様式 C-19

科学研究費補助金研究成果報告書

平成 22 年 5 月　17 日現在

研究種目：若手研究（B）
研究期間：２００８～２００９
課題番号：20700093
研究課題名（和文）経験を表現する個人コンテンツの暗黙的リンク構造に基づくデスクトップサーチ手法
研究課題名（英文） A Method for Desktop Search Based on Implicit Link Structure of Personal Content about User's Experience
研究代表者
牛尼　剛聡（USHIAMA TAKETOSHI）
九州大学・大学院芸術工学研究院・助教
研究者番号：50315157

研究成果の概要（和文）：近年，個人が管理する電子メール，デジタルカメラで撮影された写真等の個人コンテンツの飛躍的な増加に伴い，個人コンテンツを対象とした効果的な検索手法が注目されている．個人コンテンツの重要な特徴の一つとして，それが個人の経験と密接に関連していることがあげられる．本研究では，個人コンテンツを個人経験の表現として捉え，個人コンテンツ検索に於いて利用者はコンテンツ自体ではなく，経験を検索していると仮定する．この仮定により，個人コンテンツ間に存在する明示的なリンクと暗黙的なリンクを発見し，与えられたキーワードに関連する経験表現の適切さという観点から，コンテンツ間のリンク構造を利用して個人コンテンツをランキングする手法を開発した．

研究成果の概要（英文）：Recently, according to the remarkable increasing of the number of personal contents, such as e-mail messages and digital photographs and so on, it is required to develop an effective search technique for personal contents. One of the characteristic features of personal contents is that many of personal contents are related to personal experiences. In this work, we treat personal contents as representations of personal experience, and we assume that a user requests to obtain various types of information about personal experiences instead of personal contents on personal contents search. Based on this assumption, we have developed a novel ranking technique for personal contents. This technique uses explicit links and implicit links among personal contents for ranking. The ranking criterion is how sufficiently targets represent personal experiences.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	**1,800,000**	**540,000**	**2,340,000**
2009 年度	**1,500,000**	**450,000**	**1,950,000**
年度			
年度			
年度			
総　計	**3,300,000**	**990,000**	**4,290,000**

研究分野：情報学
科研費の分科・細目：
キーワード：デスクトップサーチ，個人コンテンツ管理，暗黙的リンク

様式 C-19

科学研究費補助金研究成果報告書

１．研究開始当初の背景

近年，ユーザが個人的に管理するデジタルコンテンツ（個人コンテンツ）が増大している．この背景として，デジタルカメラ，デジタルビデオカメラ等の記録装置が広く普及したこと，電子メール，Blog，SNS 等のテキストを利用したコミュニケーションが一般化したこと，音楽や映像などをデジタル化して携帯電話や携帯型オーディオ機器にダウンロードして使用する形態が普及したこと等が挙げられる．こうした中で，大量の個人コンテンツを効果的に管理することの重要性が増大している．

従来，個人コンテンツは，ファイルとして，他のシステムファイルと同様に階層構造を利用して管理されることが一般的であった．しかし，コンテンツが大量化するに従って以下の問題点が存在する．

(1) 大量のコンテンツを管理するためには階層が肥大化し，利用者が階層の全体像を把握できない
(2) 単一のコンテンツがユーザの利用コンテクストに応じて複数の側面を持ち，単一のコンテンツが複数の分類に所属する場合に適切に対応できない

これらの問題点を解決するために，デスクトップサーチツールを利用した個人コンテンツのフリーワード検索が注目されている．従来型のデスクトップサーチ手法では，コンテンツに含まれているテキスト情報を利用してコンテンツに索引付けを行う．具体的には，コンテンツ中に含まれるテキストの他に，ファイル名，メタ・データ，アノテーション等に含まれるテキストを形態素解析し，含まれている単語を索引語として利用する．しかし，検索対象となる個人コンテンツの種類は多様であり，検索対象となるすべての個人コンテンツがテキスト情報を持つわけではない．また，テキスト情報を持っていたとしても，必ずしも期待するテキスト情報を含んでいるとは限らない．

文書に対するランキングには，文書の統計的な特徴を利用することが多く，代表的な手法に TF*IDF 法がある．個人コンテンツのランキングに TF*IDF を適用する場合，個人コンテンツは，論文や Web ページに比べて単語数が少ないため，適切な重み付けができないことが多い．また，メールのようなコミュニケーションコンテンツでは文脈に応じて省略が発生することが多く，検索対象となるメールに関して利用者が与えた検索語を含まないため，適切に検索できないという問題点がある．

２．研究の目的

本研究では，個人コンテンツを，経験コンテンツと非経験コンテンツに分類する．経験コンテンツとは，ユーザ自身の実世界での経験を外化したコンテンツである．一方，非経験コンテンツとは，ユーザ自身の実世界での経験コンテンツであり，ユーザの経験として内化されるものである．本研究では，検索対象を経験コンテンツに限定し，コンテンツからユーザの経験に関する表現を抽出し意味的な内容に基づく構造化を実現した．

本研究で開発する検索手法の概要は以下の通りである．

(1) 個人コンテンツを個人経験の表現として捉える
(2) 個人経験の表現という観点から個人コンテンツ間に関係性を発見する
(3) 発見した関係性に基づいて構造化する
(4) 構造を利用して経験表現の適切性の視点からランク付けする

本研究では，まず，複数の個人コンテンツからユーザの経験を表す表現を抽出する手法を開発する．次に，抽出した表現に基づいて個人コンテンツ間にリンクを設定し，リンクによって得られるコンテンツネットワークの構造から個人コンテンツの重要度を決定する手法を開発する．

本研究では，利用者が個人コンテンツを検索することの目的は，利用者が経験について情報を獲得することであるとする．すなわち，検索対象は個人コンテンツ自体ではなく，個人コンテンツに記録された経験であると考える．これは，利用者が，個人コンテンツを介して，経験に到達するというイメージである．検索において，利用者は検索対象をイメージし，その特徴をキーワードとして与える．つまり，利用者は，個人コンテンツの特徴ではなく，経験に関する特徴を指定すると考える．従来型のコンテンツ検索手法では，コンテンツ自体を検索対象として考えてきたのに対して，本研究では，コンテンツが表す内容を検索対象とする．

従来のテキスト検索においては，利用者が与えたキーワードとテキストの関連の度合いによって重み付けを行う手法が一般的である．それに対して，本研究では，利用者の検索要求は経験であるとするため，利用者が想定する経験に関する情報が最も詳しく述べられている個人コンテンツが最も重要であると考える．検索対象となる内容である経験をモデル化することにより，個人コンテンツに対して，従来型の検索手法とは異なった新しいランキングのアプローチを適用できる．また，これにより，テキスト以外のコンテンツもキーワードで検索可能とする．また，テキスト情報を持たない画像，映像などのコンテンツに対しても，利用者が与えたキーワ

ードから連想的に検索を行い，テキスト情報を持たないコンテンツも検索可能とする．
また，この手法は，単語の頻度だけでなく，PageRank と同様にリンク構造に基づいて検索結果をランク付けする．

本研究では，個人コンテンツは以下の 2 種類の役割を果たす．

(1) 利用者の経験の幾つかの側面を記録し，利用者が情報を得る表現としての役割である．
(2) 利用者のクエリを変換する知識としての役割である．

これまで，多くのコンテンツに関する検索手法ではコンテンツは表現としての側面のみが協調されていた．しかし，本研究では，コンテンツは，検索対象であると同時にクエリを変換する知識としての役割を果たす．

従来型のデスクトップサーチツールによって行われる検索は，コンテンツが含むテキスト情報を利用した検索技術が中心的な要素である．これらは，基本的にコンテンツが含むテキストに対して全文検索を提供するが，テキスト情報を持たないコンテンツに対する検索や，連想的な検索は提供していない．本研究では内容に基づいた暗黙的なリンクによりこれらを提供する．

３．研究の方法

本研究の目標は，ユーザの実世界上での経験を記録した個人コンテンツを対象とした検索に於いて，ユーザにとっての経験記録の適切性という観点からランキング付け可能な検索手法を開発することである．この目標のために，まず，個人コンテンツに含まれる経験表現を洗い出し，コンピュータでそれらの経験表現を自動的に抽出する手法を開発する．また，個人コンテンツ間に存在する意味的な関係を洗い出し，抽出した経験表現に基づいて個人コンテンツ間の意味的な関係（リンク）を自動的に導出する手法を開発する．経験表現の抽出と関係の導出に関しては，実コンテンツを利用して性能を定量的に評価した．さらに，コンテンツから抽出されたリンク構造に基づいて，検索結果となる個人コンテンツの重要度を決定するアルゴリズムを開発する．

４．研究成果

本研究では，個人コンテンツの検索に於いて検索結果のランキングの実現を対象としている．検索結果のランキングには様々な基準が考えられる．従来のテキスト検索においては，利用者が与えたキーワードとテキストの関連の度合いによって重み付けを行う手法が一般的である．それに対して，我々の手法では，利用者の検索要求は経験であるとするため，利用者が想定する経験に関する情報が最も詳しく述べられている個人コンテンツが最も重要であると判断する．

個人コンテンツがどの程度経験を適切に表現しているかを評価するために，経験をモデル化する．利用者が実世界上で行った活動を経験と呼ぶ．経験は，時間(when)，場所(where)，人物(who)，対象(what)，理由(why)，手段(how)によってモデル化する．これらの要素は 5W1H と呼ばれ，一般的に，実世界上の出来事を記載する際に 5W1H が分かるように記載することにより，読者に内容を正確に伝えることが出来ると考えられている．ここで対象としている利用者の経験も事実である．経験を構成する要素を 5W1H で考える．ここで，5W1H を経験要素と呼ぶ．経験要素は，個人コンテンツ中に表現として出現する．個人コンテンツ中に表現された経験要素を経験要素表現と呼ぶ．

個人コンテンツを検索する際の注意点として，一つの経験に関する経験要素表現が複数の個人コンテンツに分散して存在することが挙げられる．単一の個人コンテンツには，経験の特定の側面だけが記載されることが多い．本研究では同一の経験要素表現を含む個人コンテンツや，関連が強い経験要素表現を含む個人コンテンツは同一の経験について表現している可能性が高いと考える．そして，経験要素の関連に基づいて，利用者が希望する経験に関する情報を最も多く含んでいると考えられる個人コンテンツに対して高い評価を与える．例として，利用者が検索語として与えたキーワードを含む 3 つの文書 A,B,C が存在する状況を考える．文書 A には時間情報のみが含まれているとする．文書 B には場所の情報のみが含まれているとする．文書 C には時間と場所の情報が共に含まれているとする．このとき，ユーザが与えたキーワードに関しては，文書 C が経験に関する最も多くの情報を含んでいると考える．上記の例は，単純に経験要素記述の絶対量だけを考えているが，同一の経験要素記述間で重みを伝播させるようにすることで，より高次の重み付けが可能になる．

個人コンテンツ間には 2 種類のリンク構造が存在する．一つは明示的なリンクであり，もう一つは暗黙的なリンクである．本研究では，個人コンテンツ間に存在する明示的なリンクと，暗黙的なリンクを利用する．以下にそれぞれのリンクの特徴について述べる．

(1) 明示的リンク

明示的なリンクの代表例として，Web ページのハイパーリンク，電子メールのリプライ参照，電子メールでのメッセージの部分的引用，ブログのトラックバック，SNS における

足跡等を挙げることができる．例えば，電子メールのリプライ参照は，メールが以前のメールの参照である場合，返信元のメールに対する参照がヘッダ部分に記載される．参照されたメールとは同じ内容について記載されている可能性が高い．電子メールにおける返信には，他にリンク構造を考えることができる．参照元のメールからテキストの一部分を引用することがある．引用されたテキストの先頭には，引用を表す特別な記号を付与することが多い．リプライの参照はメール単位の関連を表しているのに対して，メールの引用は文章単位の参照関係を表している．

(2) 暗黙的なリンク

コンテンツの中には，内容的なつながりを持つものがある．コンテンツが経験を表現しているとするとコンテンツの中に経験要素(5W1H)が含まれる．類似した経験要素を含むコンテンツは，同じ経験を表現していると考え，関連のある経験要素表現をリンクとして考える．経験要素表現によって設定されたリンクを暗黙的リンクと呼ぶ．暗黙的リンクは，設定の基準となった経験要素表現が存在する．基準となった経験要素表現の種類によって，暗黙的時刻リンク，暗黙的場所リンク，暗黙的人物リンク，暗黙的対象リンク，暗黙的原因リンク，暗黙的手段リンク，と呼ぶ．

本研究では，利用者は，特定の経験について興味があることを前提としている．単一の経験に複数の個人コンテンツによって表現される．いま，利用者が興味のある経験に関する個人コンテンツ集合が与えられたとする．利用者は，一つの経験に関する複数のコンテンツを見て，必要な情報を獲得する．

利用者が，検索要求を満足するために行う振る舞いを以下のようにモデル化する．コンテンツが同一の経験について記述していると考えられる場合に，２コンテンツ間に暗黙的なリンクを設定する．利用者は，興味のある経験に関する個人コンテンツを見ている．しかし，利用者が必要としている情報はそこには存在していないばあい，関連する個人コンテンツにアクセスして不足している情報を補う．暗黙的なリンクは，ある個人コンテンツから連想的に他の個人コンテンツに遷移するパスを表す．

例えば，２つのメール A，メール B が共に，ある会議（経験）に関連する内容を含んでいるとする．メール A，メール B は共に 2007/7/2 という時間表現を含んでいる．メール A は，会議の会場（場所）について確認する内容である．メール B は，会議の参加者に関する内容である．メール A を読んでいて，利用者は参加者について知りたくなるかもしれない．メール A,B は同一の時間表現を含んでいるため，それらには暗黙的時間リンクが存在する．このリンクの遷移は，メール A を見ている利用者が，同じ日付を含むメール B を見て参加者を確認する振る舞いを表している．

リンクは，利用者の連想的な個人コンテンツ閲覧の遷移を表している．リンクの元になった経験要素表現自体の特徴や，同一のコンテンツに含まれる他の経験要素表現に依存して，遷移の可能性は変化する．遷移のしやすさをリンクの重みとして捉える．

リンクの重みは概念の包含関係によって定義する．経験要素表現に対して概念領域を与え，それに基づいて重みを導出する．経験要素表現 a の概念領域を r(a)と表記する．2 個の経験要素表現 a_1, a_2 の概念領域の共通部分を $r(a_1) \sqcap r(a_2)$ とする．いま，領域 $r(a)$ の大きさを $size(r(a))$ と表記するとき．経験要素表現 a_1 から a_2 へのリンク l_{a_1,a_2} の重み $w(l_{a_1,a_2})$ を以下のように定義する．

$$w(l_{a_1,a_2}) = \frac{size(r(a_1) \sqcap r(a_2))}{size(r(a_1))}$$

概念領域は個々の経験要素表現に関して定義する必要がある．例えば，時間区間は時間軸上の区間に割り当てることが考えられる．場所概念は地理空間上の平面領域に割り当てることが考えられる．人名は，姓と名に分けることが考えられる．

いま，検索対象とする個人コンテンツ集合を $C = \{c_1, \cdots, c_M\}$ とする．いま，利用者が検索要求として与えたキーワードを含む個人コンテンツ集合を R_0 とする．また，R_0 に含まれる個人コンテンツとの間にリンクが存在する個人コンテンツを R_1 とする．本手法では，R_0 と R_1 の和集合 $R = R_0 \cup R_1$ に含まれる個人コンテンツに対して重要度を考えランク付けをおこなう．なお，R に含まれる要素の数を N で表す．

本研究では，対象とするコンテンツ集合間を利用者が閲覧しながら情報を確認するとし，重要なコンテンツにはリンクによる遷移が集中して存在確率が高くなると仮定する．利用者が，個人コンテンツ集合をリンク構造に基づいてブラウジングするとする．時刻 t において個人コンテンツ c_i を閲覧している確率を $p_{i,t}$ と表現する．時刻 t における集合 R に含まれる個人コンテンツの閲覧確率を列ベクトル $\mathbf{p}_t = (p_{1,t}, \cdots, p_{n,t})^T$ と表現する．

個人コンテンツ c_i から c_j のリンクを $l_{i,j,1}, l_{i,j,2}, \cdots, l_{i,j,N}$ と表記する．また，リンク l の重みを $w(l)$ と表記する．

時刻 t において c_i を閲覧している利用者が，時刻 t+1 に於いて c_j に遷移する確率を $e_{i,j}$ と表記し，その値を以下のように定義する．

$$e_{i,j} = \frac{\sum w(l_{i,j})}{\sum w(l_i)}$$

ここで，$\sum w(l_{i,j})$はc_iからc_jへのリンクの重みの総和を表し，$\sum w(l_i)$はc_iから出ているリンクの総和を表している．ここで，遷移確率$e_{i,j}$から構成される遷移確率行列

$$E = \begin{pmatrix} e_{1,1} & e_{1,2} & \cdots & e_{1,N} \\ e_{2,1} & e_{2,2} & \cdots & e_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ e_{N,1} & e_{N,2} & \cdots & e_{N,N} \end{pmatrix}$$

を利用すると，存在確率$\mathbf{p}_{t+1}$は以下のように表現できる．

$$\mathbf{p}_{t+1} = E^T \mathbf{p}_t$$

さらに，リンクが存在しない個人コンテンツ間で遷移が起こる確率をダンピングファクタｄとして与えると，存在確率$\mathbf{p}_{t+1}$は，以下

$$\mathbf{p}_{t+1} = d(1/N)\mathbf{p}_t + (1-d)E^T \mathbf{p}_t$$

のように表現できる．
上記の遷移を繰り返すと遷移確率ｐは一定の値に収束する．収束したそれぞれの存在確率を個人コンテンツの重要度とみなす．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 3 件）

① Taketoshi Ushiama and Toyohide Watanabe: A framework for personal content search and recommendation based on personal experiences, International Journal of Advanced Intelligence Paradigms, Vol. 1, No. 2, pp. 151-162 (2009).

〔学会発表〕（計 14 件）

① Taketoshi Ushiama and Toyohide Watanabe: "X-Web: A Data Model for Managing Personal Contents Based on User Experiences", Proc. of Int'l Conf. KES2008, LNCS 5178, pp. 798-805 Zagreb, Croatia, (2008)

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）
○取得状況（計 0 件）

６．研究組織
(1)研究代表者
牛尼　剛聡（USHIAMA TAKETOSHI）
九州大学・大学院芸術工学研究院・助教
研究者番号：50315157