2K-4 電子マニュアルの構造を利用した文書評価メトリクス

川口真司[†]　谷口真也[‡]　松下誠[†]　井上克郎[†]

[†]大阪大学大学院基礎工学研究科　[‡]NTT ソフトウェア

1　はじめに

近年, 一般社会におけるソフトウェアの重要性が認識されてきている. それに伴い, ソフトウェアの開発・利用を容易に行うために高品質なマニュアルが必要とされている. 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.

これまでに, 文献 [2] 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を行った.

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, すなわち各計測項目が文書構造の評価に有用であることが確認できた.

2　構造化文書

本節では, 文献 [2] において定義されている, 構造化文書とその構成要素の概略を述べる.

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である．構造化文書は, 基本的構成要素である**モジュール**, **情報ブロック**と, 構成要素間の関係を表す**階層**, **参照**から成り立っている [3].

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，構造化文書の基本的構成要素である. HTML マニュアルにおいては, モジュールは見出しタグ (⟨H1⟩,...，⟨H6⟩) によって分割可能な一連の情報と定義する.

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．ここでは, 段落タグ (⟨P⟩) により記述された段落を情報ブロックと定義する.

階層とは, モジュール間の上下関係を示すものである．HTML マニュアルにおける階層は次の二つのパターンで定義される．一つ目は同一ファイル間のモジュールの階層関係で，見出しタグ (⟨Hn⟩:n は 1 から 6 の整数) の大小 (整数 n の大小) によって表されるものとする．もう一つは複数のファイル間のモジュールの階層関係で，ファイル間のリンクによって表されるものとする.

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．HTML における参照はリンクによって表される.

3　構造評価手法

文献 [2] では, 構造化文書を評価するための基準として以下の 8 個をあげている.

(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にはページ外参照リンクをはる

本研究では, これら評価基準を定量的に評価するために, 次のような作業を行なった.

(1) サンプルデータの収集
(2) 評価基準を定量的に評価するための計測項目の定義
(3) データ計測および統計的手法を用いた項目数の絞り込み
(4) 異常な値を示したファイルの抽出
(5) 抽出されたファイルの分析

まず, サンプルとして 142 個の HTML ドキュメントを検索サイトを用いて集め, 評価基準を元に定めた 13 個の計測項目の実際の値を計測した. そののち, 主成分分析 [1] を用いて余分な計測項目を削除した. 結果として, 以下の 9 項目について分析を行った.

Metrics Using Electric Manual Structure

[†]Shinji Kawaguchi, [‡]Shinya Taniguchi, [†]Makoto Matsushita and [†]Katsuro Inoue
[†]Graduate School of Engineering Science, Osaka University
[‡]NTT Software

計測項目	平均	標準偏差
DoH	3.218309859	2.312826392
SL	0.795467616	0.937877775
IaRL	1.546763111	2.583110566
IrRL	4.539997172	5.865490262
L/M_Avg	3049.499294	17241.46901
L/M_SD	1189.184811	2584.57701
L/M_RSD	1.13093818	0.737258368
IB/M_Avg	27.71743483	209.9704552
IB/M_SD	7.413351076	16.4234133
IB/M_RSD	1.311888054	1.228550273
C/M_Avg	9.980392604	20.86888555
C/M_SD	8.382458725	15.30438072
C/M_RSD	0.879385607	0.714265478
L/F_Avg	13508.47483	62830.35277
M/F_Avg	12.56327214	29.2199919

表 1: 各計測項目の平均, 標準偏差

計測項目	閾値	文書数
DoH	8	5
SL	2.72	4
IaRL	6.92	9
IrRL	16.2	11
L/M_Avg	42394	2
L/M_SD	7661	3
L/M_RSD	2.68	3
IB/M_Avg	476	2
IB/M_SD	43.0	7
IB/M_RSD	4.12	4
C/M_Avg	71.5	4
C/M_SD	42.7	7
C/M_RSD	2.37	6
L/F_Avg	185366	3
M/F_Avg	78	3

表 2: 各計測項目の異常値とそれを越える文書の数

(1) 階層の深さ (DoH)
(2) 構造リンク数 (SL)
(3) ページ内参照リンク数 (IaRL)
(4) ページ外参照リンク数 (IrRL)
(5) 文字数/モジュール (L/M)
(6) 情報ブロックの数/モジュール (IB/M)
(7) 子供の数/モジュール (C/M)
(8) 文字数/ファイル:平均 (L/F_Avg)
(9) モジュール数/ファイル:平均 (M/F_Avg)

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 (Avg)，標準偏差 (SD)，変動係数 (RSD) についてそれぞれ計測する．

本研究では ±2 シグマを越える値を異常値とした．これは, 構造的な欠陥がない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．

各項目の平均, 標準偏差を表 1 に, 閾値と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 個であるが, 重複を取り除くと 48 個となった．

これら異常な値を示したマニュアルを調査したところ, 複数のファイルに分割するべきであると思われるマニュアルや, 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があったり, 何らかの構造上の不具合が確認できた．

4 考察

「階層の深さ」や「文字数/モジュール: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効力を持つものと考える．

しかし,「ページ外参照リンク数」では読み手の便宜のために各モジュールにリンクをつけているだけで, 構造的欠陥を抱えているとはいえないマニュアルが異常値を示した．この項目については, 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という最初の仮定があてはまらないものと思われる．

5 まとめ

評価基準に基づいて計測項目を定義し，多数の HTML マニュアルに対して各項目の計測と分析を行なった．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がわかった．これにより, 各計測項目は文書構造の品質評価に有用であると考えられる．

今後の課題としては，HTML 以外の構造化文書フォーマット, 特に XML への対応が考えられる．

参考文献

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

[2] 谷口，川口，松下，井上: “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)．

[3] 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)．