経験コンテンツ検索における意味的リンク構造を利用したランキングに関する考察

A Study on Ranking of Experience Content Search Based on the Semantic Link Structure

牛尼 剛聡♥ 渡邊 豊英♦
Taketoshi USHIAMA Toyohide WATANABE

近年，個人が管理する電子メール，デジカメ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または暗黙的なリンクを利用し，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の可能性について考察する．

Recently years, 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 paper, 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 discuss 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.

1. はじめに

近年，デジタルカメラ，デジタルビデオカメラ，カメラ付き携帯電話などの記録装置が普及し，気軽にデジタルイメージを撮影できるようになった．また，電子メールは一般的な通信手段として認知され，Blog や SNS によるコミュニケーションも一般化し，個人のコミュニケーションを記録したテキストも増加の一途を辿っている．さらに，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利用形態が広く普及した．一方，ハードディスクの大容量化及び低価格化により，個人で大量のデジタルコンテンツが蓄積しておくことが現実的になった．しかし，大量のデジタルコンテンツが蓄積可能になった反面，大量のデジタルコンテンツの中から必要なコンテンツを探し出すことが困難になるという新たな問題が生じている．こうした背景の下，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[1]．なお，本論文では，個人コンテンツとは個人が所有し，個人で管理するデジタルコンテンツを指すものとする．

現在，我々は，個人コンテンツを効率的に管理し，効果的に活用可能なシステムの開発に取り組んでいる．本論文では，検索による個人コンテンツの効率的な管理を目的とし，個人コンテンツを，ユーザの検索要求に基づいてランキングする新しいアプローチを提案し，その可能性について議論する．

2. 個人コンテンツと検索

2.1 経験表現とコンテンツ

個人コンテンツには様々な種類があり，様々な分類が考えられる．本研究では，個人コンテンツを以下の２種類に分類する．

(1) 自分の経験と深く関係するコンテンツ
(2) 自分の経験と関係しないコンテンツ

前者は，経験コンテンツ，後者を非経験コンテンツと呼ぶ．経験コンテンツの具体例として，電子メール，デジタル写真などがある．非経験コンテンツの具体例として，CD からリッピングした楽曲等がある．本論文では，経験コンテンツに対象を限定し，単に個人コンテンツといった場合には経験コンテンツを指すものとする．

単一の個人コンテンツは複数の経験について表現していることがある．単一の経験は複数の個人コンテンツによって表現される．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，経験に到達するイメージである．

検索において，利用者は検索対象を想像し，その特徴をキーワードとして与える．本研究では，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．提案手法では，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が適用できる．

2.2 経験要素と個人コンテンツの評価

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い，ランキングする手法が一般的である．一方，我々は，利用者の検索要求は経験であるとし，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

♥ 正会員 九州大学大学院芸術工学研究院
ushiama@design.kyushu-u.ac.jp
♦ 正会員 名古屋大学大学院情報科学研究科
watanabe@is.nagoya-u.ac.jp

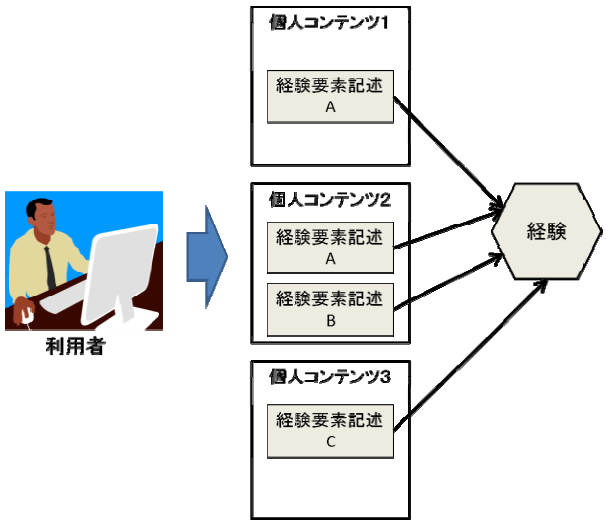

図 1: 個人コンテンツ内の経験要素記述と経験の関連

Fig.1: Relationship between an experience unit and experience element descriptions in personal contents

であると考える．利用者から見た個人コンテンツ，経験要素記述，経験の関係を表す概念図を図 1 に示す．

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．本論文では，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．この考えに基づき，経験を構成する要素を 5W1H で捉える．ここでの 5W1H を経験要素と呼ぶことにする．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．この例では，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．次節で，この考え方に基づいたランク付け手法について述べる．

3. 個人コンテンツ間のリンク

本論文では，経験に基づいた個人コンテンツのランキング手法を提案する．本手法は個人コンテンツ間に設定されたリンク構造に基づいて重要度を判断する．

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，もう一つは暗示的なリンクである．提案手法では，個人コンテンツ間に存在する明示的なリンクと，暗示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．

3.1 明示的リンク

個人コンテンツ中の明示的なリンクについて述べる．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，リプライ参照の他にも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．

3.2 暗示的なリンク

一方，コンテンツの中には，内容的な関連を持つものがある．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る．

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示的リンクと呼ぶ．暗示的リンクは，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，暗示的時刻リンク，暗示的場所リンク，暗示的人物リンク，暗示的対象リンク，暗示的原因リンク，暗示的手段リンク，と呼ぶ．経験要素表現を含むコンテンツは多数存在する．

3.3 経験要素表現の自動抽出

暗黙的なリンクを自動的に設定するためには，経験要素表現を自動的に抽出する必要がある．自動的抽出できる要素およびその方式は個人コンテンツの種類によって異なる．電子メールなどのテキスト情報に関しては，時間概念や空間概念を自動的に抽出するための手法が提案されている[2][3]．また，対象は人名以外の固有名詞，手段は動詞を中心に抽出することが考えられる．具体的な抽出手法の開発とその評価は今後の課題である．

4. リンク構造に基づくランキング

4.1 リンクの意味

提案手法では，利用者は，特定の経験について興味があり，単一の経験は，複数の個人コンテンツによって表現されることを前提とする．

いま，利用者が興味をもつ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．提案手法では，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コンテンツ間に暗示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場合，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．暗示的なリンクは，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，それらには暗示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．

4.2 リンクの重み

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$\mathrm{r}(a)$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$\mathrm{r}(a_1) \sqcap \mathrm{r}(a_2)$とする．いま，領域$\mathrm{r(a)}$の大きさを$\mathrm{size(r(a))}$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$\mathrm{w}(l_{a_1,a_2})$を以下のように定義する．

$$\mathrm{w}(l_{a_1,a_2}) = \frac{\mathrm{size}(\mathrm{r}(a_1) \sqcap \mathrm{r}(a_2))}{\mathrm{size}(\mathrm{r}(a_1))} \quad (1)$$

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．本論文では，概念範囲設定の具体的な手法については言及しない．

4.3 リンクの重みに基づいた重要度の計算

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$R_0 \subseteq \mathrm{C}$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$R_1 \subseteq C$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とする．

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．このアプローチはWebページの重要度を計算するPageRankアルゴリズム[4]と同一である．PageRankアルゴリズムと提案手法の比較は5節で行う．

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$\boldsymbol{p}_t = (p_{(1,t)}, \ldots, p_{n,t})^{\mathrm{T}}$と表現する．個人コンテンツ$c_i$から$c_j$のリンクを$l_{i,j,1}, l_{i,j,2}, \ldots, l_{i,j,N}$と表記する．また，リンク$l$の重みを$\mathrm{w}(l)$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．

$$e_{i,j} = \frac{\sum \mathrm{w}(l_{i,j})}{\sum \mathrm{w}(l_i)} \quad (2)$$

ここで，$\sum \mathrm{w}(l_{i,j})$は$\mathrm{c_i}$から$\mathrm{c_j}$へのリンクの重みの総和を表し，$\sum \mathrm{w}(l_i)$はc_iから出ているリンクの総和を表している．

遷移確率$e_{i,j}$から構成される遷移確率行列

$$\boldsymbol{E} = \begin{pmatrix} e_{1,1} & e_{1,2} & \cdots & e_{1,N} \\ e_{2,1} & e_{2,2} & \cdots & e_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ e_{N,1} & e_{N,2} & \cdots & e_{N,N} \end{pmatrix} \quad (3)$$

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

$$\boldsymbol{p}_{t+1} = \boldsymbol{E}^{\mathrm{T}} \boldsymbol{p}_t \quad (4)$$

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$\boldsymbol{p}_{t+1}$は，以下のように表現できる．

$$\boldsymbol{p}_{t+1} = d\left(\frac{1}{N}\right)\boldsymbol{p}_t + (1-d)\boldsymbol{E}^{\mathrm{T}}\boldsymbol{p}_t \quad (5)$$

上記の遷移を繰り返すと遷移確率 $\boldsymbol{p}$ 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. 考察

5.1 提案アプローチの特徴

提案手法では，個人コンテンツは2種類の役割を果たしている．一つ目の役割は，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．二つ目の役割は，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．提案手法では，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している．知識としての役割を果たす事ができるためには，利用者の情報要求に対するに対するモデルが必要である．本研究では，利用者の情報要求は利用者の経験であるとし，経験を 5W1H でモデル化することにより，クエリを知識として利用することが可能となった．

現在，個人コンテンツ管理のために幾つかのデスクトップサーチツールが提供されている．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．提案手法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．

5.2 Web ページ検索との比較

Web ページ検索ではロボットを利用した全文検索型のサーチエンジンが広く利用されている． Web ページを対象にして全文検索を行う場合，利用者が与えた検索語を含むページは膨大であることが普通である．つまり，再現率は高いが適合率が低いと考えられる．Web ページ検索においては，適合率を向上させるための技術が重要である．

個人コンテンツ検索では，検索結果は大量ではあったとしても，ウェブページほどではない．個人コンテンツ検索では，再現率を向上させることが重要である．コンテンツによっては，利用者が与えたキーワードにマッチしないような対象でも関連するコンテンツを検索結果に含める事が必要である．

しかし，単に再現率を向上させるだけでなく，候補を多数提供しつつ，それらに対して適切なランク付けを行うことが重要である．

5.3 PageRank アルゴリズムとの比較

提案手法はコンテンツ間に設定したリンク間で重みを伝搬させ，複数回遷移させて収束した値をコンテンツの重みであると判断している．このアプローチは，代表的な Web 検索エンジンである Google において利用されているランキングアルゴリズムである PageRank アルゴリズム[4]に類似している．

5.3.1 PageRank アルゴリズム

PageRank はハイパーリンクに基づくページのリンク構造を利用して，Web ページの重要度を計算するアルゴリズムである．PageRank の基本的な考え方は，ページ間に存在するリンクに基づいて重みを伝播させるものである．例えば，リンク元ページp_0がn種類のリンクを有するとし，p_0に与えられたランクがrであるとする．リンク先のページに分配される重みはr/nとなる．リンクが存在しないところに対しても，一定の確率（ダンピングファクタ）で重みを分配させる．このようなリンクに基づく重みの伝搬を再帰的に計算すると一定の値に収束する．この収束した値をそのページの重要度と見なす．

5.3.2 PageRank との相違点

提案手法は，基本的に PageRank と同様にページのリンク構造に基づいて重みを伝搬させる．個人コンテンツ間には Web ページのように明示的なリンクは存在しないため，提案手法では内容に基づいた暗示的リンクも利用する．PageRank アルゴリズムでは，リンクの質を考慮せず，すべてのリンクに対して公平な価値が与える．すなわち，一つのノードから複数のエッジが出現する場合，それらに分配される重みは同一であるとし，リンク元ページからリンク先ページに平均的に分配する．提案手法では，リンクに重みをつけて，伝播させる重みの量を変化させている．

5.3.3 その他の関連研究

甲谷ら[5] らはパーソナルコンピュータに蓄積されたコンテンツに対して PageRank を適用する手法を提案している．この手法では，Web ページの PageRank から，文書の類似度を利用して，リンク構造を持たないテキストの重要度を計算する．このアプローチは，インターネット上で重要であると判断されたページに，類似した個人コンテンツは重要であるという仮定に基づいている．この手法は，客観的な事実を記述した文書では有効であると思われるが，個人の経験を検索する場合には適用できないと予想される．

中谷ら[6]は文書間類似度とキーワードを利用して，Web ページ間に関連リンクを自動的に生成する手法を提案している．この手法では，単語の出現頻度に基づいてページ間の類似度を判断し，類似したページ間にリンクを生成する．この手法では，一般的な Web ページを対象としているため，ページ間の類似度という 1 種類のリンクしか考えていない．これに対して，本手法では，個人の経験についての情報を得るための個人コンテンツ検索を対象とすることにより，5W1H という 6 種類の関連を考えている．個人コンテンツはテキストデータが短く，Web ページのように単語頻度に基づく類似度が正確に獲得できないことが多いため，提案手法の方が個人コンテンツを対象とした場合には有効であると考えられる．

向ら[7]は利用者の Web ページの閲覧履歴に基づいて PageRank における Web ページの遷移確率を変更する手法を提案している．この手法では，履歴情報を利用するため，新規に導入されたコンテンツに対して適用できない．また，個人コンテンツは Web ページのようなパブリックコンテンツと異なり，同一のコンテンツを利用する頻度が少ないため十分な履歴情報を取得することが困難であるため，個人コンテンツ検索への適用は難しい．

6. まとめ

本論文では，利用者の経験表現という視点から個人コンテンツ間にリンクを設定し，リンク構造に基づいて個人コンテンツ検索結果をランク付けする手法を提案した．

今後，コンテンツに含まれる経験要素の量について調査し，自動的な経験要素表現の抽出方法を開発する予定である．また，本研究では個人コンテンツ間の暗黙的なリンクをランキングに利用したが，個人コンテンツの有効的な活用という視点から考えると，ランキング以外の利用も考えられる．例えば，暗黙的なリンクに基づいて利用者が関連するコンテンツを閲覧可能な対話環境を提供することが考えられる．これらの可能性についても検討していく．

[文献]

[1] Teevan, J., Jones, W. and Bederson, B. B.: Personal Infromation Management, Communications of the ACM, Vol. 49, No. 1, pp. 40-43 (2006).
[2] 石川幹直，細川宜秀，高橋直久：ドキュメント・データを対象にしたタイムコーディングシステムの実現方式， 第 14 回データ工学ワークショップ(DEWS2003) (2003).
[3]相良毅，有川正俊，坂内正夫：ジオリファレンス情報を用いた空間情報抽出システム，情報処理学会論文誌，Vol.41, No. Sig6, pp. 69-80 (2000).
[4] Page, L., Brin, S., Motwani, R. and Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web, Technical report, Stanford Digital Library Technologies Project (1998).
[5] 甲谷優，湯本高行，小山聡，田島敬史，田中克己：Web ページの PageRank 値に基づくローカルコンテンツの品質推定, DEWS2006 予稿集 (2006).
[6] 中谷圭吾，鈴木優,川越恭二：文書間類似度とキーワードを用いた Web リンク自動生成手法, DBSJ Letters, Vol. 4, No. 1, pp. 85-88 (2005).
[7] 向亨，成凱，上林彌彦：利用履歴に基づく PageRank アルゴリズムの改良,DEWS2002 予稿集 (2002).

牛尼 剛聡 Taketoshi USHIAMA
九州大学大学院芸術工学研究院助教．1999 名古屋大学大学院工学研究科情報工学専攻博士課程後期課程単位取得退学．博士（工学）．情報処理学会．電子情報通信学会，IEEE-CS, ACM 会員．

渡邉 豊英 Toyohide WATANEBE
名古屋大学大学院情報科学研究科教授．1974 京都大学大学院工学研究科修士課程修了．1975 同大学工学研究科博士課程中退．工学博士．電子情報通信学会（フェロー），情報処理学会，日本ソフトウェア科学会，人工知能学会，システム制御情報学会，ACM，IEEE-CS，AAAI，AACE 各会員．