解説

テラバイト級データを蓄積し検索表示可能にした ドキュメント管理システム

—3Dグラフィックと画像検索とDVDチェンジャ制御技術の融合—

森 吉弘^{*1}, 新田博之^{*1}, 井上充司^{*1}, 木村孝司^{*1}
島本 出^{*1}, 伊藤浩晴^{*1}, 北町篤志^{*1}

[著者抄録] コンピュータ, デジタルカメラ, スキャナーなどを利用して文書や画像を作成したり紙データをデジタルデータとして取り込んだりすることが一般的になった。その結果, 蓄積されるデジタルデータが加速度的に増加している。部署単位, 会社単位で生成される膨大なデジタルデータを, 単なるファイルとして, 共有サーバに蓄積することは量的にも利用効率の面からも限界に達しつつある。これらの問題を解決するために, 様々な検索(自己組織化画像検索・全文検索・関連検索)技術とDVDチェンジャ制御技術, 3次元グラフィック技術を統合したドキュメント管理システムを構築した。このシステムにより, テラバイト級データを簡単に蓄積・検索・表示, さらにWebをも統合することで, 広範囲にデータの配信が可能となった。

[著者付与キーワード] 自己組織化, 全文検索, 画像検索, データベース, DVD, ドキュメント管理, Web

Commentary article

A document management system that can handle over terabyte order data

—An integration of self organized picture search, 3D graphics and DVD changer control technology—

MORI Yoshihiro, NITTA Hiroyuki, INOUE Mitsuji, KIMURA Koji
SHIMAMOTO Izuru, ITO Hiroharu, KITAMACHI Atsushi

[Author Abstract] Creating digital document by scanning paper or using a digital camera or using a computer is a daily task at every office. Digital document is increasing at high pace. The quantity of digital document is almost beyond the maximum capacity of the online storage and is destroying searching efficiency. To solve these problems, we developed a document management system (ChronoStar) by integrating various searching methods (Picture, Full-Text, Related), 3D graphic and a DVD changer.

[Keywords by Author] self-organization, full-text search, picture search, database, DVD, document management, Web

^{*1} クボタコンプス(株) 情報システム部 (〒541-0059 大阪府大阪市中央区博労町4-2-15)
Tel. 06(6245)8921 E-mail: mori@Kubota.co.jp
Kubota Comps Corporation, Information System Division
(2-15, Bakuromachi 4-chome, Chuo-ku, Osaka-shi, Osaka, 541-0059)

1. まえがき

コンピュータの普及とあいまって文書や画像がデジタルデータとして蓄積されるのが一般的になっている。一つの部署や会社で生成されるデジタルデータは膨大なものとなり、個々のファイルをファイル名だけで部署・会社内に設置した共有サーバに蓄積することは、量的にも効率的にも限界に達している。

本稿では、量的な限界を解決するために、最大9.6テラバイトの容量を保持するDVDチェンジャを設置して、効率的な限界を解決するために様々な検索方法を利用可能としたドキュメント管理システムに関して述べる。

2. システム概要

開発したドキュメント管理システムの仕様を表1に示す。本システムは構成図（図1）に示すよ

うにクライアント・サーバ方式で、クライアントからサーバにデータを登録し検索する。検索クライアントはWebブラウザも使用することができる。

3. 解決すべき課題

ドキュメント管理システムを構築する上で解決しなければならない課題がいくつかあるが、以下の4点が重要であると考え（表1）。

表1 システム仕様

データ容量	オフライン機能により無限 (チェンジャ内は 1,850 枚, 9.6 テラバイト)
クライアント数	5~1,000 (サーバハードウェア性能に依存)
検索方法	全文検索, 類似画像検索, 属性 検索, 関連検索
検索時間	10 万件から 1 件を 1 秒以内
登録データ種	Windows 上で生成される文書, 表計算, 画像, 動画, CAD 図面

図1 システム構成図

3.1. 登録コストの低減

効率よく検索するためには、検索のための属性をデータに付与することが必須である。しかしながら、属性を付与することに非常に手間がかかり、データがあまり増加せずに使われなくなったドキュメント管理システムも少なくない。入力コストを下げることは、使えるドキュメント管理システムを構築するための最重要課題である。本システムでは、画像から自動で特徴を抽出し検索に利用する方法、ならびに簡単な操作で複数の属性を付与できる階層フォルダを開発することで、入力コストの大幅な削減を図った。

3.2. 多様な検索方法の用意

自分以外の人が登録した膨大なデータの中から必要とするデータを見つけるためには、一つの方法で検索に失敗した場合に、異なる検索方法で検索可能でなければならない。つまり、様々な方法で検索が可能でなければならない。本システムでは、属性検索、関連検索、全文検索、画像検索の4種類の検索を利用可能とした。本論文では、関連検索と画像検索に関して詳しく記述する。

3.3. 飛躍的なデータ増加への対応

ハードウェアの進化により、作成されるデジタルデータのサイズは予想を超えるサイズになっている。特に、動画を蓄積すればコンピュータのオンライン記憶領域では支えきれない。本論文では、どこまでも拡張可能なリムーバブルメディアをデータ格納領域として使用することを提案し、この課題を解決した。

3.4. 広範囲へのデータ配布手段

大量なデータが蓄積できた。検索手段も豊富になった。残る課題は、たくさんの人が利用できる環境の構築である。本システムは、Webブラウザを検索用のプログラムとして利用可能とし、インターネットへ接続さえできれば登録データの検索閲覧が可能となるようにした。本論文ではこの機能の詳細は説明しない。

4. 効率的限界を解決するために

4.1. 属性階層フォルダ

通常、データベース内のデータは3段階（大分類、中分類、小分類）程度の階層で分類が行われる。使用者は階層ツリーをたどることで目的のデータを見つけ出す。従来システムでは、ドキュメント管理システムを運用する前に固定された階層を構築し、使用者に提供していた。本システムでは、データベース管理者がデータベース内の属性項目から属性を選択し、その選択した属性を階層ツリーのどの層に表示するかを決定して階層ツリーを構築する。例えば、図書管理を行う場合に、Dublin Core¹⁾の15個の分類属性中からCreatorを第1階層に、Publisherを第2階層に、そしてTypeを第3階層に選択すれば、この順番でたどることのできる階層ツリー（図2）を構築できる。

図2 分類のための階層ツリー

データベース内の属性で分類ツリーを構築することで以下の二つの利点が生まれる。

4.1.1. 属性自動登録

本システムでは、操作者はデータを登録するために、登録を望むファイルをマウスで選択して階層ツリー上のフォルダへドラッグ&ドロップする。この操作により、最上層からそのフォルダまでの経路上にある属性を登録中のデータに設定する。つまり、ドラッグ&ドロップするだけで、検索に使用できる複数の属性を自動でデータに設定できたことになる。

4.1.2. 属性階層交換

ドキュメント管理システムの管理者は、データベース内に存在する属性を選択して階層属性ツリーを構築する。この機能は、運用が開始された

後でも実行することができる。図2の分類体系の大分類であるCreatorを第2階層へ、そしてPublisherを第1階層へ移動させた図3の体系で運用を続行できることになる。この運用後に体系を変更できる機能により、運用開始までの時間を大幅に短縮できる。

図3 階層変更後の分類ツリー

4.2. 類似画像検索

検索対象が文書である場合には、ドキュメント管理システムへの登録時に文書内に含まれる単語を抽出記憶することで、属性を付与することなく文書の検索を可能とする全文検索技術がある。しかしながら、画像に関しては属性を自動で付与する方法が確定しておらず、文書と画像の登録効率には大きな隔たりがあった。これを解決するために自己組織化類似画像検索方式を考案し、画像に対しても属性を付与するコストが発生することなく登録し検索可能とした。システムに実装した類似画像検索のアルゴリズムを以下に示す。

- (1) 登録される画像を横4分割（図4）し、各領域内のRGB情報を平均化しLAB情報に変換した後、これを画像の属性としてデータベース内のLABテーブルに格納する。LABテーブルは、ファイルID、L、A、Bの4個のフィールドで構成される。縦方向にも同様の処理を行い、 $4 \times 3 \times 2$ で24次元（ $X_1 \sim X_{24}$ ）の属性が抽出される。
- (2) この特徴を、Kohonen²⁾³⁾により提案された式[1]で学習する細胞を2次元上にならべた自己組織化マップ（図5）に入力し、最も入力に近い状態を持つ細胞とその近傍の細胞を学習させる。繰り返し学習を行うことで、似た状態を持つ細胞が近くに集合する。近傍の定義としては、注目している細胞からのユークリッド距離がある閾値以下であるものとした。

図4 横分割

図5 自己組織化マップ

- (3) 変化が収束したところでマップを固定し、データベース内の各画像がマップのどの位置に属するか決定し、各画像のマップ位置としてデータベース内のIPテーブルに記憶させる。IPテーブルは、ファイルID、X、Yの3個のフィールドから構成される。
- (4) 検索時にユーザは、任意の画像を指定し、その画像と似た画像を検索するように指示する。ドキュメント管理システムは、指定された画像とマップ上で同一位置に属する画像をIPテーブルから検索し類似画像として結果を返す。

$$m_i(t+1) = m_i(t) + \alpha(x(t) - m_i(t)) \quad [1]$$

$m_i(t)$: 時刻 t における細胞 i の状態

α : 学習係数

$x(t)$: 時刻 t における細胞刺激（入力）

この認識方式を製品に組み込むにあたって、認識精度を保ったまま、この学習が夜中（最長8時間）のうちに終わるようにするべく、以下のような高速化を行った。

- (1) 自己組織化マップのサイズを認識対象ファイ

ル数に応じて可変とした。サイズは8×8から開始し、1エリアあたりにマップされるファイル数が40を超えないように増加させた。このアルゴリズムにより、データ量に比例した最適なマップサイズが用意できることになる。つまり少ないデータ数の時に大きなマップを用意するような非効率な計算を回避することが可能となった。また、マップの大きさをデータ量に比例させたために、マップのエリア間の相互影響をあるレベル以上に保つことができたため、偏りのないマップの作成に寄与した。実装したアルゴリズムでは、完全にデータに比例してサイズを可変にする方法ではなく、認識対象が2,500枚までは8×8で、10,000枚までは16×16で、それ以上は48×48で対応する段階的なものとした。

- (2) ファイル登録時に既存の自己組織化マップ上での位置を計算し、現在位置として記憶し、認識に利用する。
- (3) (2)で計算し記憶した位置を、次の自己組織化時の近傍計算開始地点として利用する。計算初期段階で最小値に近い値が求まり、各エリアとの距離計算時に、この値を超えた場合に計算を打ちきめることで計算量を少なくする。

これらの高速化後に、Celeron400MHzを搭載するコンピュータで24次元の特徴ベクトルを持つ1,000枚の画像を16×16の自己組織化マップで1,000回学習したところ400秒を要した。自己組織化に必要な時間は以下のような式で表すことができる。

(計算時間) =

$$C \times TF \times VD \times MS \times MS \times LI \quad [2]$$

定数 : C

認識対象画像数 : TF

特徴ベクトル次元 : VD

マップサイズ : MS

学習回数 : LI

100,000枚の画像に対してマップの大きさを50×50にして学習を行えば、約1,000倍の計算量が必要となり約108時間かかることになる。これ

を8時間以内に収めるためには、10倍以上のさらなる高速化が必要となるが、本手法が本質的に並列計算⁴⁾なので、4CPUで動作クロックの数倍速いサーバを用意すれば、ほぼこれを満たすことができる。つまり、認識対象が10万枚程度まではこの方法で対処できることを意味している。また、本手法の認識に要する時間は認識対象数の影響を受けないので、使用者にとっては認識対象数が増えても検索効率は変わらないという大きな利点がある。

本機能により、画像に対しても属性を付与することなく検索を行うことが可能となり、大量の画像データを検索可能なデータとして低コストで登録することが可能となった。

4.2.1. 関連検索

本システムに登録されるデータは、関連するデータを指定することができる。例えば、ある画像ファイルとその画像を説明するためのドキュメントファイルを関連のあるデータとして設定できる。つまり、検索者は登録者が設定した関連を追跡できることになる。いいかえれば、登録者が埋め込んだ知識を追跡することができる。

5. 量的限界を解決するために

ドキュメント管理システムに登録されるデータの量は飛躍的に増加しているため、それを保持する記憶領域は使用者の要求に応じて簡単に増加させられるものでなくてはならない。この要求を満たすことができる記憶媒体は、記憶媒体を容易に交換できるMOやDVDのリムーバブルメディアである。本システムでは、DVDチェンジャを以下のような機能で統合し、データ量の飛躍的な増加に対処しつつ、その保管したデータが再利用可能な状態に保てるようにしている。

5.1. DVDチェンジャ連携

5.1.1. 信頼性エージェント

メディアに保管したデジタルデータが数年後に読み出せることはメディアが保証する。しかし

ながら、その読み出したデータが確実に再利用、再配置できるかどうかは保管したフォーマットを理解できるアプリケーションが存在するかどうかにかかっている。本システムでは、登録時や登録後にデータフォーマットをチェックするエージェントを動作させることで、利用できなくなるであろうフォーマットがデータベース内に存在することを防いでいる。エージェントとしては以下の3種類が存在する。

(1) フィルターエージェント

データの保存時に動作するエージェントで、データベース管理者が設定したデータ変換テーブルを参照して、保存にふさわしいデータかどうかを判定する。

(2) コンバータエージェント

データの変換を行うエージェント。ユーザが起動する場合と、フィルターエージェントから自動で起動される場合がある。

(3) ウォッチャーエージェント

データベース内を巡回して、管理者により設定された変換を行う。

5.1.2. 均質な検索速度

チェンジャから取り出されて、実在の棚に置いて保管されるDVDメディアをオフラインデータと呼ぶ。オフラインデータに対しても検索可能とするために、検索のために必要な属性を、ハードディスク内に保持した。これにより、本システムでは、検索に必要な時間が記憶媒体に関係なく均質になっている。

5.1.3. 記憶階層間移動

検索に必要とされる時間は、前項の機能により媒体非依存となったが、データを開く時間は媒体依存である。必要なデータをあらかじめ記憶階層間で移動させておくことは、多数のデータにアクセスする場合には非常に重要な機能となる。例えば、明日、古い冷蔵庫に関するコラムを執筆することがわかっているならば、1997年より前に作成された文書で、“冷蔵庫”という単語を含む文書が高速にアクセスできる媒体上に集められていれば、

非常に効率よくデータにアクセスすることができる。これを可能とするために、本システムでは、ある検索条件に合致する結果を一括してDVDメディアからハードディスク上へ転送することができるようになっている。その逆方向で不要になったデータを、ハードディスクからDVDメディアへ転送することも可能である。また、一括転送を行う時間を指定することで、必要なデータを比較的コンピュータの負荷が低い夜中に転送できるようになっている。

5.2. 3Dビューワ

大量のデータを表示するためには、人間の視覚に優しく、しかも使用者に記憶されやすい形で表示することが必要である。本論文では3次元グラフィック技術を活かした3種類の表示方法を提案する。

5.2.1. スライダービューワ

このビューでは、検索結果を円筒の表面に貼り付けて表示する。このように表示することで、目的のデータが大量データのどのあたりに存在していたかを記憶しやすくなる。また、画面手前のデータを詳細に表示し、画面奥のデータを小さく表示することで一度に大量のデータを注目しているところの詳細を失うことなく表示できる。

5.2.2. ブックビューワ

人間にとって優しい表示とは、普段の生活において慣れた形態での表示である。特に不特定多数の人々が集まる場所（図書館、美術館、博物館、駅構内など）で使用される場合には必須の機能となる。本システムでは、3Dグラフィックによりページめくりも行うことができるバーチャルブックを表示し、初めてでも直感的に操作できるように図7に示すブックビューワを実装した。1ページに表示できるデータの数には1～16枚で可変とし、ページをめくる速度は使用者が高速・通常速、低速から選択できるようにした。また、自動めくりモードを準備することで、膨大なデータをコストをかけずにチェックすることを可能とした。

図6 円筒状スライダービューワ

図7 ブックビューワ

5.3. ウォークスルービューワ

自己組織化画像検索では、すべての画像は $N \times N$ の2次元マップのどこに属するか決定されている。この2次元マップを3次元上に表示したのがウォークスルービューワである。操作者は、操作者の意図した方向（上下・左右）にこの3次元空間を進んでいくことができる。操作者の主観的類似性の軸に従って画像の中をウォークスルーすることができる。

6. 公共施設で使用した場合の利点

本システムを図書館・美術館・博物館などで運用した場合の利点を示す。

6.1. 利用者の利点

(1) マルチメディア（文書、画像、動画、音声）の

検索、表示ができる。

(2) 多彩な検索（属性検索、画像検索、関連検索、全文検索）により、様々な視点から検索可能である。

(3) 臨場感のある表示（ブック表示、円筒状スライダー表示、ウォークスルー表示）により、初めてでも使いやすい。

(4) Web機能により、開かれた情報源となる。

6.2. システム運用者の利点

(1) デジタル複合機器と連携しているため、紙データ入力が非常に簡単である。

(2) DVDチェンジャ連携しているため、大量データを長期保存できる。

(3) 関連検索機能をリファレンスサービスとして利用することで、登録者の持っている知識を公開できる。

図8 ウォークスルービューワ

7. むすび

本論文では、画像から抽出した色情報ベクトルを基に自己組織化マップを作成し認識に利用することを提案するとともに、10万枚の画像を実時間認識できるシステムを開発した。さらには膨大な検索結果を人に優しく表示するために、3次元グラフィック技術を応用した3種類のビューを開発し、検索結果データの中を操作者の主観類似軸に沿ってウォークスルーできることを示した。

現在は、ウォークスルービューをさらに進めて、3次元表示する機能とデータベースの統計情報を重ね合わせた3次元ナビゲーション機能を開発中である。また、今回提案するシステムは、認識特徴として色情報だけを扱うものであったが、認識システムの特徴抽出部分はプログラムとして独立しているため、他の特徴抽出方法と置き換えることが可能であり、医療画像やCAD図面を認識するシステムが構築可能になる。

～～～ 参 考 文 献 ～～～

- 1) Dublin Core Metadata Element Set.
(<http://purl.org/dc/documents/rec-dces-19990702.htm>)
- 2) T. Kohonen Self-Organizing Maps. vol30,
Springer Info-Sciences
- 3) ATR国際電気通信基礎技術研究所編. ニュ
ーラルネットワーク応用. オーム社, 1995,
227p.
- 4) 森 吉弘. PDPモデルによる手書き漢字認識.
信学論. Vol.J73-D-II, No.8, 1990, p.1268
-1274