数にする
   子供の数/モジュールの平均，標準偏差，変動係数がこの評価基準に関連する計測値である．同じ因子に属する標準偏差，変動係数で検出される文書の違いについては 1. で

も述べたとおりである．よって，メトリクスとしては各計測値で外れた値をとった文書の和集合をとることにする．

6. 1 ファイルに記述されるのは 1 モジュールこれに関連する計測値は，モジュール/ファイルのみであり，これをそのままメトリクスとして利用する．

7. 1 モジュールにつき 1 つのページ内参照リンク
   文字数/ファイルとページ内参照リンクが関連する測定値である．評価基準からメトリクスとしては，文字数/ファイルが 5000 以上で，かつ、ページ内参照リンクが 1 以上を基準として用いる．

8. ファイルサイズとページ外参照リンクのバランス
   この評価基準に関連する測定値は，ページ外参照リンクと文字数/ファイルである．しかしながら，これらの測定値において外れ値となったファイルを実際に検証してみたが，特に問題は感じられなかった．この理由としては，参照がモジュール間の相互関連を示すためのものであり，その性質が文書にとって必要十分なものではないからと推定できる．

基準	検出データ	特徴
1	2 件	見出しタグが誤って使われている
2	6 件	極端に大きいモジュールが存在する モジュールのばらつきが大きい
3	2 件	段落タグが大量に使われている
4	2 件	文書に線形な部分が含まれる
5	14 件	子供が大量にある文書が含まれる
6	3 件	サイズの大きい単一ファイルで記述されている
7	2 件	複数のモジュールをもつファイルが含まれる サイズの大きいモジュールが含まれる

表 1: メトリクスにより検出されるデータ件数とその特徴

5.4 文書構造の修正ガイドライン

各評価基準に対するメトリクスにより検出されたデータの修正するためのガイドラインを定めた．

1. モジュールサイズは 1 ウィンドウ程度

   - 見出しタグが使われていない，あるいは誤って使われている
     見出しタグを利用してモジュールに分割する．ファイル規模が大きい場合は，さらに構造リンクを用いた構造化を行う.

2. 各モジュールのサイズは均等

   - 極端に大きいモジュールが存在する
     そのファイル内を見出しタグを用いてモジュールに分割し，分割されたモジュールを，ファイル単位で構成しなおす.
   - モジュールのばらつきが大きい
     文書内容を再考し，モジュールを再度分割しなおす.

3. モジュールは複数のブロックから構成

   - 段落タグが大量に使われている
     文書内容を再考し，意味のあるまとまり単位で段落を構成しなおす.
   - 段落タグが使われていない
     意味のあるまとまり単位で段落タグを利用した記述を行う.

4. モジュールが構成する階層は基本的に 3 階層

   - 文書に線形な部分が含まれる
     線形になっている部分を木構造に構成しなおす．ただし，線形なっている部分の最初が親ではないときには，新たに親となるモジュールを作らなければならない.

5. 各モジュールの子は適切な数

   - 子供が大量にある文書が含まれる
     内容的にまとめられる子供があれば，それらの親となるモジュールを作成し，そのモジュールを元の親の子供とする.

6. 1 ファイルに記述されるのは 1 モジュール

   - サイズの大きい単一ファイルで記述されている
     1 モジュール単位にファイルを分割する.
   - モジュールをファイルに分割する基準が一定でない
     モジュールの階層構造中の位置といった明確な基準でファイルの分割を行う.

7. 1 モジュールにごとに 1 つのページ内参照リンク

- 複数のモジュールをもつファイルが含まれる
  モジュール単位でページ内参照リンクを記述する．サイズの大きいモジュールがある場合は，約 1200 文字ごとにページ内参照リンクを記述する．
- サイズの大きいモジュールが含まれる
  サイズの大きいモジュールがある場合は，約 1200 文字ごとにページ内参照リンクを記述する．

5.5 構造化文書の修正例

本節では，各基準に関する構造化文書の修正ガイドラインに従って文書を実際に修正した例について説明する．

• 基準 1

モジュールサイズは 1 ウィンドウ程度あるという基準で検出される品質が低いマニュアルは全部で 2 件である．そのうちマニュアル「Guideline for the Prevention of Surgical Site Infe ction」について，その問題点と修正法について述べる．

このマニュアルは単一のファイルから構成されており，図 3 に示したように，見出しに相当する「第一部：手術部位感染 (ＳＳＩ):」に見出しタグが使用されていないため，ユーザは文書の内容がどこからどこまででまとまっているかを認識することが難しい．

```
<P ALIGN="JUSTIFY">第一部：手術部位感染（ＳＳＩ):</P>
<P ALIGN="JUSTIFY">概要</P>
<P ALIGN="JUSTIFY">A．はじめに</P>
<P ALIGN="JUSTIFY">　１９世紀半ばまで手術患者は、通常術後に発熱をきた
し、手術創からの排膿があり、重症の敗血症となり、時には死亡した。</P>
<P ALIGN="JUSTIFY">手術後の感染による死亡が減少するようになったのは Joseph
 Lister が１８６０年代の後半に抗菌という原理を導入してからである。</P>
<P ALIGN="JUSTIFY">Lister の仕事によって、外科手術は感染と死を伴う作業
から病気を終わらせて生命を永らえる技術へと劇的に変わった。</P>
```

図 3: Guideline for the Prevention of Surgical Site Infection ソース修正前

よって，評価基準 1 の修正ガイドラインに従い，図 4 に示したように見出しに相当する部分全てに見出しタグを記述し，文書内をモジュールに分割した．この修正の結果，ユーザは文書がどこで分割されているか認識しやすくなり，モジュール単位で文書を読むことが可能

となる．

```
<H1>第一部：手術部位感染（ＳＳＩ）:</H1>
<P ALIGN="JUSTIFY">概要</P>
<P ALIGN="JUSTIFY">A．はじめに</P>
<P ALIGN="JUSTIFY">　１９世紀半ばまで手術患者は、通常術後に発熱をきた
し、手術創からの排膿があり、重症の敗血症となり、時には死亡した。</P>
<P ALIGN="JUSTIFY">手術後の感染による死亡が減少するようになったのは Joseph
 Lister が１８６０年代の後半に抗菌という原理を導入してからである。</P>
<P ALIGN="JUSTIFY">Listerの仕事によって、外科手術は感染と死を伴う作業
から病気を終わらせて生命を永らえる技術へと劇的に変わった。</P>
```

図 4: Guideline for the Prevention of Surgical Site Infection ソース修正後

• **基準 2**

各モジュールのサイズは均等であるという基準で検出される品質が低いマニュアルは全部で 6 件である．そのうちマニュアル「ImageViewer-J ヘルプ」について，その問題点と修正法について述べる．このマニュアルは 16 のファイルから構成されているが，そのうちの 1 つのファイルが FAQ という一つの大きなモジュールとなっており，基準 1 の場合と同じくユーザは文書の内容がどこからどこまででまとまっているかを認識することが難しい．よって，評価基準 2 の修正ガイドラインに従い，各質問に見出しタグをつけ，さらに質問ごとにファイルを分割した．この修正の結果，ユーザは文書がどこで分割されているか認識しやすくなり，モジュール単位で文書を読むことが可能となる．

• **基準 3**

モジュールは複数の情報ブロックから構成するという基準で検出される品質が低いマニュアルは全部で 2 件である．そのうちマニュアル「樹脂ペレット漏出防止マニュアル」について，その問題点と修正法について述べる．

このマニュアルは単一のファイルから構成されており，図 5 に示したように，番号つきリストを段落タグを用いて記述しているためにモジュール内が大量の情報ブロックに分割されてしまう．

よって，評価基準 3 の修正ガイドラインに従い，図 6 に示したように番号つきリストを記述するためのタグを用いて記述してやることで意味のあるまとまり単位で情報ブロックを記述することができる．

```
<P><FONT FACE="ＭＳ 明朝">　1．本年度中に、石油化学工業協会始め当連盟会
員団体の推進体制を整備し、具体的活動に入る</FONT></P>
<P><FONT FACE="ＭＳ 明朝">　2．次年度中にプラスチック関連業界全体に亘る
運動に拡大する</FONT></P>
<P><FONT FACE="ＭＳ 明朝">　3．国内での積極的推進・展開と合わせ、政府及
び民間関係方面の協力を得て、近隣諸国を中心とする海外へ普及推進に着手す
る</FONT></P>
<P><FONT FACE="ＭＳ 明朝">の三段階にて強力に取進め、逐次実効あるものへ
と展開致したいと考えております。
</FONT><BR></P>
```

図 5: 樹脂ペレット漏出防止マニュアル ソース修正前

```
<P><OL>
<LI><FONT FACE="ＭＳ 明朝">本年度中に、石油化学工業協会始め当連盟会員
団体の推進体制を整備し、具体的活動に入る</FONT></LI>
<LI><FONT FACE="ＭＳ 明朝">次年度中にプラスチック関連業界全体に亘る運
動に拡大する</FONT></LI>
<LI><FONT FACE="ＭＳ 明朝">国内での積極的推進・展開と合わせ、政府及び民
間関係方面の協力を得て、近隣諸国を中心とする海外へ普及推進に着手する
</FONT></LI>
</OL>
<FONT FACE="ＭＳ 明朝">の三段階にて強力に取進め、逐次実効あるものへ
と展開致したいと考えております。
</FONT><BR>
</P>
```

図 6: 樹脂ペレット漏出防止マニュアル ソース修正後

• 基準 4

モジュールが構成する階層構造の深さは 3 であるという基準で検出される品質が低いとされるマニュアルは全部で 2 件である．そのうちマニュアル「山地酪農の技術指導書」について，その問題点と修正法について述べる．

マニュアルは 86 個のファイルから構成されたマニュアルであり，一見しただけでは特に問題のない構造化文書に思える．しかし，ある手順部分に関して，その説明を図 7 のように HTML 的に線形に記述されている．

実際には，マニュアル中の「草地生産性向上対策事業の手引き」について記述した部分のトップページが図 8 に示したように，手引きの最初の部分にのみリンクがはられ，トップページ以降はトップと次ページにのみリンクがはられている．このため，ユーザはその手順の概要を把握しづらく，また，各手順を直接参照することができなくなっている．

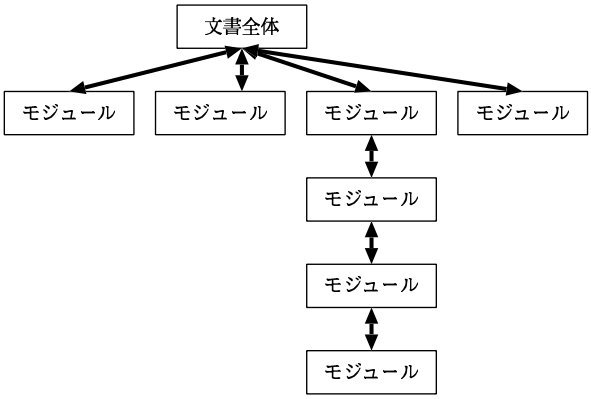

図 7: 山地酪農の技術指導書がもつ文書構造の模式図 (修正前)

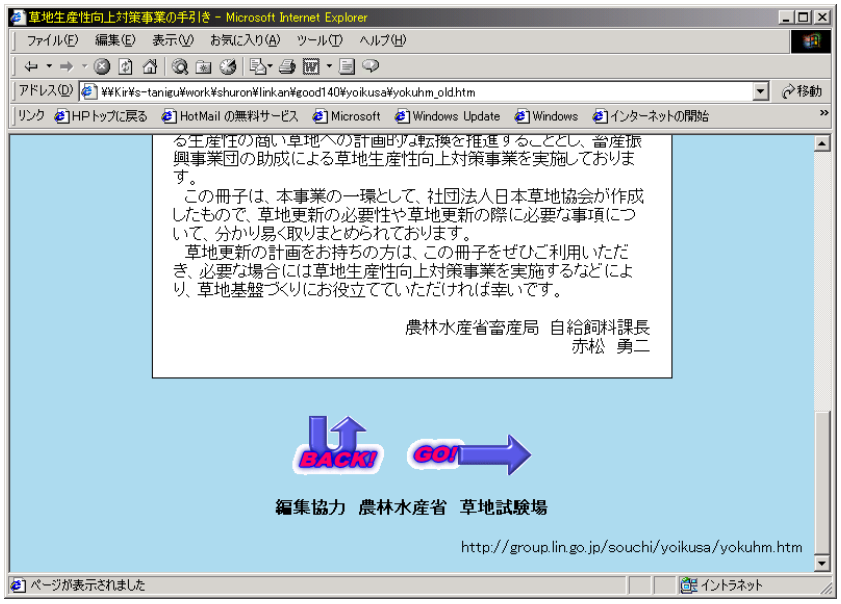

図 8: 草地生産性向上対策事業の手引き トップ (修正前)

よって，評価基準 4 のガイドラインに従い，文書構造を修正する．このマニュアルは，線形に記述されている部分の最初のモジュールが残りのモジュールの親となる要素を含んでいるため，木構造に再構成する際には特に親のモジュールを作ることなく最初のモジュールを親にすることで木構造に再構成することが可能である．

今回の場合は，図 10 に示したようにトップページに各手順へのリンクを記述することにより，階層構造を構成した．これによって，ユーザはトップページを見ることにより草地生

産性向上対策事業の手引きの概要を知ることができ，また，各手順を直接参照することが可能となった．また，修正前の線形に記述された構造リンクは参照リンクとして保存され，各手順を順を追って閲覧することももちろん可能である．

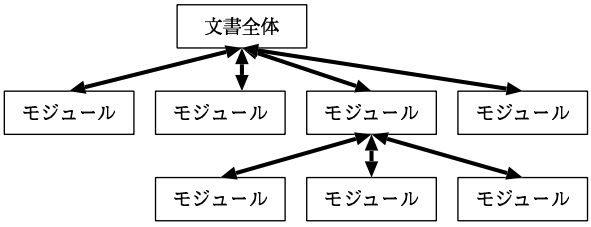

図 9: 山地酪農の技術指導書がもつ文書構造の模式図 (修正後)

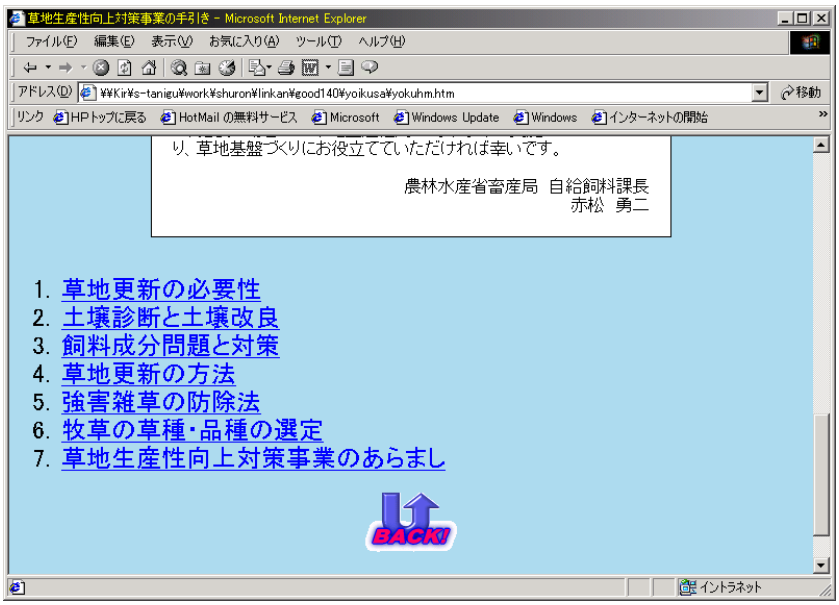

図 10: 草地生産性向上対策事業の手引き トップ (修正後)

● 基準 5

各モジュールの子供は 1 桁以内であるという基準で検出される品質が低いマニュアルは全部で 14 件である．そのうちマニュアル「The Perl5 Manual」について，その問題点と修正法について述べる．

このマニュアルは 202 のファイルから構成されており，図 11 に示したように，トップページから子ノードだけでなく孫ノードにまでリンクがはられている．よって，ユーザは文書の階層構造を誤認する可能性がある．

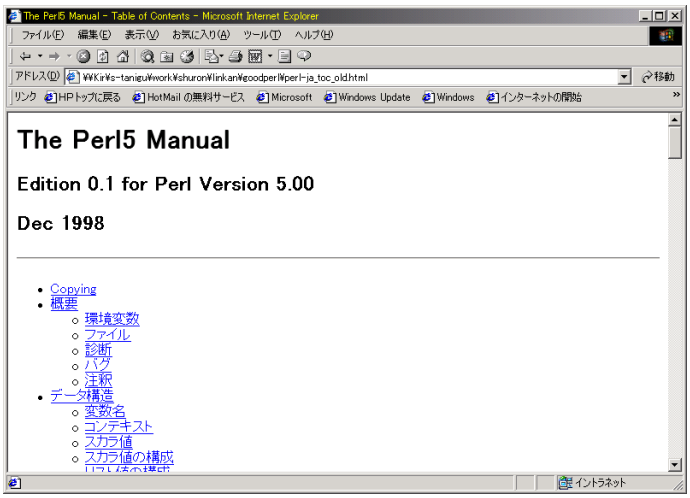

図 11: The Perl5 Manual トップ (修正前)

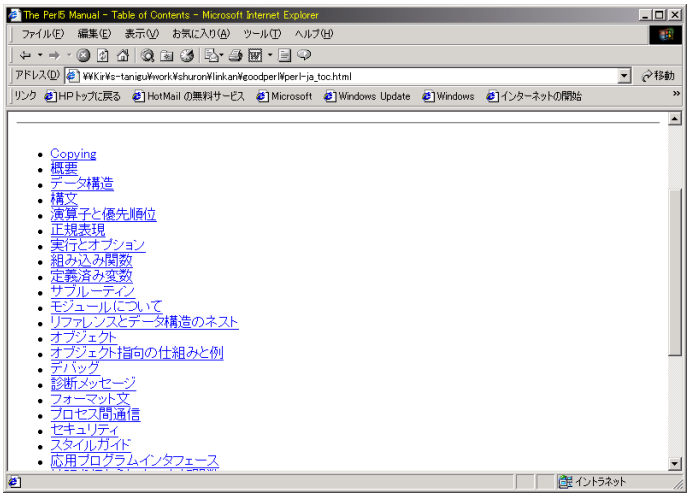

図 12: The Perl5 Manual トップ (修正後)

よって，評価基準 5 の修正ガイドラインに従い，図 12 に示したようにトップページからのリンクを子ノードのみにした．これによってユーザが文書の階層構造を誤認する可能性がなくなる．また，各モジュールを直接参照できるようにコンテンツページを新たに作成した．

• 基準 6

各モジュールの子供は 1 桁以内であるという基準で検出される品質が低いマニュアルは全部で 3 件である．そのうちマニュアル「GCC マニュアル」について，その問題点と修正法について述べる．

このマニュアルは 1 のファイルから構成されているが，そのサイズは 1.5M に及び，その中に 264 個のモジュールが含まれている．よって，この文書を修正する際には目的のモジュールを発見することでさえ非常な労力を要することになる

よって，評価基準 6 の修正ガイドラインに従い，図 13 に示したようなトップページを作成し，以下 264 のモジュールをそれぞれ一つのファイルに記述し階層構造をなすように再構成した．

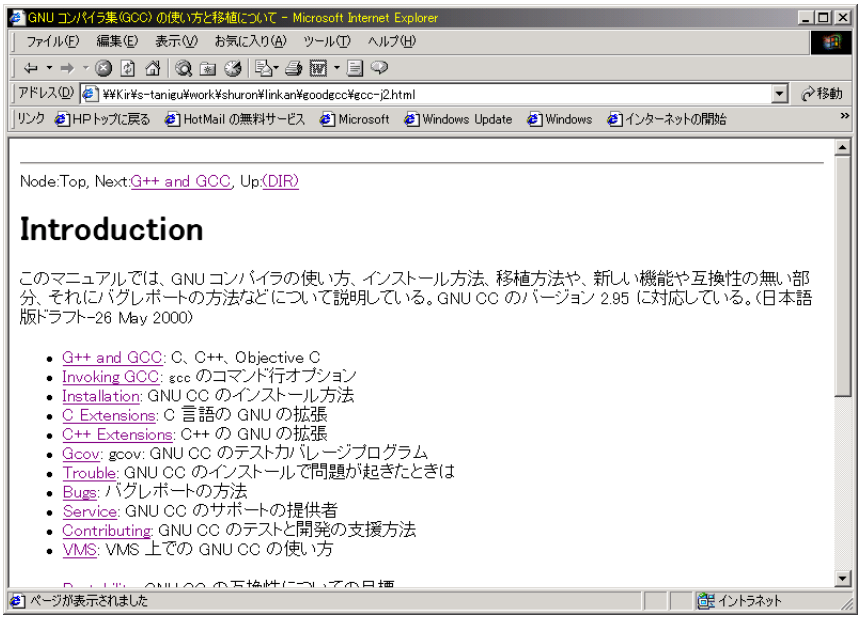

図 13: GCC マニュアル トップ (修正後)

この結果，この文書を修正する際には目的のモジュールが含まれるファイルがわかればそのファイル内に含まれるモジュールを修正するだけで十分である．

• 基準 7

1 モジュールにごとに 1 つのページ内参照リンクで検出される品質が低いとされるマニュアルは全部で 9 件である．検出されたデータのうち，「染めごころ 操作マニュアル」について，その問題点と修正法について述べる．

このマニュアルは単一のファイルから構成されており，12 個のモジュールを持っているが，図 14 に示すようにページ内リンクが全くはられていなため，ユーザはモジュール間の移動が非常にやりづらくなっている．

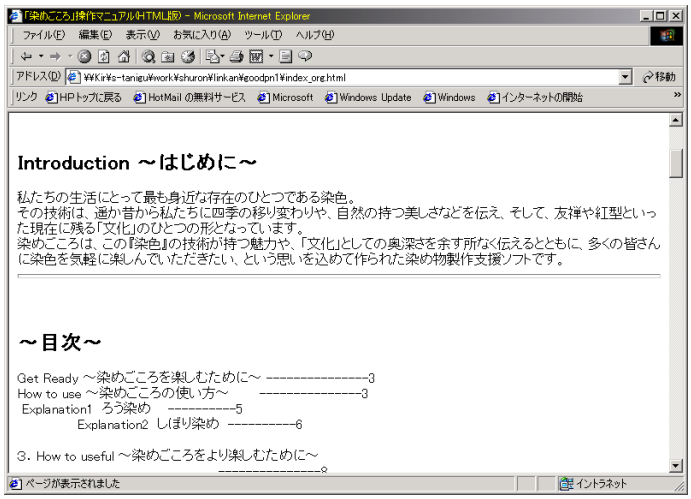

図 14: 染めごころ 操作マニュアル (修正前)

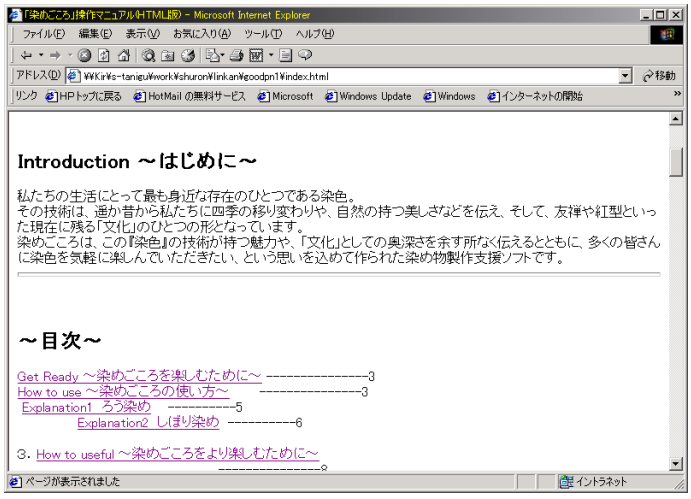

図 15: 染めごころ 操作マニュアル (修正後)

よって，評価基準 7 の修正ガイドラインに従い，図 15 に示したように目次の部分から書くモジュールへのページ内リンクを作成するとともに各モジュールから目次へのページ内リンクを記述した．

これによって，ユーザは目次を利用して参照したいモジュールに直接飛ぶことができるようになり，モジュール間の移動が非常にやりやすくなる．

6　考察

本手法では，ほとんどの評価基準において，品質の低い文書を検出することができた．また，その結果得られた品質の低い文書を今回示した修正法に従うことで，それらの文書の持つ構造の品質を高くなる．

しかし，基準 3，5，7 においてデータの検出が適切な結果を得られない場合があった．各基準について適切な結果が得られなかった理由について考察する．

基準 3 で検出されたデータは段落タグが必要以上に利用された文書である．この評価基準ではそのような文書の他にモジュール内に含まれる情報ブロックが極端に少ない文書の検出も意図していたが，実際には検出されなかった．これは，今回収集した HTML マニュアルに段落タグをあまり利用していないものが数多く含まれていたため，情報ブロックを段落タグで表現するという定義が不十分であっためと考えられる．

基準 5 において，実際に検出されるのは、texinfo に代表されるようなトップページに全てのノードへのリンクが張ってある文書である．これは，今回利用したツールが HTML の構造リンクによる階層と，モジュール間の論理構造から構成される階層のうち，前者を優先して判断しているために，孫ノードのように，本来ページ外参照リンクであるものまで構造リンク、すなわち子ノードへのリンクとして捉えているためである．

基準 7 では，大規模であるにも関わらず，単一ファイルに内容が全て記述された文書が検出された．この評価基準では，そのような文献のほかにモジュールをファイルに配置する基準が一定でない文書の検出も意図していたが，そのような文書は検出されなかった．原因としては，この基準に対して用いたモジュール/ファイルというメトリクスでは，ファイル数とモジュール数との関連を十分に示すことができていないことが考えられる．

今回，Web 上に存在する様々な種類の HTML マニュアルを収集し，その構造の評価を行った．その結果，現在提供されている HTML マニュアルには文書構造の品質が低いものが少なからずあることが確認できた．よって，本研究で提案した評価メトリクスを用い文書構造の品質が低いマニュアルを検出し，ガイドラインに従ってより品質の高いものにすることは重要である．

7　まとめ

本研究では，HTML で記述された電子マニュアルを対象として，文書構造の良さを定量的に評価するためのメトリクスを，既存の文書に基づいて統計的手法を用いて導出した．また，提案するメトリクスによって検出された，品質の低い電子マニュアルに対する改善手法を提案し，具体的に修正を行うことによって文書構造を改善できることを示した．

今後の課題としては，まず第一に，本研究では，142 件のデータを対象に評価を行ったが，より大量の HTML マニュアルを収集し，分析を加えることにより，さらに正確な評価メトリクスを得ることができると考えられる．また，十分な数がそろった段階で各マニュアルが対象とする分野単位で分析を行うことによって，マニュアルが対象とする分野に特有の問題点を検出することができると思われる．

次に，Web ブラウザが HTML の柔軟な記述を許しているために構文的な正しさの度合いが様々な HTML 文書が存在していることをふまえ，文書構造の品質と HTML 文書の構文的正しさとの関連性を調べてみたいと考えている．

また，文書構造の品質が実際にどの程度の影響を与えているかどうかを調べるために，実際に構造の品質が様々な文書に関して再利用を行い，その効率を調べる評価実験を行いたい．

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，ご協力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 川口 真司 君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] 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)

[9] 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)

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

[11] 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)

[12] 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)

[13] 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).

[14] 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).

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

[16] “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).

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