I－2 電子納品に向けた文書同一性判定システムの研究開発

Research of Document Data Identity Verification Component for Electronic Delivery

池辺正典[1]・田中成典[2]・古田均[2]・中村健二[3]

Ikebe Masanori, Tanaka Shigenori, Furuta Futoshi, Nakamura Kenji

抄録： 土木・建築分野における電子納品は，国土交通省の主導により，標準規格が策定され実施体制が整いつつある．しかし，実際の電子納品の現場では，受発注者間のデータ交換の時に同一性保証の問題や類似データの混在による作業コストの増大が問題となっている．現在，これらの問題を解決するために専用のアプリケーションで文書処理が行われているが，課題を根本的に解決する手法ではない．

そこで，本研究では，様々なデータ形式で作成される文書データに対して，レイアウト解析を行い，構造情報を作成した後に，自然言語処理を用いて意味情報を解析することで文書間の差分情報を抽出し，同一性判定を行う．さらに，詳細な差分情報を提示することで，文書管理を効率的にするシステムの研究開発を行う．

Abstract: In electronic delivery of engineering works and construction field, the standard is settled and the base is adequately maintained by the initiation of the Ministry of Land, Infrastructure and Transport. But, on the actual electronic delivery, identity verification and coexisting similar data between order and ordering suppliers increase the work cost. And a special application doing document processing is introduced as means to solve these problems. But, it is not a technique for fundamentally solving the problem.

In this research, the layout of the document data made by a variety of data forms was analyzed, and structural information was made. The substantial difference between documents is extracted by analyzing meaning information using the natural language processing, and the identity is verified. In addition, document management is made efficient by presenting detailed difference information.

キーワード： 電子納品，文書処理，レイアウト解析，自然言語処理，同一性判定

Keywords : Electronic Delivery, Document Processing, Layout Analysis, Natural-Language Processing, Identity Verification

1．はじめに

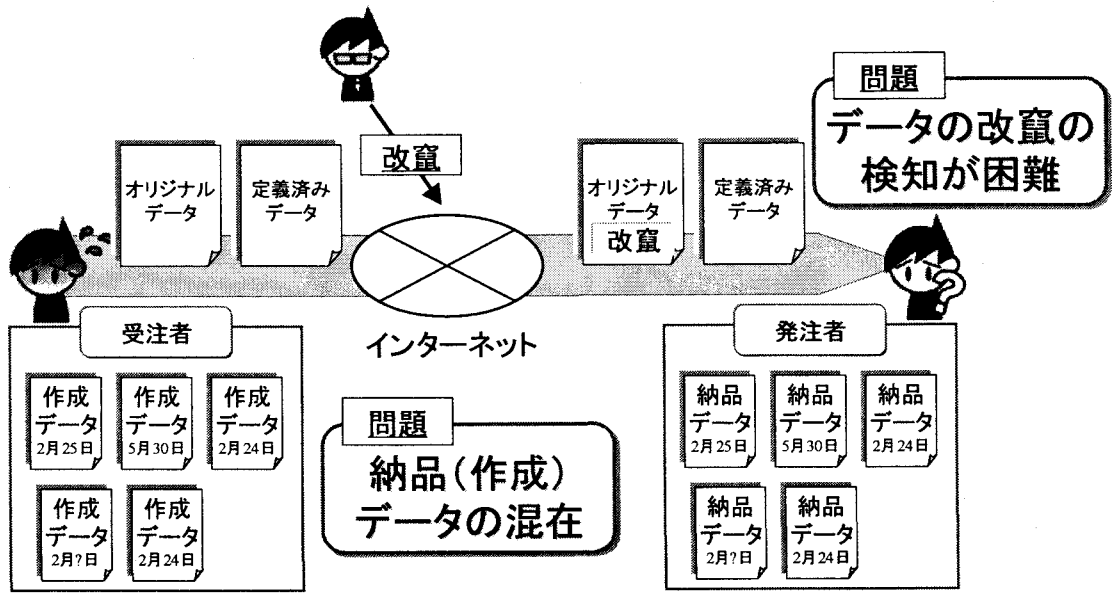
近年の情報技術の発達に伴って，電子データを用いた商取引が社会に広く普及している．こうした流れの中で，官公庁では，電子納品への対応や，標準規格の策定に積極的な取組みが行われた．電子納品が行われることで，「ペーパーレス・省スペース化」，「事業執行の効率化」，「品質の向上」などの効果が見込まれる．土木・建築分野においては，国土交通省の直轄工事で，2001 年から電子納品が実施された．電子納品の実施当初は，工事の契約予定金額に制限があり，3億円以上が対象となった．その後，契約予定金額の制限は，年々低下し，2004 年には全ての工事がその対象となった[1)]．また，全国の自治体における電子納品の実施状況は，2002 年度調査では，56.0%であったのに対し，2003 年度調査では 63.8%となり，自治体の電子納品への関心が高まっていることが分かる[2)]．このように，発注者側となる官公庁では，電子納品の普及に向けた動きが活発[3)]である．一方，受注者側となる一般企業においても，電子納品への対応準備が着実に進められており，2002 年度では 79.4%が対応可能との調査結果が得られている[4)]．このように，受発注者の双方において電子納品に対する関心が高まっている．しかし，電子納品が普及するにつれて，実際の工事を行う現場では，電子納品特有の様々な問題が表面化している．

実際の工事において，電子納品や電子媒体による受発注者間でのデータ交換は，受注者が，発注者にデータ形式が定められていないオリジナルデータとデータ形式が定められている定義済みデータの2種類のデータを納品する．これらのデータには，作成日付が付加される．これが，発注者において，納品データとして蓄積される．このような作業が，日常的に行われるようになると，以下のような問題(図－1)が発生する．

① オリジナルデータの形式が Microsoft Word などの変更可能なデータ形式であった場合，容易に改竄される危険性があり，データの改竄を検証することが困難である．

② 受発注者間で複数回にわたって修正が繰り返され

1 ：学生会員 情修 関西大学大学院 総合情報学研究科
(〒569-1095 大阪府高槻市霊仙寺町 2-1-1, Tel :06-6309-9696, E-mail : ikebe@kansai-labo.co.jp)
2 ：正会員 工博 関西大学 教授 総合情報学部
3 ：学生会員 情学 関西大学大学院 総合情報学研究科



図－1　電子納品における問題

る場合，データのバージョン管理が適切に行われていないために，類似データが混在し，データの巻き戻りが発生する[5]．

③　複数人で作業を行う受注者の場合，多くの類似データが作成されるため，発注者に提示する適切なデータの判断には，多大な人的コストを要する．

これらの問題を分類すると，セキュリティ的な問題とデータ管理に関する問題に分類することができる．

セキュリティ的な問題については，電子文書の原本保証として，完全性確保，機密性確保，見読性確保，存在証明，長期安全保管の5つの項目が重要視される[6]．上記で述べたデータ改竄に関する問題は，完全性確保，機密性確保に該当し，データが受注者で作成されてから発注者の手に渡るまで，内容に変更が加えられていないことを保証するものである．セキュリティに対しては，暗号化通信や，電子署名を付加する形式で対応することが可能である．しかし，実際の電子納品では，受発注者間において，数回のデータのやり取りが行われるため，その都度に上記のように適切なセキュリティ対策を施すことは，費用面でも時間面でも困難である．

データ管理に関する問題については，類似データの混在が主な問題である．この問題を解決するために，実際の工事では，文書管理において適切なファイルを判別する時に，ファイルの中身を目視もしくは，文書比較ツールによって確認する必要がある．ファイルの中身を目視で確認する場合，納品されたデータや作業中のデータから，適切なデータを選別する段階で，多大な人的コストを要する．また，文書比較ツールを利用して確認する場合には，2つの問題が発生する．1つ目は，現在公開されている文書比較ツールは，データ内のテキストを対象としてものが多く，レイアウトや文章の意味内容などの判別が困難な点である．このため，レイアウト情報を含む文書を従来の文書比較ツールに適用した時は，比較対象位置がずれた時点で以降の内容が全て修正されていると認識するような非常に低い精度でしか差分情報を抽出することができない．2つ目は，複数間のオリジナルデータにおいてファイル形式が異なる場合，文書の比較には，内容情報の解析が必須となるため，比較を正常に行うことができない点である．

そこで，本研究では，電子納品において，新たなセキュリティ向上手法と異なったデータ形式間の同一性の判定と詳細な差分検出を行うことのできる手法を提案する．

2．研究の概要

（1）研究の目的

現在の電子納品における問題点として，セキュリティや類似データの混在が挙げられる．本研究では，これらの2つの問題に対して解決策を提案するものである．

セキュリティの問題を解決するためには，既存研究として，第三者認証[7]や電子署名[8]を付加する方式の研究がなされている．しかし，これらの研究は，同一のバイナリデータの改竄検知や原本保証を行うものであり，異なるデータ形式で同じ構造情報を持つ文書を判定するものではない．また，同一の拡張子を持つデータである場合も，データを作成したツールの種類やバージョンによって，異なったデータ形式となる．このため，本研究では，文書の内容情報を解析することで，異なるデータ形式であっても，同じ構造情報を持つデータを同一と判定することで，データの安全性を確保することが可能となる．

類似データが混在する問題への対策としては，電子納品ツールや文書管理ツールを利用することで解決すると

考えられる．しかし，これらはいずれも文書作成時に文書に関する説明文等の情報を付加する必要があり，既に完成している文書が対象となる場合は，その効果が期待できない．また，文書データがメール等の手段によって受発注者間で交換される場合は，文書データがこれらのツールの管理外となる．そのため，情報を付加することができず文書単体で情報の内容や改定情報を判断する必要性がある．そこで，本研究では，同一判定で差分が検出された時には，文書間の詳細な差分情報を提示する．このことにより，改定箇所の容易な特定や改定意図を類推可能な情報を利用者に提示することで，適切なデータの判断を支援することを目的とする．

（２）本論文が解決する問題

本研究において，データ交換時に異なった2種類のデータフォーマットの文書を送付することで，受信者はこれらの2文書を比較し，データの改竄が行われていないと判断することができる．この方式は，従来のセキュリティ対策とは異なった形式を取るために，従来方式と重複して利用することが可能となり，セキュリティをより高くすることが可能である．さらに，従来では比較が困難であったレイアウト情報を含む文書の比較が可能になることで，利用者の文書管理にかける労力を省力化するだけでなく，紙媒体と電子媒体が混在する施行段階においても，電子納品ツールにデータ入力を行う時に本研究を適用することで，電子化作業の労力を軽減できるなど，実際の工事に適した形で文書管理を行うことができると考えられる．

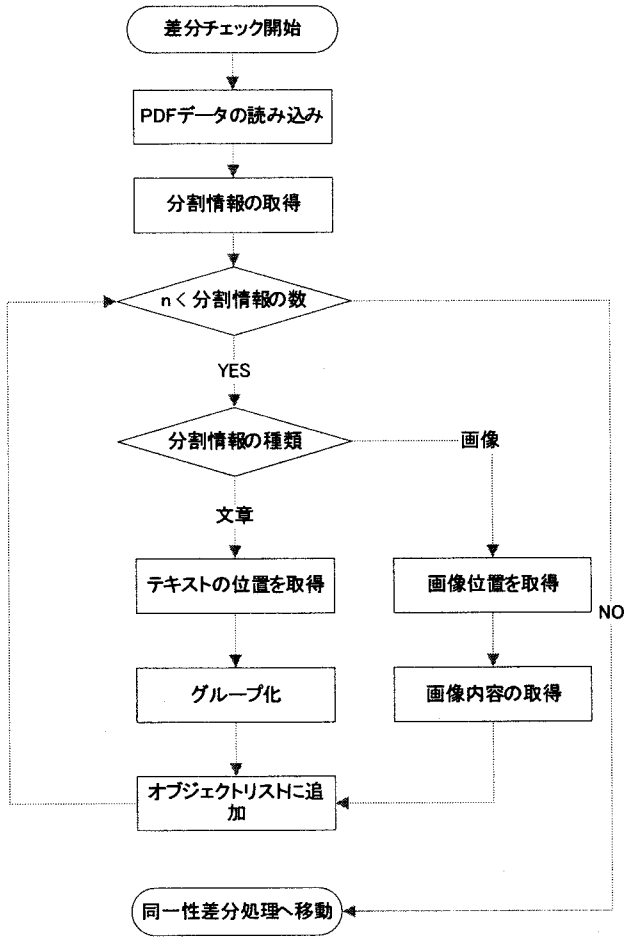
３．システムの詳細

本研究で開発するシステムは，比較を行う文書を解析し，処理単位であるオブジェクトに分割する文書解析機能と，比較を行う文書間においてオブジェクトの対応関係を抽出する同一箇所判定機能，そして，2文書においての差分結果をオブジェクト単位で詳細に表示する差分出力機能から構成される．

（１）文書解析機能

本機能では，比較対象となる変更前，変更後の2つの文書を読み込み，解析することで，文書の処理単位となるオブジェクト単位に分割する．また，オブジェクトには，レイアウト等の情報が付加されるため，これを付加情報として同時に生成する．本機能の処理の流れを図－2に示す．

レイアウト情報の抽出には，オブジェクトを大きな処理単位に分割した後に，処理単位の内部をさらに詳細に解析するという手法[9)10)]が一般的であるが，本研究で利用するデータは，情報の保持形式が最小の処理単位まで分割された形式で保存されるPDF(Portable Document Format)を基準とするために，最小の分割単位をグループ化し，より大きな分割単位を生成するという逆のプロセスを採用する．



図－2　文書解析の流れ

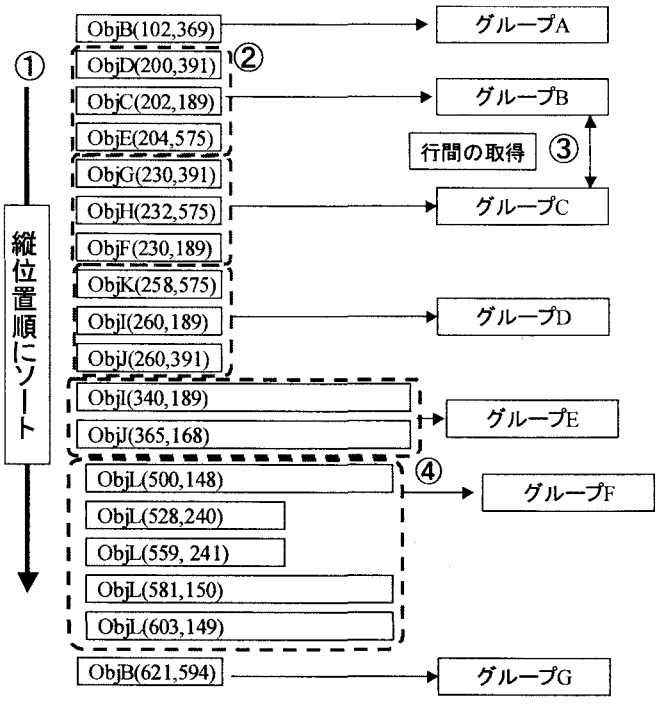
a）文書解析処理

本システムの入力データは PDF 形式の文書ファイルを対象とする．このため，PDF 以外の文書形式のデータは，PDF データに変換を行った後に本システムに入力する．

PDF 文書において，文書情報がテキストの場合は，文字単位でテキスト内容が分割されて保存される．このため，本処理では，入力データからオブジェクト生成の前段階となる分割されたテキスト内容についての情報を取得する．取得する情報は，文書の左上を原点とした時の縦位置と横位置である．また，解析内容がテキストの場合は，これに加えて，フォント種別，文字サイズ，文字色，行間サイズ等のレイアウトに関する情報とテキスト内容を取得する．そして，画像の場合は，実際の画像のバイナリデータと画像の横幅，縦幅についての情報を取得する．本処理は，「Xpdf」を改良したツール「pdftohtml」で PDF 文書のテキスト化を行った．「pdftohtml」により得られるテキストの状態は，一定間隔で表示されるテキストをグループ化した情報である．

b）オブジェクト生成処理

文書解析により得られた情報は，一定の纏まりを持つが，段落や表，箇条書き等の単位ではないため，これをグループ化する必要がある．



図－3　グループ化の手順

オブジェクトのグループ化では，まず，文書におけるオブジェクトの縦位置，横位置およびオブジェクト間の位置関係によってグループ化する．オブジェクトの位置は，オブジェクトの左上を基準とする．次に，グループの種別を表，箇条書き，段落の3種類に分類する．最後に，段落のテキスト内容を結合する．オブジェクトのグループ化の手順(図－3)を次に示す．

① オブジェクトを縦位置の順にソートする．

② 縦位置が一定範囲内のオブジェクトをグループ化する．

③ スタイル単位に基準となる行間を取得する．

④ 上下のオブジェクトを比較し，行間が基準値範囲内ならグループ化する．

上記の処理を高さが同様のオブジェクト数が0になるまで繰り返し行う．次に，グループの種別を判定する．グループ種別の判定手順(図－4)を以下に示す．

① グループ内でオブジェクトを横位置の順にソートする．

② 複数のオブジェクトがある行が2個以上続いた場合は表候補とする．

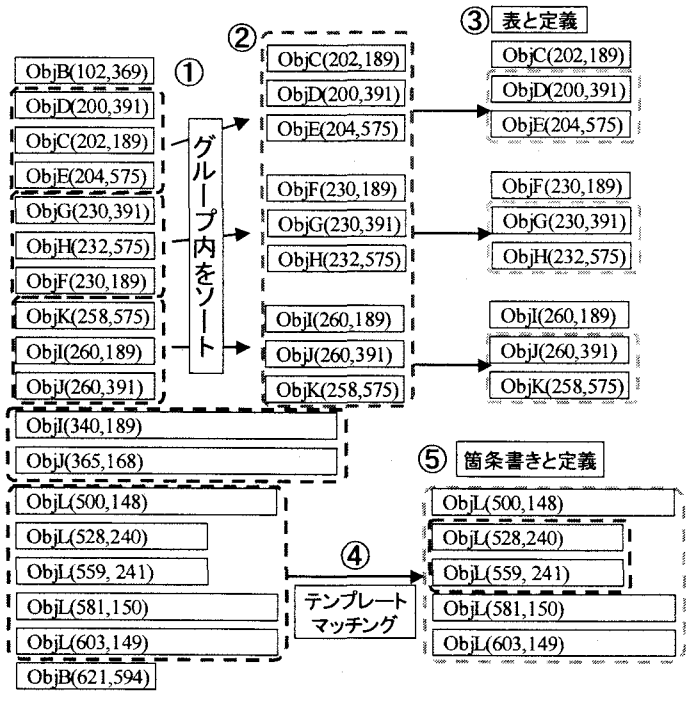
③ 同一行の2番目以降のオブジェクトにおいて横位置が同一のオブジェクトがあった場合は表と定義し，行数と列数を取得する．

④ 表以外のオブジェクトについて，行頭の文字から箇条書きインデントのテンプレートとマッチングを行い，一致した場合は，箇条書きとして定義する．

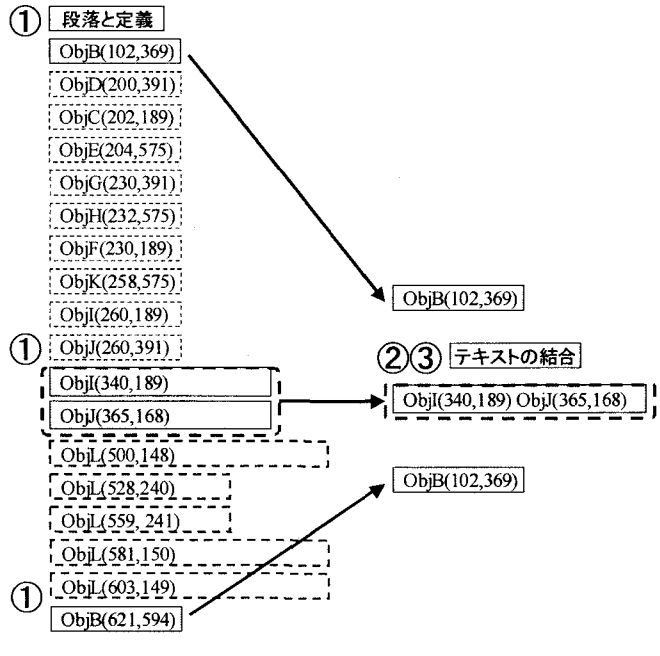
⑤ 箇条書きと定義されたオブジェクトについて，横位置の情報を元に箇条書きの包含関係と個数を取得する．

最後に，通常のテキストを段落としてグループ化する処理の手順(図－5)を次に示す．

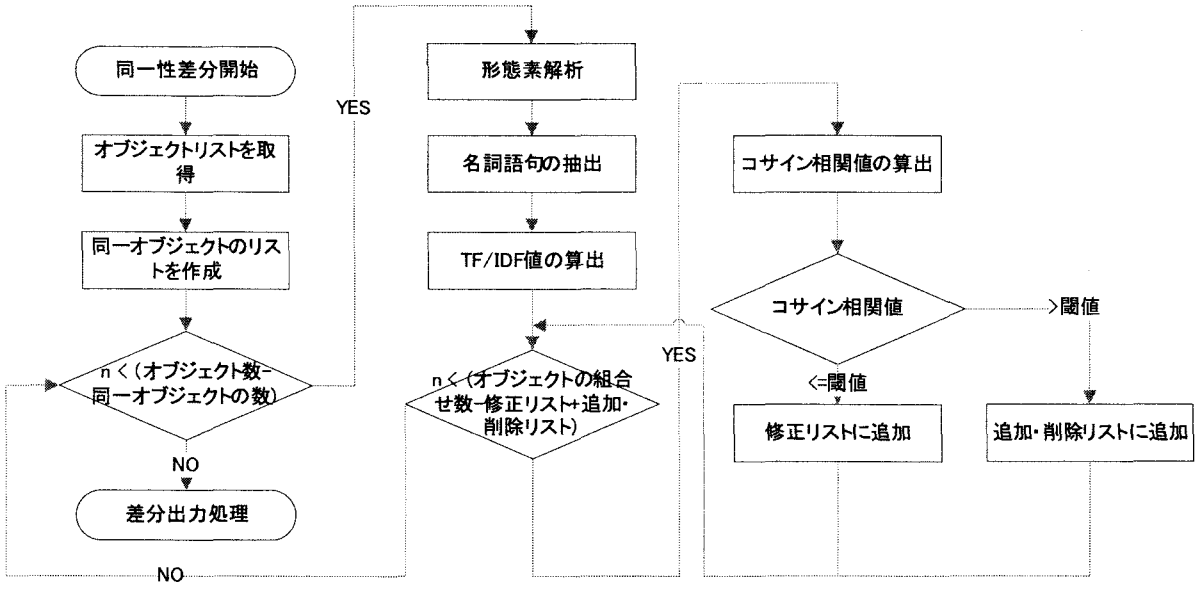
① 表と箇条書きに分類されていないオブジェクトを段落候補として定義する．



図－4　グループ種別の判定手順



図－5　段落のグループ化手順



図－6 同一箇所判定の流れ

② オブジェクトの縦位置によってグループ化されているテキストの結合を行う.

③ オブジェクトの行間情報によってグループ化されているテキストの結合を行う.

上記の処理を未分類のオブジェクト数が0になるまで繰り返し行う.

（2）同一箇所判定機能

本機能では, 変更前と変更後の文書のオブジェクト単位に分割されたデータを比較し, 同一箇所の特定を行うことで, オブジェクトの追加・修正・削除の検出を行う. 同一オブジェクトの特定では, TF-IDF (Term Frequency-Inverse Document Frequency)法による特徴ベクトルの生成を行い, 特徴ベクトルを多次元空間に展開し, コサイン相関値を算出するベクトル空間法を利用する. 文書の分類手法としては, 既方式 [11][12]以外に事例ベース手法 [13]による分類がよく用いられるが, 本処理は, 単一の文書データのみで文書の分類を行うため, 事例ベースの適用が困難である. このため, 本論文では, TF-IDF 法とベクトル空間法による手法を採用した. 本機能の処理の流れを図－6に示す.

a）同一オブジェクト判定処理

同一オブジェクトの判定では, 変更前と変更後の文書で対応するオブジェクトを取得するために, 最初に正規一致による同一リストの生成を行う. 正規一致処理は, 変更前文書を構成する各オブジェクトと変更後文書を構成する各オブジェクトの全ての組合せに対して, 完全に同一となる場合に, その組合せを同一リストに追加する.

b）特徴ベクトル生成処理

特徴ベクトル生成処理では, 同一オブジェクト判定処理で生成された同一オブジェクト以外のオブジェクトを変更オブジェクト候補として, 対応するオブジェクトの有無を探索する. 具体的な処理としては, オブジェクト単位に形態素解析を行い, その結果から名詞を抽出する. ここでは, 形態素解析システム「茶筅(ChaSen)」を用いる. そして, 形態素解析で得られた各名詞に対して TF-IDF 値を算出する. TF-IDF 値は, 式(1)のように定義する.

$$w(t) = tf(t) \times idf(N,t) = tf(t) \times \log \frac{N}{df(t)} \quad \cdots \text{式(1)}$$

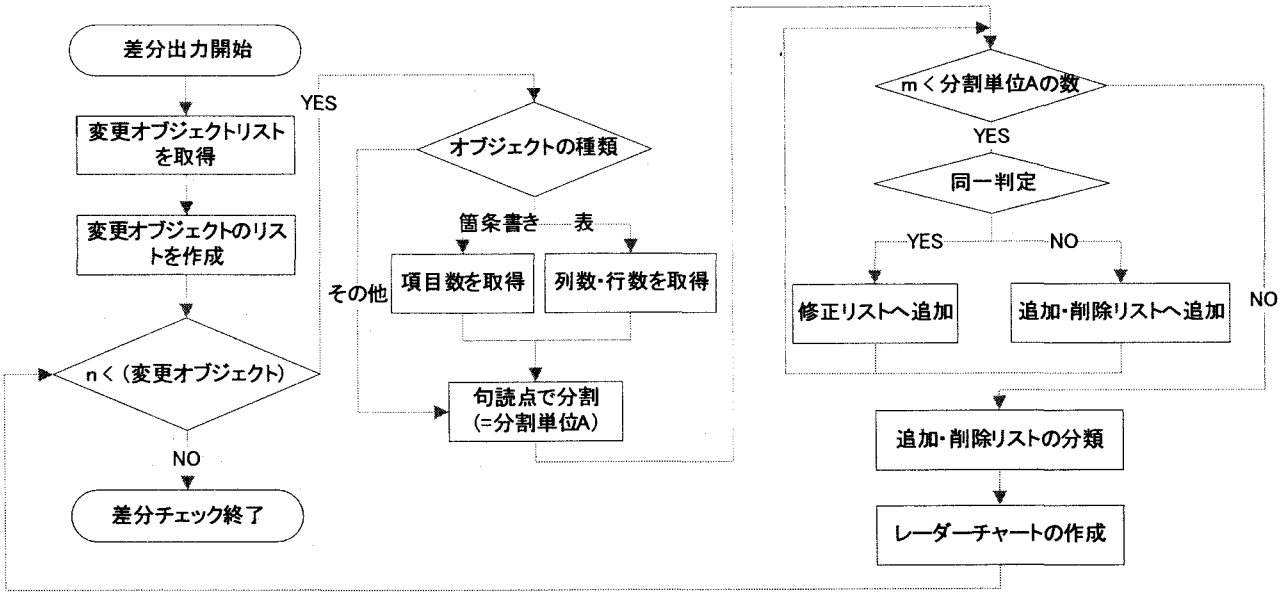
ここで, $tf(t)$ は, 解析を行うオブジェクト中に含まれる単語 t の出現回数を表す. また, N は解析文書における総オブジェクト数を示し, $df(t)$ は, 解析文書における単語 t が出現するオブジェクト数となる. この時に, オブジェクト X に含まれる名詞群を $Ox = \{Tx_1, Tx_2, Tx_3, ..., Tx_n\}$, オブジェクト X の TF-IDF 値群を特徴ベクトルとして $Vx = \{Wv(Tx_1), Wv(Tx_2), Wv(Tx_3), ..., Wv(Tx_n)\}$ と定義する.

c）相関オブジェクト判定処理

内容が修正されたオブジェクトの判定を行うために, 特徴ベクトルの算出を行った全てのオブジェクトの組み合わせについて, D 次元の仮想空間に特徴ベクトルを展開する. ここで用いる D は, オブジェクト X とオブジェクト Y を比較した時に, 式(2)で得られる集合の個数と定義する.

$$D(x,y) = (Ox \cup Oy) \quad \cdots \text{式(2)}$$

このため, 仮想空間に展開された特徴ベクトルは, $D(x,y)$ 個の要素を持つ必要がある. しかし, Vx で定義



図－7 差分情報抽出の流れ

された特徴ベクトルは，$D(x,y) \geq n$ となる n 個しか値を保持しない．このため，次元数をあわせるために，それぞれのオブジェクトの特徴ベクトルについて，値が存在しない特徴要素を0として補完する．次に，オブジェクト間の関連性を算出するために，特徴ベクトルからコサイン相関値を算出する．コサイン相関値 $sim(Vx,Vy)$ は，式(3)のように定義する．

式(3)において，右辺の分子はオブジェクト X とオブジェクト Y の特徴ベクトルの内積を表し，分母は各特徴ベクトルの原点との距離の積である．コサイン相関値は，文書の特徴ベクトル同士の比較を角度として算出するもので，0～1の範囲で結果が得られる．この結果が1に近いほど2文書の分類が近いということが言える．そして，全てのオブジェクトの組合せにおいてコサイン相関値の算出が完了すると，得られたコサイン相関値が一定の閾値を超えた組合せを修正が加えられたオブジェクトであると判定する．そして，コサイン相関値が1に近い順に，その組合せを修正リストに追加し，閾値を超えた組合せが0になるまで繰り返す．そして，残りのオブジェクトを追加・削除リストとする．

（3）差分出力機能

本機能では，同一判定処理で得られた変更オブジェクトのリストに対して，差分箇所の詳細な情報を抽出し，利用者に提示する．差分出力する情報は，全体の変更点が一目で分かるようにオブジェクトの追加・修正・削除を提示する他に，変更があった各オブジェクトに対して，詳細情報を参照することができる．また，変更意図が添付されていないケースを想定し，利用者が文書の変更意図を判断する時の手助けをする情報として，オブジェクトの意味情報をレーダーチャートで表現し，意味内容の変更を可視化する．本機能の処理の流れを図－7に示す．

a）差分情報抽出処理

差分情報の抽出では，同一箇所判定機能で抽出された変更オブジェクトリストから文の追加・削除の情報を抽出する．変更オブジェクト内の差分情報抽出処理の手順を次に示す．

① 変更オブジェクトの種類を判別する．
② 差分出力情報として箇条書きの場合は項目数，表の場合は行数と列数を取得する．
③ オブジェクト内の文章を句読点で分割する．
④ 分割単位毎に同一判定を行い，文が同一の場合は修正リスト，異なる場合は追加・削除リストに追加する．
⑤ 変更前文書のリストに存在し，変更後文書のリストで存在しない分割単位を削除，変更前文書のリストに存在せず変更後で存在する分割単位の情報を追加として，追加・削除リスト内の分割単位を分類する．

上記の処理を修正オブジェクト毎に繰り返し行い，すべての修正情報に関して差分情報を抽出する．

b）結果出力処理

$$sim(Vx,Vy) = \frac{Wv(Tx_1)\times Wv(Ty_1) + Wv(Tx_2)\times Wv(Ty_2) + \ldots + Wv(Tx_n)\times Wv(Ty_n)}{\sqrt{Wv(Tx_1)^2 + Wv(Tx_2)^2 + \ldots + Wv(Txn)^2} \times \sqrt{Wv(Ty_1)^2 + Wv(Ty_2)^2 + \ldots + Wv(Tyn)^2}} \quad \cdots \text{式(3)}$$

結果出力処理では，追加・修正・削除の件数をオブジェクト単位で表示する．また，修正されたオブジェクトに関しては，詳細情報を閲覧できる形式である．詳細情報に含まれる情報は，追加件数と削除件数，実際に追加・削除が行われた内容を文書中に明示する．結果出力処理では，これらの情報を利用者に提示するためのインターフェイスを提供する．

4．実証実験

（1）文書解析精度

a）実験方法

本実験では，文書解析機能を用いて，文書データを読み込み，細分化されたオブジェクトをグループ化し，オブジェクト構成とする．その後に，オブジェクトの種別を画像とテキストに分類を行い，また，テキストに関しては，表，箇条書き，段落の個数が正確に取得できたかを検証する．

b）実験対象

文書解析機能の精度を測定するために5つの文書データを用意した．本研究は，土木・建築分野における電子納品を対象としているために，実験に利用した文書は，全て土木関連の技術文書である．各文書の構成を表－1に示す．実験で用いる文書は，10～21 ページの文書で，含まれるオブジェクト数は，72～150 個である．

c）実験結果

5つの文書を解析した結果，グループ化により合計で636 個のオブジェクトを構成することに成功した．また，表－1に示す通り，実験に用いた5つの文書を目視で確認したオブジェクトの総数が 636 個であるために，正常にグループ化が行われたことを証明できた．さらに，オブジェクトの種別判定については，全オブジェクト 636 個を画像 47 件，表 29 件，箇条書き 26 件，段落 455 件に分類することができた．この結果においても，目視確認によるオブジェクト種別の個数と合致するために，オブジェクトの種別判定も正確に行われたと判断できる．

d）考察

本実験により，文書解析機能のオブジェクト生成機能の精度が実用に耐えうることを明らかにした．本実験では，一般的な文書での利用頻度の高い画像，表，箇条書き，段落の4つのオブジェクト種別に対して，個別にグループ化手法を提案し，有用な精度を確認することができた．また，上記4つのオブジェクトの種別を判定する手法も提案し，正確な分類が行われていることを確認した．このため，これらの種類より構成される文書は，本研究を適用することが可能であると考えられる．

（2）同一箇所判定の精度

a）実験方法

本実験では，同一箇所判定機能を用いて，2つの文章を読み込み，追加・削除オブジェクトと修正オブジェクトの判定を行った．また，同一箇所の判定基準は，相関値が0.7 以上のものとした．

b）実験対象

同一箇所判定の精度を測定するために，文書解析精度の実験で利用した文書1に対して，「文章，画像と箇条書きブロックの追加・修正・削除」，「表の行列数の変更」を行い，比較対象となる文書を作成した．各変更の詳細を表－2に示す．

c）実験結果

実験の結果，48 個の同一オブジェクト，25 個の追加オブジェクト，37 個の修正オブジェクトと7個の削除オブジェクトを正確に認識することができた．

d）考察

同一判定内容の，修正リストに追加したオブジェクトの組合せについて，相関値の範囲は 0.716～0.999 である．これに対して，追加・削除リストに追加したオブジェクトの組合せで最大となる相関値は 0.535 である．そして，相関値の分布を見た場合，0.5～0.7 の範囲について，点の密集率が比較的低かったことを確認できた．このため，相関値0.7を分割の基準としたが，この数値が適切であったことが確認できた．

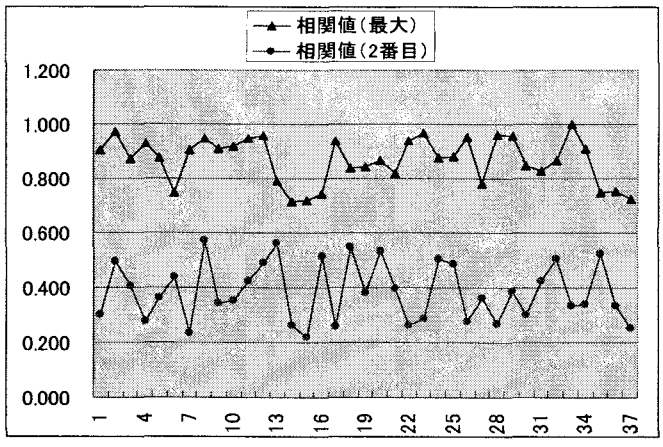
また，修正リストのオブジェクトの組合せにおいて，各オブジェクトの相関値の最大と2番目の値を比較すると，相関値の差は平均で0.484となった．相関値の差を図－8示す．以上の結果から，意味情報によって同一と判定したオブジェクトは，他のオブジェクトと意味的に大きな違いがあることを検出できた．

表－1　文書解析対象の内部構成

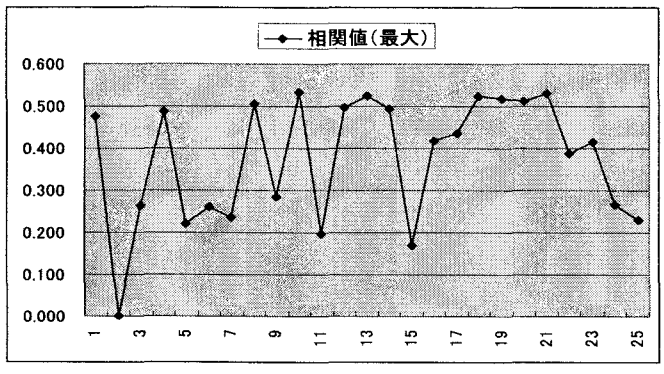
文書	総ページ数	画像	表	箇条書き	段落
文書1	17	12	1	4	86
文書2	16	8	16	8	101
文書3	10	3	10	2	57
文書4	15	10	2	3	84
文書5	21	14	0	9	127
計	79	47	29	26	455

表－2　比較対象文書の変更数

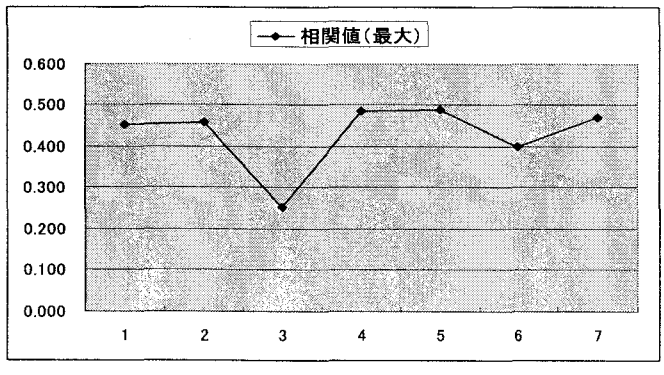
オブジェクトの種類	追加	変更	削除	同一
文章(段落)	21	33	6	44
表	1	1	0	2
箇条書き	3	3	1	2
計	25	37	7	48



図－8 修正オブジェクトの相関値



図－9 追加オブジェクトの相関値



図－10 削除オブジェクトの相関値

さらに，追加リストの相関値を図－9，削除リストの相関値を図－10に示す．追加・削除リストに含まれるオブジェクトは，他のいずれのオブジェクトとの相関も低く，最大でも 0.535 であった．このため，追加・削除を修正と判断する過剰認識は発生せず，正確な件数を認識することができた．

5．おわりに

本研究では，文書内のテキスト情報を意味的に解析することで，これまで困難であったデータ形式の異なる電子文書の比較を可能にした．また，実証実験により，比較精度が有用な水準であることも確認することができた．それに伴い，本研究を利用することで，電子納品における受発注者間のデータ交換時に文書の同一性を保証することでセキュリティを向上するだけでなく，文書の差分箇所を詳細に提示することで，受発注者間の修正作業を円滑化し，文書管理の効率化を実現することができる．

しかし，電子納品では，情報活用があまり検討されていない等の問題[14]が依然残されている．このため，本研究により得られた成果から発展研究を進めることで，納品データの意味情報を加味した検索システムを構築する等の展開により，情報活用のフェーズにおいても有用となる研究を進める予定である．

参考文献

1) 国土交通省：電子納品運用ガイドライン(案)，2004 年 10 月．
2) 社団法人日本土木工業協会：情報化実態調査報告書，2002 年 12 月．
3) 松本喬：工事成果品の電子納品における現状と問題点，土木技術，Vol.58，No.7，pp.34-41，2003 年 7 月．
4) 日本土木工業協会：電子納品対象工事実施状況調査結果，2003 年 10 月．
5) 奥谷正，有冨孝一：電子納品情報を活用した業務改善(BPR)に関する研究，土木技術資料，Vol.45，No.3，pp.38-43，2003 年 3 月．
6) 総務省：インターネットによる行政手続きの実現のために，共通課題研究会報告書，2000 年 3 月．
7) 吉川信雄，小谷誠剛：電子文書保証ソリューション-原本性保証・長期保証-，FUJITSU，Vol.55，No.1，pp.68-73，2004 年 1 月．
8) 金井洋一，谷内田益義，小川雅也：原本性保証電子保存システム(TrustyCabinet)の開発，Ricoh Technical Report，No.26，pp.97-103，2000 年 11 月．
9) Seong-Whan Lee，Dae-Seok Ryu：Parameter-Free Geometric Document Layout Analysis，IEEE Transactions on Pattern Analysis and Machine Intelligence，Vol.23，No.11，pp.1240-1256，2001 年 11 月．
10) Tapas Kanungo，Song Mao：Stochastic Language Models for Style-Directed Layout Analysis of Document Images，IEEE Transactions on Image Processing，Vol.12，No.5，pp.583-596，2003 年 5 月．
11) 上田芳弘，加藤直孝，林克明，成田仁志，南保英孝，木村春彦：テキストマイニングと強化学習を用いた電子メール自動分配，電子情報通信学会論文誌，Vol.J87-D-1，No.10，pp.887-898，2004 年 10 月．
12) Andy Dong，Andrew W. Hill，Alice M. Agogino：A Document Analysis Method for Characterizing Design Team Performance，Transactions of the ASME. Journal of Mechanical Design，Vol.126，No.3，pp.378-385，2004 年 5 月．
13) Lam Wai：Modeling Textural Document Classification，Proceedings of 1999 IEEE International Conference on Systems，Man and Cybernetics，Vol.3，pp.946-949，1999 年
14) 有冨孝一：電子納品の現状と課題，建設マネジメント問題に関する研究発表・討論会講演集，土木学会，Vol.20，pp.103-106，2002 年 11 月．

(2005.5.20 受付)