

構造化文書処理におけるユーザ別未確認部分明示機構

2S-3

河野章博 長島正明 山川正

キヤノン株式会社 情報メディア研究所

1 はじめに

近年、紙などのハードコピー上で扱われていた文書情報を、情報の電子化、オンライン化の流れから、端末上でそのままソフトコピーとして扱うというような要求が出てきている。また、グループウェアや共同執筆などが注目されているが、文書情報をソフトコピーとして扱うためには、ソフトコピー上で情報のやりとりができる機能を取り込む必要がある。本研究では、今回、構造化文書処理におけるユーザ別未確認部分明示機構について提案し、その有効性について検証した。

2 文書更新管理

我々は、構造化文書を、文書の単位ではなく、文書に含まれる文書要素及び文書要素の内容を単位として文書情報を扱う、DoD(Document Oriented Database) 環境の開発を行なっている。ここで言う文書情報とは、文書要素及び文書要素内容を単位として構成され、スタイル情報を付加することで最終的に目に見える文書の形として得ることができる情報群のことである。本研究のユーザ別未確認部分明示機構は、このデータベースの最新情報のビューとして見える文書を見ながら、過去に参照した時点からの変更部分を明示するものである。

現在、文書のバージョン管理などを行なう更新管理を行なうようなものがあるが、これらは文書情報更新の度に昔の文書に戻すための手続きである reverse delta を保持する必要がある。このため、長期間に渡って文書を管理する場合、つまり膨大なバージョンが存在する場合には、reverse delta も膨大になる。このため、バージョン管理機構を未確認部分明示に利用しようとしても、以下の問題がある。

第1に、ユーザの参照時点のバージョンを記録

Presentation Capability of User-Specific Unnotified Document Elements in a Document Database

Akihiro KOHNO, Masaaki NAGASHIMA, Tadashi YAMAKAWA

Media Technology Laboratory, Canon Inc.

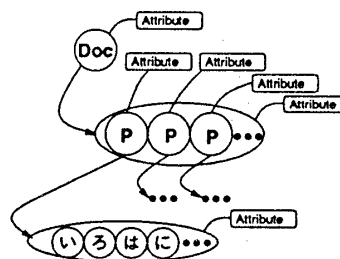


図 1: DoD/UMM 文書モデル

してにおいて、現在の最新バージョンの文書と比較して、変更部分を抽出しようとする、バージョンの隔たり分だけ、古いバージョンの文書を再生する手続きが必要になってしまうこと。第2に、参照時点の文書が再生されて現在の文書と比較されたとしても、文書情報のある部分が移動したり、大量に挿入削除が繰り返されていた場合、実際にどの部分に変更されたのか、抽出できないことである。

3 DoD/UMM 機構

本研究では、前述の問題を解決するために、再生手続きなどのための、時間及び記憶容量が必要なく、変更部分を限定できるような機構を提案する。

本研究の DoD/UMM (Updating Management Mechanism) では、ネストを持つ構造化文書情報で利用でき、構造を利用して変更部分の絞り込みを行う。また、ユーザ毎に異なる参照時点からの変更部分が、現在の最新文書情報を用いるだけで明示できるように以下の機構を用いた。

すなわち、我々の DoD の特徴である、OoD_{ET}[1] で提案した文書モデルを用い(図1)、文書要素と文書要素内容を分離させた。そして、更新された時刻情報のみを文書更新情報として、各文書要素及び文書要素内容毎に属性として持たせることによって、文書更新情報管理を行う。

DoD/UMM では、文書要素の更新を要素の状態更新時間、文書要素の内容の更新を要素の形態更新時間、という属性で扱う。これらの更新時間属性は、文書情報が編集される度にその時刻に設

	地理対象の親の形態	地理対象の親の形態	地理対象の形態	地理対象の形態	地理対象の中の全ての値
要素	挿入	nop	nop	update	nop
	削除	nop	update	nop	nop
文字	挿入	nop	nop	update	update
	削除	nop	update		

表 1: 属性値

定されながら変化する。

一方、ユーザが文書を参照した際は、その時点での文書情報を確認したと考え、参照時点の時刻を参照時間情報としてユーザ別に保持する。次回参照時には、この保持されている参照時間よりも新しい更新時間情報を持つ文書要素、及び文書要素内容が、そのユーザにとって未確認の文書情報となっている。

実際に未確認部分を抽出し、表示する機構について述べる。DoD/UMM では、以下の更新時間設定法を採用した。

- 1) 文書要素が挿入された場合は、その文書要素の状態更新時間属性に挿入時刻をセットする
- 2) 文字が挿入された場合は、その文字を子要素として持つ親文書要素の形態更新時間属性に挿入時刻をセットする
- 3) 文書要素及び文字が削除された場合は、その文書要素あるいは文字を子要素として持つ親文書要素の形態更新時間属性に削除時刻をセットする
- 4) 文書要素が削除され、カットバッファなどに保持される場合、その文書要素の状態更新時間属性、及びその中に含まれる全ての更新時間属性を保持する
- 5) 文書要素が複写され、カットバッファなどに保持される場合、その文書要素の状態更新時間属性、及びその中に含まれる全ての更新時間属性を複写時間にセットする

以上により、表 1 に示したような属性が付けられる。表中 nop は no operation の略である。

ユーザ別参照時間よりも新しい更新時間属性を持つ文書要素、あるいは文書要素内容を網掛け表示するようにした UMM 表示例を、一つの文書要素の内容を例に図 2 に示す。

この表示では以下のことを示すことが可能で

		新しい	古い
新しい	要素形態新	This content is new. <new-elm> new-content </new-elm> end of the content. A	This content is new. <old-elm> new-content </old-elm> end of the content. B
	要素形態旧	This content is new. <new-elm> old-content </new-elm> end of the content. C	This content is new. <old-elm> old-content </old-elm> end of the content. D
古い	要素形態新	This content is old. <new-elm> new-content </new-elm> end of the content. E	This content is old. <old-elm> new-content </old-elm> end of the content. F
	要素形態旧	This content is old. <new-elm> old-content </new-elm> end of the content. G	This content is old. <old-elm> old-content </old-elm> end of the content. H

図 2: UMM 表示

ある。

- 文書を書き足したような場合 (図 2 の A)
- 文書の構造を変えずに、文書要素の内容の文字のみを変更したような場合 (図 2 の B, D, F)
- 文書要素内容に、文書要素 (new-elm) が挿入された場合で、挿入された文書要素が Cut&Paste, つまり移動された場合 (図 2 の C, G)
- 文書要素内容に、文書要素 (new-elm) が Copy&Paste で挿入された場合や、移動された文書要素内容が変更された場合 (図 2 の A, E)
- 何も変更のなかった場合 (図 2 の H)

4 まとめ

文書要素と文書要素内容を区別し、それぞれに更新時間属性を持たせることによって、容易にユーザ別の未確認情報を明示する機構を実現できる。また、DoD/UMM 機構で規定した属性の付け方によれば、高々更新時間しか情報がなくとも、ユーザ別で、移動の行なわれた文書要素や、変更の行なわれた文書要素を明示することもできた。

参考文献

- [1] 河野章博ほか：内容の部分共有を許す文書データベースにおける更新機構, 情報処理学会第 47 回全国大会 (1993)