【論文】

論文作成のためのウェブコラボレーションシステム

田崎 潔志* 深見 空斗* 松永 智揮* 天野 善一† 藤田 毅*

Web collaboration system for writing thesis

Kiyoshi TASAKI* ,Takato FUKAMI* ,Tomoki MATSUNAGA* , Zenichi AMANO†
and Takeshi FUJITA*

Abstract- Wiki is a kind of web contents management system that allows users to easily add and edit contents and is especially suited for collaborative writing.

SmartDoc is a tool for creating documents based on XML that can convert SmartDoc format document to HTML4.0 format and LaTeX format.It is suited for making technical and scientific documentation.

We developed a system called wiki2sdoc that can convert Wiki documents to SmartDoc documents. So we can use both Wiki and SmartDoc features to do collaborative design and discussion in our laboratory.

Furthermore, we added the following features to the system.

- The ability of mathematical expression like LaTeX.
- Marking descriptions for authors, figure captions and bliographies.
- Indexing automatically using term recognition and extraction.

Especially we adopted a method based on single-noun statistics calculated with single-noun bigrams.

Keywords: Wiki, SmartDoc, document generator, collaboration system, automatic indexing

1 はじめに

Web コンテンツ管理システム Wiki は Web サーバ上で動作し、ブラウザさえあれば誰でも簡単に文書を作成し、編集できる。Wiki の最大の特徴は複数人でひとつの文書を作り上げていく点である。通常、複数人で共同して文書を用意する場合はデータの共有、交換などを行う必要がある。これに対して Wiki で作業する場合は、Web サーバ上で文書の一括管理が行われるため、文書の共有、交換の必要はない。また、Web サーバ上で動作するためインターネットが使用できる環境であれば時と場所を選ばず更新できるという特徴もある。しかし残念ながら Wiki は Web ページ作成を目的としているため、論文のような技術文書を作成するための十分な機能は備えていない。

Wiki とは別に論文を作成する方法として、整形された文書を生成する文書生成システム SmartDoc がある。SmartDoc 文書は技術文書向けに設計された XML に基づく文書形式であり、HTML や LaTeX 形式の出力が得られるため，論文の記述に適している。このように様々な形式の文書に変換できるため、印刷用文書や Web ページをひとつの文書から作ることができる。

Wiki と SmartDoc を組み合わせると簡単で効率の良い論文作成が可能になる。この目的で我々の研究室で Wiki 文書を SmartDoc 文書に変換するアプリケーションプログラム Wiki2sdoc を開発した。Wiki 文書を SmartDoc 文書へ変換することにより Wiki 及び SmartDoc 双方の利点を生かした文書作成を行うことができる。

さらに我々は Wiki2sdoc を改善するため，以下のような論文作成上の利便性の向上を図った。

- Wiki での数式表示機能を導入した。高度な数式表現能力を備えた LaTeX のコマンドを用いて数式を Wiki 文書の中に埋め込むことができる。
- 図表への説明を書くキャプション記述、著者名記述、参考文献記述などの機能を追加した。
- 自動索引付け機能。

中でも特に索引の自動生成に我々は注目した。論文には索引が記載されていることが望ましいが、論文中

*電気工学科
†工学研究科電気工学専攻

の索引とする用語は人手によって抽出しなければならない。この索引生成を自動で行うことができれば、より良い論文作成への手助けとなると思われる。

我々は Wiki2sdoc に加えて新たに自動索引生成システムを開発するため、湯本ら [6] が提案した名詞 (単名詞と複合名詞) の出現頻度と連接頻度を用いた専門用語抽出法を採用した。この中の FLR 法と MC-Value 法と呼ばれる 2 つの方法を用いて索引生成システムを作成し、Wiki2sdoc の中に組み込んだ。また抽出結果を基に抽出精度の評価も行った。

2 システムの機能と使い方

2.1 システム利用の流れ

Wiki2sdoc は Web コンテンツ管理システム Wiki のひとつである「PukiWiki」用に開発した拡張機能[1]で PukiWiki で作成した Wiki 文書を SmartDoc 文書へ変換する。Wiki 文書から他の文書形式への変換は図 1 のような流れとなる。

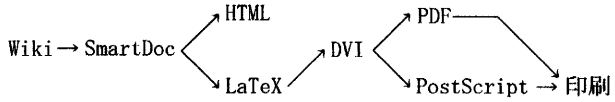

図 1: Wiki 文書から他の文書形式への変換

SmartDoc 文書は HTML、LaTeX などの文書形式に変換出来る。Wiki 文書を SmartDoc 文書に変換することは Wiki 文書をこれらの形式に変換できるようにすることである。

HTML は Web ページの作成に用いられる文書形式であり、Wiki によって作成された Web ページとほぼ同等のものであるが、図、表、数式などの参照番号や索引が追加されている。

LaTeX は学術論文に多く用いられる数式表現に優れた高品位印刷用の文書形式であるが、この文書形式はそのままでは閲覧や印刷を行うことが出来ない。閲覧や印刷を行うためには LaTeX 文書を DVI という形式に変換する必要がある。印刷の場合はさらに PDF もしくは PostScript 形式に変換して行う。

[1] PukiWiki ではプラグインと呼ばれる

2.2 Wiki 文書から SmartDoc 文書への変換

以下の手順に従い Wiki 文書を SmartDoc 文書へ変換を行う。

1. PukiWiki の下部に設置されているツールバー (図 2) の SmartDoc の部分をクリックする。

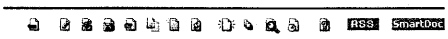

図 2: PukiWiki ツールバー

2. Wiki2sdoc の設定画面 (図 3) が出てくる。変換の際の設定ができる。

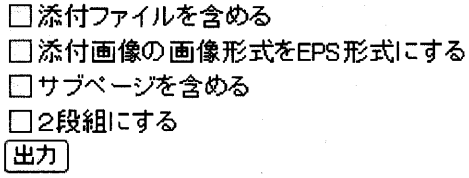

図 3: Wiki2sdoc の設定

それぞれの項目は以下のような意味を持っている。

- 添付ファイルを含める
  文書内に画像がある場合、その画像も一緒にダウンロードされる。文書内にある画像が手元にない場合でも別途ダウンロードする必要がなくなる。
- 添付画像の画像形式を EPS 形式にする
  文書と一緒にダウンロードする画像を EPS 形式に変換する。LaTeX 文書に変換する場合には EPS 形式に変換する必要がある。
- サブページを含める
  サブページ (変換するページ/で始まる名前のページ) も変換する。複数の文書を処理することができる。
- 2 段組にする
  文書を 2 段組のレイアウトにする。この設定は LaTeX 文書に対してだけ適用される。

3. 出力ボタンを押すとファイルの保存画面 (図 4) が開かれる。この画面や操作の方法はブラウザによって異なる。ブラウザの指示に従って、ファイルをハードディスクに保存する。

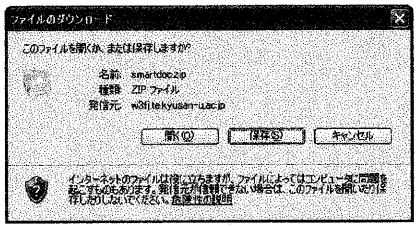

図 4: ファイルの保存画面

指定した場所に zip 形式の圧縮ファイル、smartdoc.zip が保存される。

3 システム構成

3.1 サーバのシステム構成

サーバ上に構築される論文作成システムは「PukiWiki」及び拡張機能によって構成される。

- PukiWiki, Wiki2sdoc, 数式表示, 著者名表示
- キーワード抽出
- 茶筌

PukiWiki は論文作成システムの中心となる Web コンテンツ管理システムである。PukiWiki の拡張機能として Wiki2sdoc を導入することで Wiki 文書を SmartDoc 文書に変換することができるようになる。また、Wiki には論文を記述するために十分な機能を備えていないため、それを補うための拡張機能として、数式表示機能、著者名表示機能を導入する。

Wiki2sdoc はキーワード抽出プログラムに Wiki 文書の内容を渡して、キーワード抽出プログラムから文書の中から索引として適切と判別された単語列を受け取った後、この抽出用語に基づいて索引付けを行う仕組みになっている。

キーワード抽出プログラムは外部プログラムとして作成したプログラムで索引付けを行うために Wiki2sdoc に組み込む必要がある。また、このプログラムを利用するためには形態素解析システム「茶筌」が必要となる。

3.2 利用者に求められる環境

論文作成システムの利用者は少なくとも SmartDoc が利用できる環境である必要がある。SmartDoc 文書を LaTeX 形式に変換する場合は SmartDoc だけでなく LaTeX が利用できる環境が必要となる。これらのソフトウェアはフリーソフトウェアとして公開されている。

4 キーワード抽出の方法とその評価

4.1 形態素解析

テキストの言語解析を行うには最初に形態素解析を行う必要がある。形態素解析は自然言語処理の基礎技術のひとつで自然言語で書かれた文を意味のある最小単位の形態素の列に分割し品詞を判別することである。形態素解析には奈良先端科学技術大学院大学松本研究室で開発された日本語形態素解析システム「茶筌」を使用した。索引の対象となるもののほとんどは名詞である。形態素解析されたものの中からそれ以上分割出来ない名詞の最小構成単位である単名詞、複数の単名詞で構成される複合名詞を候補として抽出する。

4.2 抽出用語に対するスコア付け

ここでは我々が採用した専門用語抽出法 [6] の原理を述べる。

索引語は文章内において重要な語である。したがって、文章中に表れたすべての名詞が索引語ではない。そこで我々は以下の方法を用いて名詞にスコアを付けることで、重要な語であるかを判別した。

抽出すべき資料に現れる単名詞 N が連接する状況、すなわち単名詞バイグラムを一般に以下のように表わす。

$$[LN_i\ N](\sharp L_i)\ \ [N\ RN_i](\sharp R_i)$$

$LN_i(i = 1, \ldots, n)$ は、単名詞バイグラム $[LN_i\ N]$ における単名詞 N の左方に連接する単名詞を、$RN_i(i = 1, \ldots, m)$ は単名詞バイグラム $[N\ RN_i]$ における単名詞 N の右方に連接する単名詞を表わす。また、() 内の $\sharp L_i(i = 1, \ldots, n)$ は N の左方に連接する単名詞 LN_i の頻度を、$\sharp R_i(i = 1, \ldots, n)$ は N の右方に連接する単名詞 RN_i の頻度を表わす。単名詞バイグラム $[LN_n\ N]$

や $[N\ RN_m]$ はより長い複合名詞の一部分である場合もある。

単名詞バイグラムを特徴付ける要因として、頻度情報 $\sharp L_i$、$\sharp R_j$ が挙げられる。次式で表わせるような頻度情報 $\sharp LN(N)$、$\sharp RN(N)$ の総和をとることで、単名詞バイグラムに重みを付ける事ができる。

$$\sharp LN(N) = \sum_{i=1}^{n} (\sharp L_i) \tag{1}$$

$$\sharp RN(N) = \sum_{j=1}^{n} (\sharp R_j) \tag{2}$$

ここで $\sharp LN(N)$, $\sharp RN(N)$ は、それぞれ N の左方、右方に連接して複合名詞を形成する全単名詞の頻度である。

単名詞の左右に連接する単語の頻度を用いたスコアの定義を行ったが、これらの左右のスコアを組み合わせて単名詞そのもののスコアを定義する必要がある。一方、複数の単名詞から構成される複合名詞にもスコア付けの定義をする必要がある。複合名詞の重要度は、複合名詞の構成単名詞数に依存するという考え方と、依存しないという二通りの考え方ができるが、複合名詞の構成単名詞数が複合名詞の重要度を決定するという根拠はない。

そこで、単名詞 $N_1, N_2, \ldots, N_L$ がこの順番で連接した複合名詞を CN と表わし、前節のスコア関数から求められた各単名詞の左右のスコアの平均をとることとする。これによって複合名詞の構成単語数に依存しないスコア関数を定義することが出来る。さらにここで前節のスコア関数の左方スコア $\sharp LN(N)$ を $FL(N)$、右方スコア $\sharp RN(N)$ を $FR(N)$ と表わすことにする。単名詞バイグラムを用いたスコア付けの方法として頻度情報 $\sharp L_i$、$\sharp R_j$ を利用する以外にも、単名詞 N の左方または右方にくる単名詞の種類の異なり数を利用する考え方があるためである。

CN のスコアの平均には相乗平均を採用し、スコアが 0 になるのを避けるために次式で CN のスコア $LR(CN)$ を定義する。

$$LR(CN) = \left\{ \prod_{i=1}^{L} (FL(N_i)+1)(FR(N_i)+1) \right\}^{\frac{1}{2L}} \tag{3}$$

$LR(CN)$ は、形態素解析後に抽出された用語候補集合内における統計的性質より導き出した。一方、用語候補には純粋に文書内で出現した頻度という別種の情報が存在する。つまり、前者が用語候補集合における構造の情報、後者が文書内における個別用語候補の統計的性質であり、これらは別種の情報として扱うべきである。そこで、用語候補である単名詞あるいは複合名詞が単独で出現した頻度を考慮すべく、式 (4) を補正して、次のように $FLR(CN)$ を定義する。

$$FLR(CN) = f(CN) \times LR(CN) \tag{4}$$

ここで、$f(CN)$ は候補語 CN が単独で、つまり他の複合名詞に包含されることなく出現した頻度である。

4.3 MC-Value 法によるスコア付け

単名詞バイグラムによらないスコア付けの方法として、C-value がある。C-value は次式で表される。

$$\begin{aligned} C-value(CN) &= (length(CN)-1) \\ &\times \left(n(CN) - \frac{t(CN)}{c(CN)} \right) \end{aligned} \tag{5}$$

この式において、CN は複合名詞[2]、$length(CN)$ は CN の構成単名詞数、$n(CN)$ は本文における CN の出現回数、$t(CN)$ は CN を含む複合名詞の出現回数、$c(CN)$ は CN を含む複合名詞の異なり数である。

しかし、この式では CN が単名詞の場合 $length(CN) = 1$ となりスコアが 0 となってしまうため、適切なスコア付けが出来ない。そこで湯本らは C-value の定義を次式のように変更し、Modified C-Value 略して MC-Value と名づけた。

$$\begin{aligned} MC-Value(CN) &= length(CN) \\ &\times \left(n(CN) - \frac{t(CN)}{c(CN)} \right) \end{aligned} \tag{6}$$

4.4 索引用語の選定

スコア付けされた用語の上位どこまでを索引用語として選定するのか決める必要がある。ここまでの過程で候補用語の集合のそれぞれの要素には用語のスコア情報が付加される。用語のスコアは文書中での出現頻

[2]ここでは nested collocation と呼ばれる。

度など文書の内容によって決定されるため、用語のスコアを考慮し、ある値を閾値としてこの値を超える用語を索引用語として選定することで文書の内容を考慮した選定が行える。閾値の算出方法のひとつとして用語全体のスコアの平均値と分散を考慮して 50 を平均とする偏差値から求める方法がある。閾値は次式から算出される。

$$Threshold = \frac{(devi - 50) \times sdevi}{10} + average \quad (7)$$

ここで、$devi$ は閾値とする偏差値、$sdevi$ は用語全体のスコアの標準偏差、$average$ は用語全体のスコアの平均値である。

索引用語を選定する場合は平均値よりやや大きい値を閾値とするため、ここでは $devi = 55$ とした。

4.5 抽出実験

スコア付けの方法として FLR 法、MC-Value 法を用いて、いくつかの図書の索引用語抽出を行い、抽出を行った図書の索引と抽出結果を比較して、抽出精度を評価した。

抽出された索引用語が抽出を行った図書の索引用語を正解索引語とし、この中のいずれかと一致としたものを正解とした。正確さを示す「適合率」と正解数を示す「適合率」を用いた。

$$適合率 = \frac{抽出用語中の正解数}{抽出用語の総数} \quad (8)$$

$$再現率 = \frac{抽出用語中の正解数}{正解索引語の総数} \quad (9)$$

表 1: MC-Value を用いた評価

	適合率	再現率
実験 1	0.073	0.120
実験 2	0.148	0.290
実験 3	0.051	0.091
実験 4	0.194	0.245
実験 5	0.161	0.278

表 2: FLR を用いた評価

	適合率	再現率
実験 1	0.143	0.120
実験 2	0.176	0.194
実験 3	0.043	0.023
実験 4	0.190	0.163
実験 5	0.176	0.222

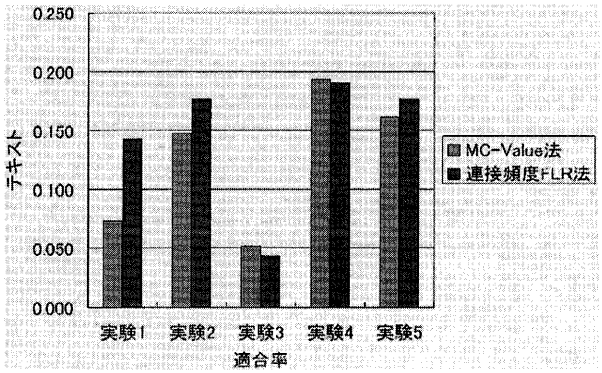

図 5: 適合率

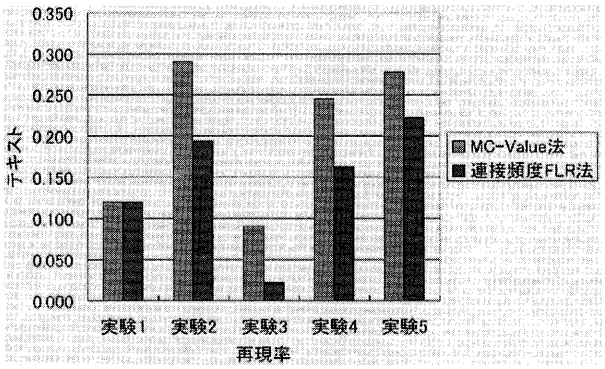

図 6: 再現率

この結果を見る限りでは、適合率はFLRの方が高くなり、再現率はMC-Valueの方が高くなっている。しかし、処理する文書によって結果が異なるため、どちらの方が優れているかは判断できない。

5 現状と問題点

現在、我が研究室ではWikiを利用して論文を作成している。この際にはWiki文書からPDF文書まで変換し、印刷を行った。WikiからPDFまでは多くの変換過程があるものの、過程の途中で文書の変更や修正を行うことなく利用することができた。

現在のシステムにまだ搭載されていない機能として以下のようなものが考えられる。

- 数式、参考文献への参照

数式や参考文献の参照にも図や表と同様に割り振られた参照番号を記述する必要があるが、現在のWiki2sdocはこの機能を備えていないため、数式や参考文献の参照はSmartDoc文書に変換した後、手作業による修正が必要となる。

さらに、自動索引生成システムにも以下のような改善すべき点がある。

- 索引用語の抽出精度

一般的に索引用語は筆者が重要とする語句であるため、その箇所は強調されていたり、各章のタイトルに含まれていることが多い。筆者が重要とする語句であると同時に読者が検索の対象として必要とするものでもある点も見逃すことができない。また、我々は名詞（単名詞と複合名詞）のみを抽出していたが、実際には名詞及び助詞からなる複数の語の組み合わせによって構成された「名詞句」が索引用語となる場合もある。したがって、抽出精度を高めるにはこれらの点を考慮した用語抽出が必要となる。

- 索引付けの速度

現在のWiki2sdocによる処理時間のほとんどは索引付けであり、その中でも索引用語候補に対するスコア付けは多くの時間を費やしている。用語抽出プログラムは開発が容易なインタプリタ型言語Rubyで作成されているため、変換中に不具合が発生することはないものの動作速度が遅い。動作速度を向上させるにはC++などのコンパイラ型言語でプログラムを作成する必要がある。

6 おわりに

Wikiに対して論文作成に必要な新機能の導入や参考文献の指定などの機能拡張を行うことで、論文をWiki上で作成できるようになった。これはWebコンテンツ管理システムによって論文が管理を行えることでもある。

まだ不足する機能も多々あるが今後、実際に運用しながら必要な機能を加えていく必要がある。また、専門用語抽出にあたっての抽出精度の評価を行ったものの、今回行った抽出精度の実験と評価はまだまだ不十分であり、今後も実験、検討を続ける必要がある。

Wikiによって論文作成を行うことは複数人での論文の作成や管理を容易にするだけでなく、作成された論文の中から必要な論文だけを検索により見つけ出すこともできるため、論文を参照する場合の利便性も向上する。

参考文献

[1] 結城 浩. 結城浩のWiki入門：YukiWikiではじめるみんなで作るWebサイト. インプレス, 第1版, 2004.

[2] 秋山 智俊. 恋するプログラム：Rubyでつくる人工無脳. 毎日コミュニケーションズ, 第1版, 2005.

[3] 浅海 智晴. XML SmartDoc公式リファレンスマニュアル. ピアソン・エデュケーション, 第1版, 2002.

[4] 天野 善一, 大西 孝史, 甲木 宏, 高尾 綾彌. 卒論作成を支援するための文書生成システム. 九州産業大学電気工学科卒業論文, 2004.

[5] 天野 善一, 藤田 毅. 卒論作成を支援するための文書生成システム. 九州産業大学 情報処理センター『COMMON』, Vol. 25, , 2005.

[6] 湯本 紘彰, 森 辰則, 中川 裕志. 出現頻度と連接頻度に基づく専門用語抽出. 自然言語処理, Vol. 10, No. 1, 1 2003.