情 報 学 基 礎　68－７
自 然 言 語 処 理 151－７
（２００２．９．１８）

構造的特徴を有するデータ間の類似度計算

佐藤 慶三　中島 誠　伊藤 哲郎

大分大学工学部知能情報システム工学科

〒870-1192大分県大分市旦野原700番地

E-mail: {skei, nakasima, ito}@csis.oita-u.ac.jp

計算機の発達や普及に伴い様々な形式のデータが利用され，それらがユーザの間で共有されることが多くなっている．しかしながら，大量かつ様々な形式が混在する共有データを効率的に管理する手法はまだ確立されていない．この点に注目し，大量の共有データの中から所望のデータを効率よく取り出せよう，データの内容を示すキーワードよる類似度計算とデータの構造的な分類情報であるパス名による類似度計算を組み合わせてブラウジングする手続きを定式化する．提案した手続きについては，検索効率の面からの有効性を実験的に示す．

Measuring Similarity Values among Data with Structural Features

Keizo Sato, Makoto Nakashima, Tetsuro Ito

Department of Computer Science and Intelligent Systems, Oita University

700 Dannoharu, Oita, 870-1192, Japan

E-mail: {skei, nakasima, ito}@csis.oita-u.ac.jp

A user comes to manage various forms of the data and share them among the other users through a LAN and/or the Internet. There is, however, no fine method of effectively managing the large volume of the shared data, which are usually stored under a tree-structured directory. We first describe two types of similarity measures: One incorporates the path names, i.e., the structural features of the data, and the other the keywords, i.e., the contents of the data. Next, we formulate the method of managing the data by arranging them according to the content similarities and then by browsing according to the two types of the similarity values. The effectiveness of the proposed method is also examined.

1. はじめに

Web におけるサーチエンジンなど情報検索技術の発達によって，ユーザはインターネットを通じて様々なデータを入手できるようになってきた．さらに，LAN の普及も進んだことにより，企業や研究機関等ではユーザ同士が，作成したデータやあるいは Web を通して収集したデータを共有しながら作業を進めていくケースも多くなってきている．

共有作業環境では，一般に数名から数十名で一つのグループを作り作業を行っている．作業内容などにもよるが，近年個人ユーザの扱うデータの量が増大している状況を考えると，共有作業環境では扱われる共有データの量は膨大なものになる．また，扱われているデータや分類の仕方も十分に把握しきれないため，共有データの中を広範囲にわたって探索していくのは困難となる．データの取り出しの効率化を考慮した管理手法が必要である．

大量の共有データを管理するための方策としては，データベースシステムの導入が考えられる．実際，企業などでは，データベースシステムによりデータの管理を行っている．しかしながら，共有データは作業内容に応じて形式や構造が多種多様である．そのため，スキーマを規定できずデータベースシステムによる管

理は難しい．厳密なスキーマの限定を必要とせずに，データの取り出しを支援するような仕組みが必要である．ここでは，大量のデータの中から効率的に所望のデータを取り出すための仕組みとして，ブラウジング支援機構[10][11]の共有データ管理への適用を考える．また，データは作業内容により分類され，データの位置関係がデータ同士の関連を反映していることから，データの分類先を示すパス名をデータの特徴の一つとして扱うことを考える．パス名は複数のラベルとその順序関係で構成されていることから，一種の構造的特徴として扱う．

以下，**2.**では共有データの管理に用いるブラウジング支援機構とパス名の導入，さらに，キーワードと併せた扱い方に関する考察を述べる．**3.**ではブラウジング支援機構と，パス名による類似度計算の導入，そしてキーワードによる類似度計算とパス名による類似度計算, 2つの類似度計算のブラウジング支援機構への組み込みとブラウジングの手続きについて説明する．**4.**では提案手法の有効性に関する実験について示す．実験では, キーワードとパス名併用の効果が認められた．**5.**ではパス名の利用の仕方についてと，データ自体の扱いについての考察を述べる．

2. ブラウジング支援とパス名利用に関する考察

2.1 ブラウジング支援に関する考察

ブラウジング支援の関連研究としては，適合フィードバック[3][5]が挙げられる．これは，ユーザが参照したデータについての適合判断結果をもとに，初期の検索式を修正して再検索する，といったアプローチである．しかしながら，扱うデータの形式が様々で，キーワードによる特徴付けが困難なものもあるうえ, 共有作業環境では携わるユーザの数も複数であるため書式にも一貫性がなく, これをそのまま適用しただけではデータの管理には不十分である．また，データの提示については, 質問との類似度によるランキングを更新して提示を行うため, 提示できる情報が局所的なものになってしまう. データ同士の位置関係も検索質問が与えられるたびに入れ替わってしまい, 共有データ管理に適しているとはいえない.

共有データ管理では, 部分的な追加や削除の操作はあってもデータの規模からすれば変化は小さく, データの分類先が頻繁に更新されることはない. 分類先に大きな変更がなければ，データ間の位置関係も保持され，参照したデータについてその位置関係を把握出来るようになる．この点から，提示の面では，利用経験を通して獲得した共有データに対する知識が生かせるような工夫が必要になる．

[10][11]は, 電子図書館による文献データのブラウジング支援について述べている. この電子図書館上では, データは静的な線形配置によって提示される．線形配置は関連の大きいデータ同士が互いに近くに配置されるような順序付けによって形成され，ユーザは配置に沿って適合判断結果にもとづいた支援機構によりブラウジングしていく．データの提示については，一瞥性と参照するデータの情報の把握を両立した提示方法が提案されており，ブラウジング支援と全体像の把握を目的としたユーザインタフェースの連携が共有データ管理に適する．

この電子図書館の技術を適用すれば，共有作業環境でデータを扱う場合でも文献データの管理と同じように効率的なブラウジングの実装が期待できる．しかしながら，適合フィードバックと同様，キーワードにもとづくブラウジング支援が主体であるため，データの管理に不十分である．これを補う方策として，次節でパス名によるデータ間の関連の把握についての考察を述べる．

2.2 パス名の利用に関する考察

ユーザが日常扱うデータは，ユーザによって作業内容に応じて分類されており，分類先を示す情報を利用すれば，キーワードとは別な観点からデータ間の関連を導き出せると考えられる. 図1にデータの分類例を示す.

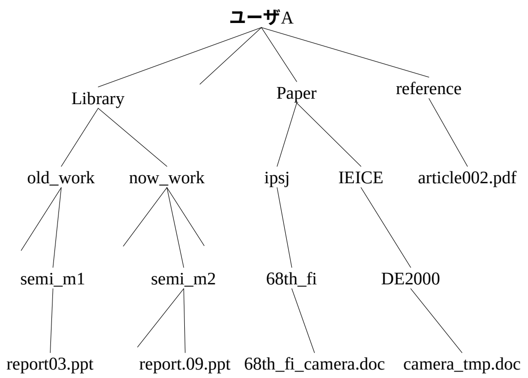

図1. データ分類の例

図1での「report03.ppt」や「report09.ppt」はともに，研究の進捗の報告に用いられたデータの例である．これらのデータでは，「研究の進捗の報告」という関連がありながら，作業内容自体の違いによりキーワードにもとづく関連性はとらえられない場合もある．しかしながら，利用される場面の共通性により分類先のフ

ォルダ名の付与に「Library」や「work」，「semi」など共通の語が用いられている．逆に，「report03.ppt」と「article002.pdf」は利用される場面が少なく，分類先のフォルダ名に共通の語はみられない．異なるユーザのデータについても，作業の内容だけでなくデータの利用目的に従って分類の際付けられるフォルダの名前に共通の語を用いる傾向がある．以上の考察を基に，分類先の情報をデータ間の関連を求めるのに用いる．

2.3 パス名とキーワードの併用に関する考察

データの構造的な特徴を利用した技術としては，XML 文書検索の研究に見られるような，いわゆる「半構造データ」の関するものがある．[6]では木構造あるいはグラフで表現されたデータ構造についての，検索技術や応用が述べられている．他の研究例としては，索引付けに関するもの[15]や半構造データの枠組みの一般化についての研究[19], さらに XML の形式で学術資料を効率的に管理するシステムの提案[8][18]がある．

半構造データの扱いでは，検索質問も構造的な表記で指定される．図2の左で示されるような構造を持つデータに対し，ユーザは図2の右で示されるような形で検索質問を与える．与えられた検索質問を含むようなデータを検索結果として出力するというのが半構造データにおける検索の一般的な操作である．

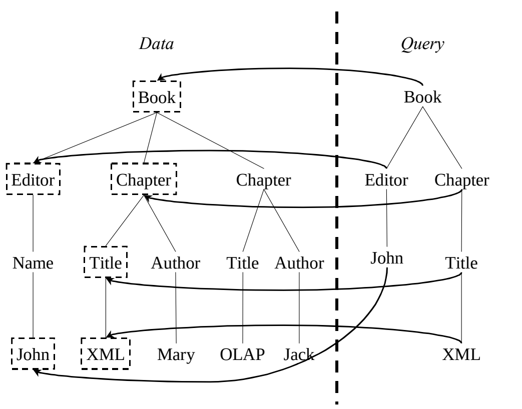

図2．半構造データの検索例

ここで，パス名の半構造データとしての扱いが妥当かどうかについて考察する．[16]では，過去の半構造データに関する研究から，一般的な半構造データの特徴について述べている．それによると，半構造データの特徴は，「大まかな構造の類似性をもちながら、部分的な欠落などで互いに異なる構造をもったデータ」という観点でまとめられている．

パス名についてこの観点で考察してみると，半構造データについては，データ間で大まかな構造の類似性があり，関連を求める手掛かりとなっているのに対して，共有作業環境におけるパス名の構造は，データの作成者によって異なるため，共有データの扱いではパスの構造の類似性だけにもとづいてデータ間の関連をとらえることは難しい．

従来の半構造データを扱う研究は，半構造的な特徴のマッチングに注目し，半構造的特徴と，キーワードに代表されるデータ内部の情報を組み合わせることでデータ間の関連性をとらえるような試みを行っていない．しかしながら，図3のようにデータ間には，パス名による関連性とデータ内部の情報にもとづく関連性の双方が存在する．そのため，テキストなどのデータ内部の情報も，データを特徴付ける重要な要素である．さらに，各データの分類先のフォルダに対する上位フォルダには，隣接フォルダとの関連性を表す情報が含まれることから，フォルダ構造での上位階層を通して得られる情報もデータの特徴付けに利用できる．これらを統合的に扱ったデータ間の類似度計算ができれば，ブラウジング支援の効果が期待できる．

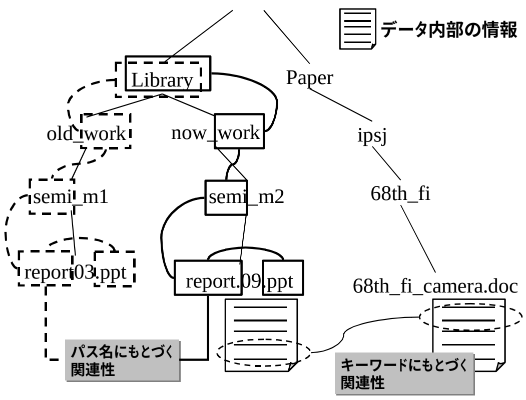

図3．フォルダ構造におけるデータ間の関連

上記の観点から，パス名による類似度とキーワードによる類似度を組み合わせたブラウジングの手続きを考える．キーワードの処理について，本稿では上位フォルダの情報は使わず，データ自身から抽出したもののみを用いた．

3. 2つの類似度計算にもとづくブラウジング

ここでは，まず，ブラウジング支援機構についての概要を説明したうえで，パス名による類似度計算，そしてキーワードによる類似度計算とパス名による類似度計算を組み合わせたブラウジング支援機構への組み込みと，ブラウジングの手続きについて述べる．説明のために，個々のデータを $d_i, d_j,$ …で，適合（不適合）

と判断されたデータは d_r（d_n）で，質問は d_q でそれぞれ記す．2データ間の類似度を求める測度は s で記す．

3.1　ブラウジング支援機構 WiB1

ユーザの適合判断結果を学習し，未参照データの適合可能性を示唆するブラウジング支援機構 WiB1が定式化されている[10][11]．WiB1では，適合判断結果をIB1[2]によって学習し，各未参照データについて適合可能性を判定する．ユーザは WiB1によって示唆された適合可能性を参考にブラウジングしていく．適合可能性の判定式は以下のようになる．

$$max\left(s(d_i,d_q),max_{d_r}s(d_i,d_r)\right)\geq max\left(\tau,max_{d_n}s(d_i,d_n)\right)$$

判定式中の τ は小さい類似度が原因で誤った適合可能性を示唆してしまうのを避けるためのパラメータである．また，WiB1による適合可能性示唆を利用したブラウジングの手続きは以下のようになる．

[手続き WiB1]

(W1) 各 d_i について W2を行う．

(W2) d_i に対し，IB1からの適合可能性示唆があれば，(W2.1)と(W2.2)を行う．

(W2.1) d_i をユーザに示し，適合判断結果を受け取る．

(W2.2) 判断結果に従い IB1で学習する．

2.で述べたように，データの配置はデータ間の類似度にもとづき，関連のあるデータ同士が互いに近くにおかれるような線形配置とする．WiB1を静的な配置の上で効率的に動かすには，少しでも多くの適合データの参照漏れを防ぐために，ブラウジングの序盤で適合データをいくつか確保できるようにする必要がある．この点を踏まえた，ブラウジングの手続きは以下の通りである．

[手続きWiB1-DL]

(L1) 次の L1.1と L1.2を行う．

(L1.1) 質問と大きい類似度を示す少数のデータ d_i, d_j,⋯およびそれらの近くに配置されたデータについて WiB1を実行する．

(L1.2) 適合判断のなされていないデータについて，それらの配置順に WiB1を実行する．

(L2) 適合判断のなされていないデータについて，質問との類似度の降順に WiB1を実行する．

(L3) 適合判断のなされていないデータについて，質問との類似度の降順にブラウジングする．

3.2　パス名による類似度計算

データ内部の情報にもとづく関連性は，共通のキーワードを多くもつ場合に大きな値を与えるコサイン測度[12]を使ってとらえることが出来る．一方，パス名中に出現するラベルは，部分一致による意味的なつながりなど，データ間の関連をとらえるのに有効であり，この性質を考慮したマッチングが望ましい．そこで，パス名のマッチングを，ラベル間の順序関係を考慮しながら，類似するものを探せるような計算方法によって求める．

部分一致を考慮した文字列パターンのマッチングと順序関係の把握を両立できる測度として，類似文検索のために定式化された構造マッチング[1]を利用する．構造マッチングでは，部分一致を考慮した処理を行うため，異なるユーザによる表記の違いにも柔軟に対応できる．

構造マッチングでは，マッチングの対象となる文は複数の句からなり，句はさらに複数の語によって構成されるとする．文間の類似度を求めるのに，句，単語単位でのマッチングを行い，その結果をもとに類似度を計算する．構造マッチングをパス名に適用する場合は，文，句，語の対応は図4のようになる．データの取り出しにおけるユーザの検索要求は，データの種別や名前も含む可能性があるため，これらもパス名のマッチングに利用する．パス名全体を一つの文として扱い，データ名，拡張子名を分割したものを句として扱う．データ名，拡張子名に加え，パスのラベルをさらに分割したものを語の羅列として扱い，マッチングを行う．

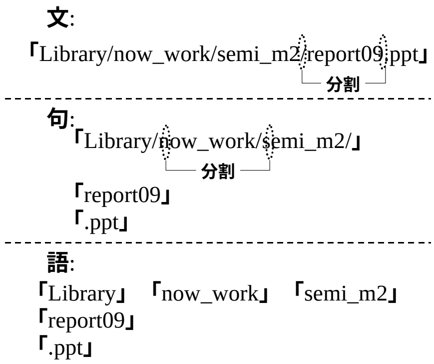

図4．パス名の解体

文間の類似度は次式を使って求める．

$$(S1+S2)/(Nx+2(Nx-1))$$

ここで,

S1： 句ごとのマッチングによる得点.
S2： 隣接2句を1組としたマッチングの得点.
Nx： マッチング対象の2文について, 句の数の最大値.

とする．句間でのマッチングでは，*S1*，*S2*は語ごとのマッチングより求め，*Nx* は語の数で求める．語間でのマッチングでは，文字数によってそれぞれ計算する．図5はいくつかのパス名について, 構造マッチングによる類似度計算の例を示したものである．

d_1： /Library/now_work/semi_m2/report09.ppt
d_2： /Paper/Ieice/DE2000/camera_tmp.doc
d_3： /Papaer/ipsj/68_fi/68th_fi_camera_03.doc
d_4： /reference_1/article002.pdf
d_5： /Library/old_work/semi_m1/report03.ppt

	d_1	d_2	d_3	d_4	d_5
d_1	1.00	0.10	0.07	0.16	0.83
d_2	0.10	1.00	0.52	0.13	0.14
d_3	0.07	0.52	1.00	0.10	0.14
d_4	0.16	0.13	0.10	1.00	0.13
d_5	0.83	0.14	0.14	0.13	1.00

図5．パス名リストと類似度計算の例

3.3 ブラウジングの手続き

ブラウジングの手続きは，WiB1-DL を基本とし，キーワードによる類似度計算とパス名による類似度計算を適合可能性の判定に組み込む．キーワードによる類似度計算での判定とパス名による類似度計算での判定の何れかで判定式が成り立てばユーザに適合可能性示唆する．パス名による類似度では，近くに分類されながら関連のないデータが多い場合にそれらを連続して参照してしまうため，ここでのデータの配置はキーワードによる類似度でみて，互いに関連のあるデータ同士が近くに配置されるようにする．以下に，ブラウジングの手続きを WiB1-DLαとして定式化する．

[手続き WiB1-DLα]

(α1) キーワードによる類似度計算のみで次のα1.1とα1.2を行う．
  (α1.1) 質問と大きい類似度を示す少数のデータ d_i, d_j,…およびそれらの近くに配置されたデータについて WiB1を実行する．
  (α1.2) 適合判断のなされていないデータについて，それらの配置順に WiB1を実行する．
(α2) 適合判断のなされていないデータについて, 質問との類似度の降順に WiB1を実行する．
(α3) パス名による類似度計算での適合可能性判定を WiB1に加え，α1.2とα2を再び行う．
(α4) WiB1-DL の後処理と同様，適合判断のなされていないデータについて, 質問との類似度の降順にブラウジングする．

WiB1-DLαでは，α1.2およびα2を適合可能性判定の仕方を変えて，2回行っている．まずキーワードによる類似度計算での適合可能性の判定結果をユーザに示唆する．2回目の処理では，キーワードによる類似度計算での判定に加え，パス名による類似度計算での判定も行う．これは，1回目でキーワードによる類似度計算での適合可能性示唆だけでのブラウジングで適合データを獲得し，次にこれら適合データとパス名の観点から関連の大きいデータを探索していくという手続きである．はじめにいくつかのデータを参照していき，必要に応じてそれぞれのデータについて周囲のフォルダをブラウジングしていくという，手動による探索に倣っている．

4. 実験的考察

先に述べた, 2つの類似度計算のブラウジング支援への組み込みについて，有効性を調べるための実験を行った．ここでは研究室の学生14名に共有作業環境で実際に利用しているアプリケーションデータを提供してもらい，これを実験用データとした．提供されたデータは，フォルダ構造に従い全体を一つのデータとしてまとめた．フォルダは，ルートから「研究テーマ→ユーザ→…」という形で構成した．実験用データの総数は3490であった（実験用データは，文書データやプレゼンテーションデータなどが主体であった）．

実験を行うために，各データについてキーワードも用意した. 各データに対するキーワード抽出の処理は, テキスト部分の抽出, 形態素解析, 重み付けからなる．テキスト部の抽出については TextPorter[4]，形態素解析には茶筌[7]をそれぞれ用いた．形態素解析の出力結果からは，一般名詞，サ変名詞，固有名詞を取り出だし，一般名詞とサ変名詞については，連続する場合は複合語として再現してキーワードとして扱った（英語表記は固有名詞として扱った）．各キーワードへの重

み付けは *tf･idf* [13]を利用した．

検索質問は研究内容などにかかわるキーワードの羅列とした．検索質問に対するブラウジングを，計算機によるシミュレートにより実施し，これをユーザによるブラウジング操作とみなした．検索質問は13用意し，それぞれ提供してもらったデータの分類されている状況に従って，研究テーマや行事等に関連するキーワードによって構成した．

各質問に対する適合データとしては，内容が質問と合致するデータおよび，それらのデータと関連のあるデータを適合データとして設定した．各質問に対する適合データの数は平均で23であった．ブラウジングの過程で出合うデータのうち，条件に合致するデータはユーザにより適合と判断されたとみなし，他のデータはすべて不適合データと判断されたものとみなした．

適合可能性の判定におけるパラメータτについて，キーワードのマッチングでは0.2と設定し，その他の類似度計算では，データ間の類似度の大きさに従って調整した．キーワードのマッチングについては，コサイン測度で類似度を求めた．ベースラインとしては，WiB1-DL を用いた．

4.1　2つの類似度計算併用効果の検証

パス名による類似度計算とキーワードによる類似度計算の導入によるブラウジング支援の効果について示す．実験内容は，WiB1-DLαに従った形でパス名による類似度計算の異なるものを，WiB1-DLαと比較する形で行った．比較のための類似度計算は，コサイン測度を利用した．図4における語を一つのキーワードとして扱い，重複のないようにキーワードの羅列としてまとめた．各ラベルの重みは一律1.0とした．実験結果を図6に示す．パス名による類似度計算をコサイン関数で求めた場合の結果については，cosine と記す．

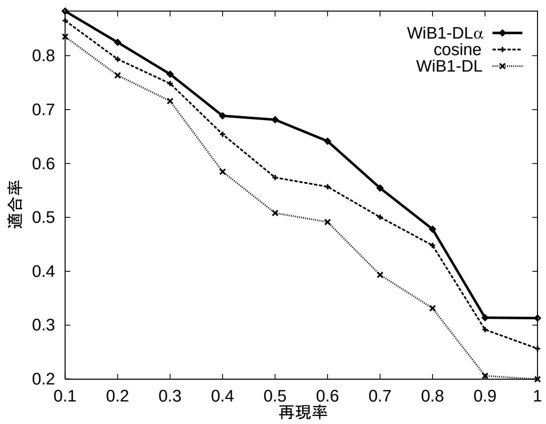

図6．パス名による類似度計算導入の効果

WiB1-DLαは WiB1-DL に対して再現率0.5～0.8のとき，統計的にみて95%の信頼度で高い適合率を示した．cosine に対しては0.5～0.6のとき，統計的にみて高い適合率を示した．実験結果から，パス名の利用によりキーワードだけではとらえられないようなデータ間の関連をとらえることができ，ブラウジング支援に効果があることが分かる．また，類似度としては構造マッチングを使ったほうが妥当といえる．

4.2　ブラウジングの手続きに関する検証

もう一つの実験として，ブラウジングの手続きに関する検証を行った．WiB1-DLαで取り入れた，手動での探索に倣った手続きの妥当性を検証するため，以下のようなブラウジングの手続きを定め，比較に利用した．

[実験用手続きWiB1-DLβ]

(β1) WiB1にパス名による類似度計算での適合可能性判定を組み込み，適合判断のなされていないデータについて，それらの配置順に WiB1を実行する．

(β2) 適合判断のなされていないデータについて，質問との類似度の降順に WiB1を実行する．

(β3) WiB1-DL の後処理と同様，適合判断のなされていないデータについて，質問との類似度の降順にブラウジングする．

WiB1-DLβは WiB1-DLαと異なり，キーワードによる類似度計算だけで適合可能性の示唆を受けるブラウジングの過程を省略している．WiB1-DLαとの比較結果を図7に示す．

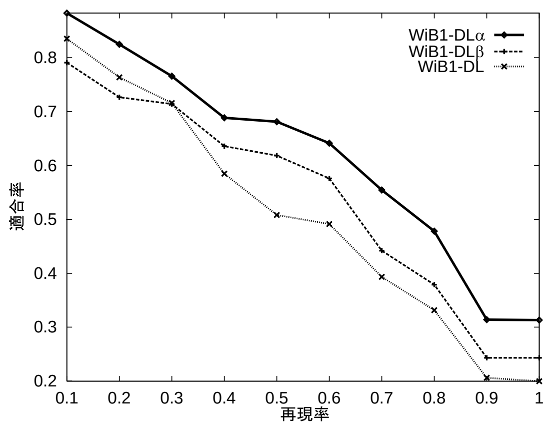

図7．ブラウジングの手続きの比較

実験結果について，再現率が0.3～0.6で両者に統計的

な有意差はなかったものの，グラフ全体として WiB1-DLαの方が上位に描かれた．このことから，共有データ管理でのブラウジングの手続きとしては WiB1-DLαを使ったほうがより妥当といえる．

5. データのもつ特徴の統合的な扱い

5.1 パス名情報のキーワードへの変換

提案したパス名の利用の仕方は，キーワードにもとづく処理と独立な扱いになっている．しかしながら，パス名の利用の仕方は個別な扱い方に限定されるものではない．

別な形でのパス名の扱いとして，各データに対するキーワードの割り当ての段階で，パス名を利用した情報を追加する方法が考えられる．パス名を構成するラベル自体や，**2.**で述べたように上位フォルダの情報を利用して関連付けられるデータのキーワードを付加することで，**3.**とは別な形でパス名を利用できる．この手順によれば，ブラウジング時はパス名に対する特別な処理を必要としないため，提案手法に比べブラウジング手順が簡略化される．ただ，パス名を手掛かりにデータ間の関係を導くという点では，**3.**で述べた手法とは本質的な違いはないため，検索効率の面での大きな改善は期待できない．

5.2 データ自体の半構造性

本稿では，データがもつ構造的な特徴としてパス名を取り上げ，キーワードによる類似度計算と併せてデータ間の関連性をとらえることについて述べた．しかしながら，データ自体については内部の構造を特に考慮せず，古典的なテキスト検索の手法を取り入れるにとどまっていた．

実際には，データ自体にも半構造的な性質が認められる．たとえば，今回実験でも取り上げた文書データやプレゼンテーションデータは，「タイトル」「著者名」「サブタイトル」「本文」など，ある程度共通する構造が見られる．大まかな構造のもとで部分的に異なるといった特徴から, XML 文書に限らず日常の作業で共有される種々の形式のデータについても半構造的な扱いの適用が可能であると考えられる．

また，データにはテキスト以外の情報も埋め込まれており，それらもまたデータを特徴付ける重要な要素である．データを構成する多様な特徴をうまく処理することで，汎用的なデータ管理の枠組みを組み立てていくという方向性が考えられる．

6. おわりに

本稿では，共有データの管理を目的として，構造的な特徴であるパス名を利用するにあたって，扱い方を半構造データの観点から考察した．さらに，パス名を利用したデータ間の類似度の計算方法と，キーワードによる類似度計算と併せたブラウジング支援機構への組み込み方について述べ，2 つの類似度計算をうまく組み合わせることで，データ間の関連性の把握に有効であることを検索効率の面から確認した．

ここでのパス名の類似度計算は，文字列の部分的な表記の違いには対応できるが，類義語のようにまったく異なる表記での概念的な関連をもつ語や日英の違いなどにはまだ対応してない．このようなケースについての対応は，[8]や[12]のような，概念辞書を利用した検索手法を導入していくことを検討する．

今後の予定としては，上で述べたマッチング法の改善に加え，データの扱いについて，タイトル部などデータ内部の構造を考慮した特徴付けの導入が挙げられる．これについては，複数のメディアで構成されたデータに対する特徴付けや検索技術に関する研究[16]もなされており，これらの技術が発展するとデータの構造を反映したキーワード抽出が可能になる．また，本稿で述べたブラウジングの手続きを実装した共有データ管理システムの構築も課題点の一つである．共有データの管理システムを実装するにあたっては，ブラウジング支援にとどまらず，共有作業に対する支援となるような機能の充実についても検討していく．

参考文献

[1] 安部 隆之, 佐藤 浩史, 重松 修一, 中島 誠, 伊藤 哲郎, “構造マッチングによる文献の知的検索と結果の色空間表示,” 情報処理学会研究報告，人文科学とコンピュータ, vol.29-5, pp.25-30, Jan. 1996.

[2] D.W. Aha, D. Kibler, and M.K. Albert, “Instance-based learning algorithms,” Machine Learning, vol.6, pp.37-66, 1991.

[3] J. Allan, “Incremental relevance feedback for information filtering,” Proceedings of the 19th Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR’96), Zurich, Switzerland, pp.270-278, Aug. 18-20 1996.

[4] Antenna House, TextPorter. http://www.antenna.co.jp/.

[5] M. Iwayama, “Relevance feedback with a small number of relevance judgments: Incremental relevance feedback vs. document clustering,” Proceedings of the 23rd Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR’00), Athens, Greece, pp.10-16, July 24-28 2000.

[6] Raghav Kaushik, Pradeep Shenoy, Philip Bohannon, Ehud Gudes, “Exploiting Local Similarity for Indexing Paths in Graph-Structured Data,” Proceedings of the 18th International Conference on Data Engineering (ICDE'02), IEEE Computer Society 2002, San Jose California, USA, pp 129-140, Feb. 26-March 1 2002.

[7] 松本 裕治, 北内 啓, 山下 達雄, 平野 善隆, 松田 寛, 浅原 正幸, “日本語形態素解析システム『茶筌』version2.0 使用説明書第二版,” Information Science Technical Report NAIST-IS-TR99012, Nara Institute of Science and Technology, Dec. 1999. http://chasen.aist-nara.ac.jp/.

[8] 中島 誠, 伊藤 哲郎, “質問との概念的関連性をとらえるための文献内容表現の扱い,” 電子情報通信学会論文誌D-I, vol.J85-D-I, no.5, pp.436-444, May 2002.

[9] 大山 敬三, 影浦 峡, 神門 典子, 木村 優, 丸山 克巳, 吉岡 真治, 高橋 一道, “大規模学術情報データベースに適した情報検索システムの開発,” 電子情報通信学会論文誌D-I, vol.J84-D-I, no.6, pp.658-670, June 2001.

[10] Yanhua Qu, Keizo Sato, Makoto Nakashima, and Tetsuro Ito, “Browsing in a Digital Library Collecting Linearly Arranged Documents,” Proceedings of the 24th Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR’2001), New Orleans, USA, pp.426-427, Sept. 9-13 2001.

[11] 曲 艶華, 佐藤 慶三, 中島 誠, 伊藤 哲郎, “電子図書館のための適合可能性示唆によるブラウジング支援,” 電子情報通信学会論文誌D-I, vol.J84-D-I, no.7, pp.1009-1020, July 2001.

[12] 榊 克彦, 宮崎 雅隆, 中島 誠, 伊藤 哲郎, “概念辞書を用いた日本語文献の索引づけ,” 情報処理学会火の国情報シンポジウム発表論文, pp.17-22, March 2002.

[13] G. Salton, Automatic Text Processing, AddisonWesley, Massachusetts, 1989.

[14] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, McGraw-hill, New York, NY, 1983.

[15] Dennis Shasha, Jason Tsong-Li Wang, Rosalba Giugno, “Algorithmics and Applications of Tree and Graph Searching,” Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2002, Madison Wisconsin, USA, pp.39-52, June 3-6 2002.

[16] 鈴木 優, 波多野 賢治, 吉川 正俊, 植村 俊亮, “複数のメディアで構成された電子文書の検索手法,” 情報処理学会論文誌データベースTOD, vol.42, no.SIG10, pp.11-21, Sept. 2001.

[17] 田島 敬史, “半構造データのためのデータモデルと操作言語,” 情報処理学会論文誌データベースTOD, vol.40, no.SIG3, pp.152-170, Feb. 1999.

[18] 高田 伸彦, 田村 武志, 大沢 一彦, “XMLによるWeb上の論文検索システムの構築,” 電子情報通信学会論文誌D-I, vol.J84-D-I, no.6, pp.650-657, June 2001.

[19] Rei-Jo Yamashita, Tetsuro Ito, Hsiu-Hsen Yao, “On Management of Semi-Structured Data for Information System Development,” Proceedings of the 5th International Conference on Information Systems, Analysis and Synthesis (SCI’99/ISAS’99), Orlando Florida, USA, pp.63-67, July 31- Aug. 4 1999.