データベースシステム　126-22
（２００２．１．２２）

# XML データ群の個人化とその構成最適化について

高野正樹[†]　鈴木優[†]　波多野賢治[†]
吉川正俊[†,††]　植村俊亮[†]

計算機の普及とネットワーク環境の整備によって，商品のカタログなどの情報が電子的に提供されるようになり，今後これらの情報が XML 形式で提供されることが予想される．提供される情報量が膨大になれば，これらの情報の中から利用者が注目しており，また新鮮な情報だけを自動的に取り出したいという要求が生まれる．そこで，本稿ではこのような要求を実現するために，利用者の情報への参照履歴や提供される情報の内容から得られる四つの特徴量を用いて，利用者が必要とする情報を取り出す手法を提案する．また，複数の特徴量をどのような評価関数によって統合すれば良いかを XML データを対象とした実験によって示す．その結果，評価関数として $p$-norm が良いことがわかった．

# XML Document Personalization based on User's Preference and Document's Freshness

Masaki Takano,[†] Yu Suzuki,[†] Kenji Hatano,[†]
Masatoshi Yoshikawa[†,††] and Shunsuke Uemura[†]

High-speed network and high-performance computer enables us to deal with many eletronic documents such as XML documents are used widely. On E-commerce, a lot of information about merchandise are broadcased over the networks, and stored into users' storages. However, users' cannot view that huge data, we need an algorithm to retrieve data which users need. In this paper, we propose a method to retrieve data which a user want automatically. In our method, we extract two features: users' preference and documents' freshness by integrating two features, respectively. We have performed experiments to find an appropriate mathematical function used in integrating score values of multipule features into single scores. As a result, we found that "the $p$-norm" exhibits the best result in our experiments.

## 1. はじめに

計算機の普及や，インターネットをはじめとするネットワーク環境の整備により，商品のカタログや広告などの情報がネットワーク上で電子的に配信されるようになりつつある．このような情報の配信方法としては，主に Web や電子メールが利用されているが，近年では，携帯端末を持つ移動中の利用者に情報配信を行うことも可能となっている．さらに，ユビキタスコンピュータやウェアラブルコンピュータが普及することで，駅のホームに貼られたポスターの内容や時刻表などがウェアラブルコンピュータに自動的に配信されるようになることも考えられる．このような環境が実現すれば，利用者の日常生活で得られるさまざまな情報の取得形態が，これまでの利用者による意識的な情報の取得から，計算機による自動取得に変化する．

上記のように，日常生活で得られる情報を自動取得することによって，利用者はこれまで以上に多くの情報を得られるようになると考えられる．しかし，携帯端末の記憶装置の容量には物理的な制限があり，画像情報や映像情報などを含むマルチメディア化されたコンテンツを全て蓄積することは困難である．また，利用者は取得した全ての情報を常に利用するとは限らない．そこで，取得した情報を大容量の記憶装置に退避させ，必要に応じて利用したい情報だけを手元の携帯端末に読み出すといった作業が必要となる．ところが，蓄積される情報量が膨大なものとなれば，このような作業を手動で行うことが困難となり，膨大な情報の中から利用者が必要としている情報を自動的に抽出する技術が必要となる．利用者が参照する情報は，すでに

† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology
†† 国立情報学研究所ソフトウェア研究系
Software Research Division, National Institute of Informatics

蓄積されていた情報や新規に取得した情報，利用者の参照頻度が高い情報や低い情報などさまざまな特徴を持っている．これらの特徴を組み合わせて，それらの情報に対する利用者の要求を数値化すれば，利用者の必要としている情報を自動的に抽出することが可能となる．ここで問題となるのは，これらの特徴をどのように組み合わせれば良いかということである．

そこで，本研究では，利用者が取得する情報から得られる特徴量として，利用者の情報への参照履歴を考慮した注目度と，配信されている情報の内容を考慮した新鮮度を利用し，これらの二つの特徴量の組み合わせる手法を提案する．本研究では，ある情報から得られた二つの特徴量を組み合わせた値を必要度と呼び，利用者がその情報を必要とする度合を表す値として定義する．また，特徴量を組み合わせる手法の評価実験を行う際に，近年さまざまなアプリケーションの情報交換の共通フォーマットとして用いられている XML を利用した．そして，XML で記述されたデータが，ユビキタスコンピュータからウェアラブルコンピュータへの情報配信において提供されるであろうという環境を想定して，利用者の必要とする情報を配信されている XML データ全体ではなく，部分文書単位で抽出した．

## 2. 基本的事項と関連研究

本節では，本研究で提案するシステムに必要な要素技術について述べる．

### 2.1 文脈ノード

本研究では， XML データの中から，利用者が必要とする情報を部分文書単位で抽出するため，XML データを部分文書に分割して蓄積する．このとき，それらの部分文書が情報の意味的なまとまりとなっていなければ，利用者が情報をまとまった形で参照することができなくなる．

我々は，文献 1) において，スキーマ情報を利用することができない XML 文書を対象とした検索手法として， XML 文書中の文脈のまとまりを表現する文脈ノードを定義し，文脈ノードを根ノードとする部分文書単位での検索が可能な「文脈検索」という手法を提案している．そこで，本研究では，情報を意味的なまとまりごとに部分文書に分割するために，文脈ノードを用いた．我々は文脈ノードを次のように定義している．

**文脈ノード**　XML 文書 $D$ の中に存在するテキストノードまたは属性ノード $n$ の文脈ノード $context(n)$ は次のように定義される．

1. $n$ が属性ノードのとき，$context(n)$ は $n$ の親の要素ノードとなる．
2. $n$ がテキストノードとき，$context(n)$ は次のように定義される．まず，$n$ の祖父母ノードを $g(n)$ とする．$D$ の最上位のノード $n_d$ と，$g(n)$ との間に存在し，同じ要素名を兄弟ノードに持つノードで最下位のノード $m$ が存在すれば，$m$ が $context(n)$ となる．
   もし，$m$ のようなノードが存在しなければ，$context(n)$ は $n_d$ となる．

文脈ノードを根ノードとする部分文書には，**図 1** の $context(n_1)$ ように，一つの部分文書内に複数の文脈ノードを含む場合がある．本研究では，内部に文脈ノードを持たない部分文書を情報の最小単位と考え，この最小単位を本研究における部分文書として定義する．

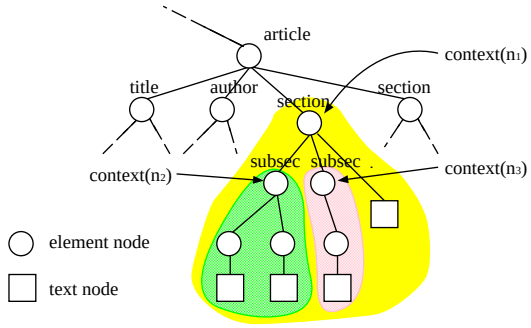


**図 1**　文脈ノード
Fig. 1　Context node.

### 2.2 部分文書の類似度

部分文書から得られる特徴量の一つである新鮮度は，他の部分文書との類似度に基づいて計算される．そのため，二つの部分文書がどれほど類似しているかを求める必要がある．

我々は文献 2) において XML 文書の類似度の計算手法を提案しており，本研究ではこの手法を用いて部分文書の類似度を計算する．

XML 文書 $D_i$ と $D_j$ の類似度 $sim(D_i, D_j)$ は次のように定義される．

#### 2.2.1 Step 1: ノードから接尾辞の抽出

XML 文書 $D_i$ はノード $\{n_1(D_i), n_2(D_i), \cdots\}$ を持ち，各ノードには根ノードからのパスが存在する．ある文書 $D_i$ に存在するあるノード $n \in \{n_1(D_i), \cdots\}$ へのパスを $a(n)$ として定義する．

$$a(n) := /n_r/a_1/a_2/\cdots/a_{k(n)-1}/n$$

ここで，$\{a_1, a_2, \cdots, a_{k(n)-1}\}$ は ノード $n$ と根ノード $n_r$ の間に存在するノードであり，$k(n)$ はルート

ノード $n_r$ からノード $n$ までのノードの数である．

また，パス $a(n)$ の接尾辞 $J(a(n))$ を次のように定義する．

$$J(a(n)) = \{J_1(a(n)), J_2(a(n)), \cdots, J_{k(n)}(a(n))\}$$

ここで，

$$J_i(a(n)) = /n_r/\alpha_1/\cdots/\alpha_{k-1}$$

つまり，ルートノードからノード $n$ までのパスのうち，$i$ 番目までのノードまでの接尾辞をとる．

#### 2.2.2 Step 2: 接尾辞を考慮した文書間の類似度

次に，部分文書の類似度を XML 文書の特徴ベクトルの類似として計算するために，全ての XML 文書から XML 文書の特徴ベクトルの基底を求める．ある文書 $D_i$ における特徴ベクトルの基底 $VB(D_i)$ は次のように定義できる．

$$VB(D_i) = \bigcup_n J(a(n))$$

つまり，文書の集合 $DB = \{D_1, D_2, \cdots, D_N\}$ における特徴ベクトルの基底 $VB$ は次のようになる．

$$VB = \bigcup_{i=1}^{N} VB(D_i)$$

この基底に従って，XML 文書 $D_i$ の特徴ベクトル $\boldsymbol{f}(D_i)$ を求めると，次式のようになる．

$$\boldsymbol{f}(D_i) = [f(J_1(a(n)), D_i), f(J_2(a(n)), D_i), \cdots]$$

あるパスを $path \in VB$ とおくと，$D_i$ の特徴ベクトルにおける各要素は，

$$f(path, D_i) = \frac{tf(path, D_i)}{|D_i|}$$

と定義される．ここで，$|D_i|$ は XML 文書 $D_i$ 中のパスの数で，$tf(path, D_i)$ は XML 文書 $D_i$ 中の $path$ の出現回数である．

最後に，XML 文書 $D_i$ と $D_j$ の類似度 $sim(D_i, D_j)$ は，コサイン相関値を用いて次の式で計算できる．

$$sim(D_i, D_j) = \cos(\boldsymbol{f}(D_i), \boldsymbol{f}(D_j))$$

### 2.3 新 鮮 度

Ceri らは，商品のカタログなどを利用者に提示する際に，利用者の過去の参照履歴などを用いて利用者が必要とするであろう情報を提供する手法を提案している [3]．また，利用者が必要であろうと思われる情報を提供する方法として，利用者が過去に参照した情報をプロファイル化し，プロファイルと類似した情報を提供したり，利用者の問合せに対して，あらかじめシステム側で用意された関連する情報を提供する方法などが Kramer によって提案されている [4]．これらの研究では，利用者の情報への参照履歴から得られる特徴を用いて利用者にどのような情報を提供するかを決定しており，提供する情報の内容から得られる特徴は，単に参照履歴に存在する情報との類似性しか考慮していない．しかし，宮崎らは，利用者が過去に蓄積した情報と，新規に取得した情報の類似性を利用して，新規に得られた情報に対する新鮮度や流行度などの新たな特徴量を定義し，新規に得られた情報が利用者にとって有用であるかを判断している [5]．

本研究で扱うデータは，商品のカタログなどの情報であり，情報の内容の更新など変化が反映される新鮮度も重要な特徴量の一つとなると考えられる．そこで，利用者の新規に得られた情報に対する必要度を求めるために，文献 5) で定義されている四つの新鮮度のうちの一部を利用した．

## 3. XML データの評価値の計算手法

利用者が携帯端末上に情報を蓄積して使用する場合，携帯端末上の記憶装置は，利用者が必要としている情報だけが蓄積された状態となっている．ところが，利用者が携帯端末上に新しい情報を蓄積したり，携帯端末上の情報への参照を繰り返すうちに，はじめは必要な情報だけしか格納されていなかった携帯端末上に不要な情報が出現したり，過去に必要だった情報が不要となることがある．

このように，新規の情報が得られた場合や，利用者の情報の利用状態によって情報の必要度が変化した場合，携帯端末上で不要になった情報をバックエンドのサーバに退避させ，新たに必要となった情報だけをバックエンドのサーバから携帯端末に読み出すといったデータ交換が必要である．

本研究で提案するシステムでは，上記のようなデータの交換を行うために，必要度を用いて利用者がある情報を必要としているかどうかを判断する．必要度は，利用者がある情報を必要とする度合を表しており，本研究では必要度を決定するための要素となる特徴量として，注目度，新鮮度の二つを用いる．注目度は利用者の情報への参照状態に依存し，新鮮度は蓄積された情報の内容に依存している．このように，利用者指向の特徴量と情報指向の特徴量とを統合することで，より利用者が必要とする情報を得ることができると考えられる．しかし，これらの特徴量から必要度を導くために特徴量をどのように統合するかを決定することは容易ではない．

また，情報が XML で配信されている場合を想定し，2.1 節で述べた文脈ノードの定義に基づいて，利用者が取得した XML データを部分文書単位で蓄積し，それぞれの部分文書ごとに注目度や新鮮度などの特徴量

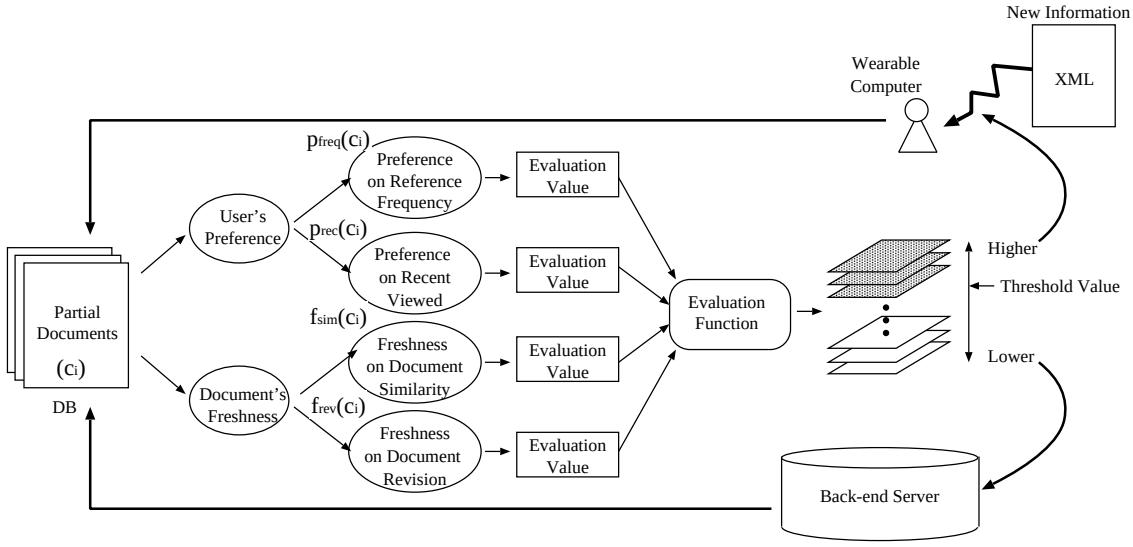


**図 2** システムの概要
Fig. 2 Overview of our system.

を計算して，部分文書単位で利用者が必要とする情報を抽出する．注目度は，ある情報に対して，利用者がどれほど注目しているかを表す特徴量であり，その値は，過去に蓄積された情報に対する利用者の参照履歴によって変化し，利用者の情報の参照状態に依存している．また，新鮮度は，利用者が新規に取得した情報が過去に蓄積されている情報と比較してどれだけ新しいものかを表す特徴量である．注目度は，利用者の情報の参照状態に依存しているため，これまで蓄積されている情報に対する評価は可能であるが，利用者がこれまで一度も参照したことのない情報に対する評価は常に低くなるという問題がある．新鮮度は，新規に取得した情報と過去に蓄積された情報との類似度に基づいて計算されるため，上記のように注目度で評価しきれない部分を必要度に反映させることができる．

### 3.1 システムの概要

本システムでは，利用者が取得した全ての情報に対し，注目度と新鮮度の特徴量が付与される．全ての情報は注目度と新鮮度を統合した必要度を用いてランキングされる．また，不要な情報を退避させるバックエンドのサーバには，過去に利用者が取得した情報が蓄積されている．利用者が新たに取得した情報を退避させる際には，新たに取得した情報と，過去に取得した情報をあわせて必要度を再計算し，再計算された必要度に基づいてランキングを行う．その結果，上位に位置する情報から，あらかじめ決められた容量の情報だけを利用者が必要な情報として取り出す．その結果，利用者が必要な情報が再構成されることになる．**図 2** に本システムの概要を示す．

### 3.2 評価値とその統合方法

本節では，必要度を求めるための特徴量として用いる注目度と新鮮度について述べる．まず，利用者が蓄積した部分文書の集合 $DB$ を次のように定義する．

$$DB = \{c_1, c_2, \cdots, c_i, \cdots, c_n\}$$

ここで，$c_i$ は $DB$ 中に含まれる部分文書である．以下にそれぞれの特徴量について述べる．

#### 3.2.1 注 目 度

注目度とは，利用者がこれまで蓄積した情報の中で，現在どのような情報に対して注目しているかを表すものである．本研究では，注目度として参照頻度による注目度 $p_{freq}(c_i)$ と，時間的注目度 $p_{rec}(c_i)$ の二つの注目度を定義し，それらを部分文書ごとに計算する．参照頻度による注目度は，ある情報を参照した回数によって決定され，参照頻度の高い情報はより注目度が高くなる．しかし，利用者の情報への参照状態は時間的に変化しており，過去に頻繁に参照していた情報でも，現在はあまり参照されていない情報は必要性が低いと考えられる．そこで，時間的に変化する利用者の情報への注目度を反映するために，時間的注目度を定義した．

**参照頻度による注目度** 参照頻度による注目度は，利用者がこれまで参照した部分文書の中から，より多く参照しているものが高い値となるように定義する．ある部分文書 $c_i$ についての参照回数を $r_{freq}(c_i)$ とすると，$c_i$ の参照頻度による注目度は次のように定義される．

$$p_{freq}(c_i) = \frac{r_{freq}(c_i)}{\sum_{j=1}^{n} r_{freq}(c_j)}$$

**時間的注目度**　時間的注目度とは，最近利用者がどのような情報を頻繁に参照しているかを表す特徴量である．蓄積された部分文書の集合を $DB$ として，最後に参照した部分文書から数えて $j$ 番目に参照した部分文書は $c_j \in DB(j = 1, 2\cdots)$ となる．また，$Time(c_j)$ は $c_j$ を参照した順序であり，$Time(c_j)$ の値が小さいほど，最近参照した部分文書であることを示す．$Time(c_j)$ を次のように定義する．

$$Time(c_j) = j$$

$c_i$ の時間的注目度 $p_{rec}$ を次の式で定義する．

$$p_{rec}(c_i) = 1 - \prod_{j=1, c_j=c_i}^{n} \left(1 - \frac{1}{Time(c_j)}\right)$$

#### 3.2.2 新　鮮　度

蓄積された部分文書集合 $DB$ に対する部分文書 $c_i$ の新鮮度として，$DB$ 内の類似する部分文書の数による新鮮度（$f_{sim}(c_i)$）と，$DB$ 内の類似部分文書との時間距離に基づく新鮮度（$f_{rev}(c_i)$）をそれぞれ部分文書ごとに求め，最後に二つの新鮮度を先に述べた二つの注目度と共に統合し，必要度とする．

**類似する部分文書の数による新鮮度**　文献 5) では，ある情報と類似する情報が，過去に蓄積された情報の中に多く存在していれば，その情報の新鮮度は低いと定義している．本研究では，この定義を部分文書に適応し，類似する部分文書の数による新鮮度を定義する．

まず，過去に蓄積された部分文書集合 $DB$ の中に存在する部分文書 $c_i$ と $c_k$ の類似度を 2.2 節で述べた方法を用いて計算する．ここで，

$$m(c_i, c_k) = \begin{cases} 0 & \text{if } sim(c_i, c_k) = 0 \\ 1 & \text{if } sim(c_i, c_k) > 0 \end{cases}$$

とすると，$DB$ 中に存在する $c_i$ と類似する部分文書の数 $M_i$ は次のようになる．

$$M_i = \sum_k m(c_i, c_k)$$

類似する部分文書の数による新鮮度は次の式で定義される．

$$f_{sim}(c_i) = \frac{1}{\log_2(2 + M_i)}$$

**時間距離による新鮮度**　文献 5) では，時間距離による新鮮度は，新規に得られた情報が，過去に同じ配信元から得られた情報と時間的に離れれば離れるほど新鮮度が高くなると定義している．つまり，新規に得られた部分文書 $c_i$ が，過去に得られた類似部分文書 $c'_j$ との平均時間距離が大きいほど，$c_i$ が新しい内容を含んでいると考えられる．$t(c_i)$ を $c_i$ が取得された時間とすると，$c_i$ の類似部分文書集合 $d = \{c'_1, c'_2, \cdots, c'_j, \cdots, c'_m\}$ との時間距離に基づく新鮮度は次の式で定義される．

$$f_{rev}(c_i, d) = \log\left(\frac{1}{m}\sum_{j=1, c'_j \in d}^{m} (t(c_i) - t(c'_j))\right)$$

#### 3.2.3 特徴量の統合方法

**偏差値を用いた特徴量の統合**　統合する四つの特徴量は，それぞれの値の範囲が評価値ごとに異なり，たとえ同じ値であってもそれぞれの特徴量ごとに価値が異なる．そのため，それぞれの評価関数から得られた特徴量をそのまま統合すると，平均的に高い値をとる特徴量の影響が大きくなり，四つの特徴量間の格差が生じる．そこで，本研究ではそれぞれの特徴量ごとに偏差値を計算し，偏差値を評価関数を用いて統合することにより，特徴量間の格差を解消した．

**評価関数**　本研究の最終的な目標は，四つの特徴量をどのように統合すれば，利用者が必要とする情報を得られるかを調べることである．ある情報から得られた特徴量に対して，利用者がどれほど注目しているか，どれほど新鮮な情報かを一意に決定することは難しい．このようなあいまいな値を統合するために，ファジィ集合演算を用いる方法が提案されている [6]．ファジィ集合理論の概念は，0 から 1 の範囲の値に適応することができるため，特徴量の統合などの方法としてファジィ集合演算を用いることができる．

そこで，本研究では，特徴量の統合方法として，文献 6) で示されている統合方法のうち，**表 1** および **表 2**，**表 3** に示した計 29 種類の統合方法を用いて特徴量を統合し，実験の結果どの評価関数が最も高い適合率となるかを考察した．これらの評価関数は，ファジィ集合演算の中で，二項演算の AND と OR の性質を満たす代表的なものである．T-演算子 では 10 種類の統合方法について，それぞれ AND と OR による統合方法があり，計 20 種類の評価関数がある．本研究では，T-演算子 に基づいた統合方法として，はじめに参照頻度による注目度と，時間的注目度を統合し，統合注目度を計算する．同様にして，類似する部分文書数による新鮮度と，時間距離による新鮮度から統合新鮮度を求める．さらに統合注目度と統合新鮮度を統合し，必要度とする．このようにして，T-演算子からは 20 種類の必要度を計算することができる．平

**表 1** T-演算子.
Table 1 T-operators.

| | AND | OR | |
|---|---|---|---|
| $T_1$ | $\mathrm{MIN}(x,y)$ | $\mathrm{MAX}(x,y)$ | |
| $T_2$ | $x \cdot y$ | $x+y-xy$ | |
| $T_3$ | $\mathrm{MAX}(x+y-1,0)$ | $\mathrm{MIN}(x+y,1)$ | |
| $T_4$ | $\frac{xy}{x+y-xy}$ | $\frac{x+y-2xy}{1-xy}$ | |
| $T_5$ | $\begin{cases} x & \text{if } y=1 \\ y & \text{if } x=1 \\ 0 & \text{otherwise} \end{cases}$ | $\begin{cases} x & \text{if } y=0 \\ y & \text{if } x=0 \\ 1 & \text{otherwise} \end{cases}$ | |
| $T_6$ | $\frac{\lambda xy}{1-(1-\lambda)(x+y-xy)}$ | $\frac{\lambda(x+y)+xy(1-2\lambda)}{\lambda+xy(1-\lambda)}$ | $0 \le \lambda \le \infty$ |
| $T_7$ | $\mathrm{MAX}(1-((1-x)^p+(1-y)^p)^{1/p},0)$ | $\mathrm{MIN}((x^p+y^p)^{1/p},1)$ | $1 \le p \le \infty$ |
| $T_8$ | $\frac{1}{1+\left((\frac{1}{x}-1)^\lambda+(\frac{1}{y}-1)^\lambda\right)^{1/\lambda}}$ | $\frac{1}{1+\left((\frac{1}{x}-1)^{1-\lambda}+(\frac{1}{y}-1)^{1-\lambda}\right)^{-1/\lambda}}$ | $0 \le \lambda \le \infty$ |
| $T_9$ | $\frac{xy}{\mathrm{MAX}(x,y,\lambda)}$ | $1-\frac{(1-x)(1-y)}{\mathrm{MAX}(1-x,1-y,\lambda)}$ | $0 \le \lambda \le 1$ |
| $T_{10}$ | $\mathrm{MAX}((1+\lambda)(x+y-1)^\lambda xy,0)$ | $\mathrm{MIN}(x+y+\lambda xy,1)$ | $-1 \le \lambda \le \infty$ |

**表 2** 平均演算子
Table 2 Averaging Operators.

| | | |
|---|---|---|
| $A_1$ | $(1-(1-w_1)\cdot(1-w_2)\cdot(1-w_3)\cdot(1-w_4))^\gamma \cdot (w_1\cdot w_2\cdot w_3\cdot w_4)^{1-\gamma}$ | $0 \le \gamma < 1$ |
| $A_2$ | $\gamma\cdot\mathrm{MAX}(w_1\cdot w_2\cdot w_3\cdot w_4)+(1-\gamma)\cdot\mathrm{MIN}(w_1\cdot w_2\cdot w_3\cdot w_4)$ | $0 \le \gamma \le 1$ |
| $A_3$ | $\gamma\cdot(1-(1-w_1)\cdot(1-w_2)\cdot(1-w_3)\cdot(1-w_4))+(1-\gamma)\cdot(w_1\cdot w_2\cdot w_3\cdot w_4)$ | $0 \le \gamma \le 1$ |
| $A_{4.\mathrm{AND}}$ | $\gamma\cdot\mathrm{MIN}(w_1\cdot w_2\cdot w_3\cdot w_4)+\frac{(1-\gamma)(w_1+w_2+w_3+w_4)}{4}$ | $0 \le \gamma \le 1$ |
| $A_{4.\mathrm{OR}}$ | $\gamma\cdot\mathrm{MAX}(w_1\cdot w_2\cdot w_3\cdot w_4)+\frac{(1-\gamma)(w_1+w_2+w_3+w_4)}{4}$ | $0 \le \gamma \le 1$ |

**表 3** 情報検索の演算子
Table 3 The Operators from Information Retrieval.

| | | | |
|---|---|---|---|
| Paice | $P_{\mathrm{AND}}$ | $\frac{\sum_{i=1}^{n}(r^{i-1}\cdot w_i)}{\sum_{i=1}^{n} r^{i-1}}$, $w_i$'s are considered in descending order | $0 \le \gamma \le 1$ |
| | $P_{\mathrm{OR}}$ | $\frac{\sum_{i=1}^{n}(r^{i-1}\cdot w_i)}{\sum_{i=1}^{n} r^{i-1}}$, $w_i$'s are considered in descending order | $0 \le \gamma \le 1$ |
| $p$-norm | $PN_{\mathrm{AND}}$ | $1-\left(\frac{(1-w_1)^p+\cdots+(1-w_n)^p}{n}\right)^{1/p}$ | $0 \le p \le \infty$ |
| | $PN_{\mathrm{OR}}$ | $\left(\frac{w_1^p+\cdots+w_n^p}{n}\right)^{1/p}$ | $0 \le p \le \infty$ |

均演算子および情報検索の演算子では，$w_1$ を参照頻度による注目度，$w_2$ を時間的注目度，$w_3$ を類似する部分文書数による新鮮度，$w_4$ を時間距離による新鮮度の値として計算する．$A_4$ および $p$-norm，Paice に関しては AND と OR による評価値を求め，その結果計 9 種類の評価値が得られる．これらを評価値と T-演算子から求められる評価値とをあわせると最終的に，29 種類の評価値が得られることになる．
また，これらの評価関数の $\gamma$ や $\lambda$，$p$ などの値は，文献 6) で最もよい結果になると言われている値を用いた．これらの値を**表 4** に示す．

**表 4** 評価関数のパラメータの値
Table 4 Parameter of Evaluation Function.

| | | | |
|---|---|---|---|
| $T_{6AND}$ | 1.5 | $A_1$ | 0.5 |
| $T_{6OR}$ | 1.5 | $A_2$ | 0.4 |
| $T_{7AND}$ | 13.0 | $A_3$ | 0.1 |
| $T_{7OR}$ | 13.0 | $A_{4AND}$ | 0.1 |
| $T_{8AND}$ | 0.8 | $A_{4OR}$ | 0.1 |
| $T_{8OR}$ | 0.8 | $P_{AND}$ | 1.0 |
| $T_{9AND}$ | 1.0 | $P_{OR}$ | 1.0 |
| $T_{9OR}$ | 1.0 | $PN_{AND}$ | 2.0 |
| $T_{10AND}$ | -1.0 | $PN_{OR}$ | 2.0 |
| $T_{10OR}$ | -1.0 | | |

## 4. 実験と考察

### 4.1 実験概要

部分集合から，利用者が必要とする情報を抽出するための特徴量の統合方法として最も良い統合方法を求めるために，実験を行った．

本研究では，利用者が取得する情報として，価格.com[7] で提供されている CSV 形式のデータを XML 化したもの用いた．価格.com で提供されているデータは，パーソナルコンピュータやその部品など

表 5　要約した再現率・適合率による評価関数のランキング
Table 5　Ranking of the Evaluation Function by the Summarized Recall and Precision.

| | $n$-point averaged precision | | | | non-interporated average precision | | | |
|---|---|---|---|---|---|---|---|---|
| Ranking | E measure | | Original | | E measure | | Original | |
| 1 | $T_{8OR}$ | 0.629991 | $PN_{AND}$ | 53.500973 | $T_{8OR}$ | 0.580903 | $PN_{AND}$ | 52.689312 |
| 2 | $T_{1OR}$ | 0.631950 | $T_{9AND}$ | 52.793239 | $T_{1OR}$ | 0.581295 | $T_{4AND}$ | 51.732621 |
| 3 | $PN_{OR}$ | 0.636029 | $T_{6AND}$ | 51.547882 | $PN_{OR}$ | 0.585278 | $T_{9AND}$ | 51.579590 |
| 4 | $P_{AND}$ | 0.640351 | $A_1$ | 51.431315 | $T_{6OR}$ | 0.594945 | $T_{6AND}$ | 50.753521 |
| 5 | $P_{OR}$ | 0.640351 | $T_{8OR}$ | 51.007348 | $T_{10OR}$ | 0.596075 | $PN_{OR}$ | 50.495505 |
| 6 | $A_{4AND}$ | 0.642185 | $T_{10AND}$ | 50.877594 | $T_{2OR}$ | 0.596160 | $T_{2AND}$ | 50.369051 |
| 7 | $T_{10OR}$ | 0.642689 | $PN_{OR}$ | 50.705950 | $T_{4AND}$ | 0.596644 | $T_{10AND}$ | 50.368855 |
| 8 | $T_{2OR}$ | 0.642756 | $A_3$ | 50.553270 | $T_{4OR}$ | 0.596908 | $T_{8OR}$ | 50.357137 |
| 9 | $T_{6OR}$ | 0.643026 | $P_{OR}$ | 50.424117 | $A_1$ | 0.597946 | $A_3$ | 50.077641 |
| 10 | $A_{4OR}$ | 0.644354 | $P_{AND}$ | 50.424117 | $A_{4OR}$ | 0.598089 | $A_1$ | 49.747148 |

についての情報であり，それぞれの商品名や仕様，値段などの値が記述されており， 33 のカテゴリの情報が提供されている．実験では， 価格.com から提供されている 34 日分のデータを収集し，それぞれカテゴリごとに異なったスキーマを作成して XML データに変換した．この段階で文脈ノードの定義に従い，XML データを部分文書単位で保存した．部分文書の総数は 765,907 個でその中から 1,512 個の文書文書を実験に使用した．

実験は次のような手順で行った．

- 利用者の情報参照履歴を人手で作成し，それに基づいて部分文書を蓄積済みデータとして保存する．
- 参照履歴に基づいて，部分文書の注目度と新鮮度の評価値を計算する．
- 得られた評価値を 29 種類の評価関数によって統合し，それぞれの統合方法における部分文書のランキングを行う．
- 参照履歴から人手で正解集合を作成し， 29 種類の評価関数によってランキングされた結果との再現率・適合率を求める．
- 求めた再現率・適合率の結果から，どの評価関数が最も良い結果となるかを調べる．

利用者の情報参照履歴は，$1,512$ 個の部分文書を，のべ $10,000$ 回参照した履歴を作成した．参照履歴中には，部分文書 ID と，部分文書が作成された時間のリストが記述されており，これらのリストは部分文書を参照した順番に並んでいる．また，正解集合は $1,512$ 個の部分文書の中から 226 個の部分文書を選んだ．

これらの参照履歴から，$1,512$ 個の部分文書について参照頻度による注目度と，時間的注目度，類似する部分文書の数による新鮮度，時間距離による新鮮度の四つの評価値を計算し，29 種類の評価関数を用いて統合して必要度を求めた後，部分文書を必要度に基づいてランキングした．そして，最後にランキングされた結果と，正解集合から，それぞれの統合方法による再現率・適合率を計算した．

さらに，上記の実験に加えて，参照履歴数と部分文書数および正解集合数は上記と同じ条件で，参照履歴中の部分文書の参照回数と順序だけを変更した同様の実験を二種類行い，三種類の結果から得られた再現率・適合率の平均を求めた．

### 4.2　実験結果と考察

実験によって，それぞれの評価関数における再現率・適合率が得られたが，類似した結果を示した評価関数が多く存在したため，どの評価関数が優れているかを明確に判別することが困難であった．そこで，結果をより明確にするため，文献 8) で示されている方法を用いて，再現率・適合率の要約を行い，最終的な結果を求めた．

文献 8) で示されている再現率・適合率の要約方法に，E 尺度と呼ばれるものがある．E 尺度は，一組の再現率と適合率の値の対から得られ，再現率を $R$，適合率を $P$ とすると，次の式で計算できる．

$$E = 1 - \frac{1}{\frac{1}{2P} + \frac{1}{2R}}$$

つまり，E 尺度の値が低いほど良い結果であると言える．

また，再現率・適合率を要約するための方法として，文献 8) で紹介されている次の二つの方法を用いた．

- $n$ 点平均精度 ($n$-point averaged precision)
- 非補間平均精度 (non-interporated average precision)

$n$ 点平均精度とは，あらかじめ決められた再現率の点について適合率を平均する方法で，非補間平均精度はランキングされた部分文書に対して，正解集合の部分文書が出現した点における適合率の平均を求める．本研究では，$n$ 点平均精度における $n$ の値を 11 とした．

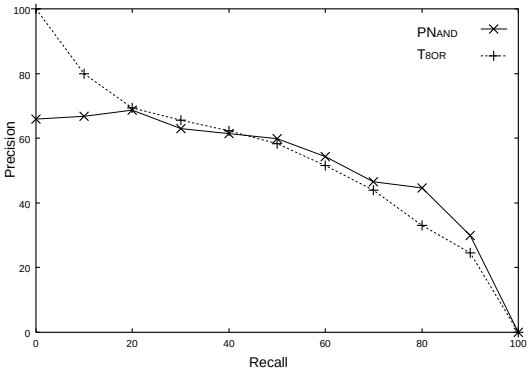


**図 3** 再現率-適合率グラフ
Fig. 3 Recall-Precision Graph.

再現率・適合率の要約は，上記の二つの方法と，E 尺度を用いなかった場合と用いた場合を組み合わせて，次の四つの方法で行った．

- $n$ 点平均精度
- 非補間平均精度
- E 尺度による $n$ 点平均精度
- E 尺度による非補間平均精度

これらの四つの方法によって求めた再現率・適合率の要約結果を用いて，各評価関数をランキングした結果の上位 10 位を，**表 5** に示す．

表 5 の結果からわかるように，E 尺度を用いた場合は $T_{8OR}$ が最も良い結果となり，E 尺度を用いない場合は，$PN_{AND}$ が最も良い結果となっている．また，$T_{8OR}$ と $PN_{AND}$ の再現率・適合率グラフは**図 3** のようになった．

このような結果となる理由は，E 尺度がとる値の特徴として，再現率，適合率のどちらか一方の値が低ければ低いほど，もう一方の値に関係なく高い値（悪い結果）を示すことになるからである．つまり，再現率の平均が同じような二つの結果を比較したとき，同じ適合率の値でも，再現率が高い点での適合率のほうが高い評価値となると言える．

上記の結果から，必要な部分文書を大量に取り出す場合には，$T_{8OR}$ による統合方法が適しているが，より必要度の高い少量の部分文書を取り出す場合には，$PN_{AND}$ による統合方法が適していると言える．

## 5. おわりに

本稿では， XML データで提供されている情報から，利用者が必要とする情報を部分文書ごとに抽出するために，部分文書ごとに注目度と新鮮度を求め，情報に対する利用者の必要度として統合する方法を提案した．また，注目度と新鮮度の統合方法として 29 種類の評価関数を用い，それらの評価関数のうちどの統合方法が最適であるかを実験によって示した．

本研究で示した特徴量とその最適な統合によって，利用者が蓄積した膨大な XML データの中から自分が必要としているものを容易に取り出すことが可能である．また， XML データを文脈ノードに基づいた部分文書単位で扱うため，一つの XML データの中から利用者が必要とする部分だけを取り出すことができる．今後は， XML で記述された商品のカタログなどの情報から，利用者が必要とする情報だけを自動的に取り出すことができるようなアプリケーションの実装を行う予定である．

## 謝　辞

本研究の一部は，文部科学省科学研究費基盤研究（課題番号: 11480088，12680417，12780309），ならびに科学技術振興事業団（JST）の戦略的基礎研究推進事業（CREST）「高度メディア社会の生活情報技術」プログラムの支援によるものである．ここに記して謝意を表す．

## 参 考 文 献

1) Hatano, K., Kinutani, H., Yoshikawa, M. and Uemura, S.: Extraction of Partial XML Documents Using IR-based Structure and Contents Analysis., *In Proc. of Data Semantics in Web Information Systems (DASWIS 2001)*, pp. 189 – 202 (2001). (to appear).
2) 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 XHTML 文書の検索手法, 情報処理学会第 62 回全国大会第 3 分冊, 情報処理学会, pp. 156 – 162 (2001).
3) Ceri, S., Fraternali, P. and Paraboschi, S.: Data-Driven, One-To-One Web Site Generation for Data-Intensive Applications, *Proc, VLDB'99*, pp. 615 – 626 (1999).
4) Kramer, J., Noronha, S. and Vergo, J.: A User-Centered Design Approach to Personalization, *Communication of the ACM*, Vol. 43, pp. 45 – 48 (2000).
5) 宮崎慎也, 馬強, 田中克己: WebSCAN:Web サイトの変更発見と放送型変更通知, 情報処理学会論文誌 : データベース, Vol. 42, No. SIG 8 (TOD 10), pp. 96–107 (2001).
6) Lee, J. H.: Analyzing the Effectiveness of Extended Boolean Models in Information Retrieval, Technical Report TR95-1501 (1995).
7) 価格.com: . http://www.kakaku.com.
8) 徳永健伸: 情報検索と言語処理, 東京大学出版会 (1999).