データベースシステム 118－6
情 報 学 基 礎 54－6
（1999．5．17）

Webグラフにおける意味的情報単位に基づく状況依存リンク

清光 英成 †
† 神戸大学経済学部
kiyomitu@econ.kobe-u.ac.jp

田中 克己 ‡
‡ 神戸大学大学院自然科学研究科
tanaka@db.cs.kobe-u.ac.jp

本論は，ハイパーテキスト文書の意味的なまとまりを製作者自身が定義し，製作者の意図どおりに利用者を閲覧させることを目的とする．そのため，製作者が予め意味的なまとまりとして意味的情報単位を定義でき，製作者の意図どおりに利用者を閲覧させることができるリンク機構を提案する．このアプローチは利用者が膨大な経路候補を情報検索システムから得て，その梗概から次の情報資源を選択するだけではなく，製作者側から経路候補を提供することで，発信される情報の意味を利用者に誤解され難くすることを目的としている．

Situation Depend Link Based on Semantic Information Unit on the Web Graph

Hidenari KIYOMITSU † and Katsumi TANAKA ‡
† Faculty of Economics Kobe University
‡ Department of Computer and System Engineering Kobe University

In this paper, we suggest a concept of the *Semantic Information Unit.* It is an idea to show the users into the web site as the hypertext documents editor hopes. A semantic information unit is a collection of documents and be put on some meenings by editor. We understand it as a path and the link structuer should be changed based on user navigation dynamically. Therefore we propose the situation depend link structure to realize the semantic information unit. Then hypertext document editors can get the power of expression to define the path of user navigation as semantic information unit explicitly.

1 はじめに

爆発的に増加するインターネット資源を構成しているハイパーテキスト文書は章・節等の文書の論理構造を持たないが，キーワードや抄録などの二次情報をもとにデータベースシステムを用いた情報検索技術や文書内走査によるアクセス技術の研究が様々な形態で行われている．

WWW(World Wide Web)は膨大なハイパーテキスト文書をリンクによって関連づけた大規模なグラフである．ハイパーテキスト文書は複数のメディアで構成されたマルチメディア情報であり，1次元の時間線上に展開できるものだけではなく，関係の有無にかかわらず多映像を描画することも可能である．また一般に，WWW 上に公開される HTML(HyperText Markup Language) 文書群はリンクによって結ばれているが，各 HTML 文書が意味的に独立していても連続していても断続的に関連してもよく，文書間の関連づけを柔軟に表現できる．

ハイパーテキスト文書は柔軟なリンク機構により情報提供を容易に実現できる一方，以下のような問題点を持つ．

- 巡行経路によって利用者が把握する意味が異なり，製作者の意図どおりに理解されない恐れがある．
- ハイパーテキスト検索の多くは，利用者が入力するキーワードから個々のハイパーテキスト文書を候補として返すことを目的としているため，複数の文書で構成された意味的な情報単位を検索するには不十分．
- 複数のハイパーテキスト文書で構成された論理的なまとまりを検索する興味深い研究が行われているが，製作者の意図を反映するアプローチではない [1][2].
- 外部からのリンクや URL による直接アクセスにより製作者が用意した閲覧順序に反した閲覧が可能で，製作者の意図どおりに巡行させる方法がない．

これらにより，ハイパーテキスト文書は製作者の意図どおりに閲覧されることが確実ではない．

本研究は，ハイパーテキスト製作者がハイパーテキスト文書の閲覧順序を指定でき，提供するハイパーテキスト文書群の一連の意味を誤解されることなく利用者に理解される枠組を提供するために，

- ハイパーテキスト文書の意味的なまとまりを，製作者が予め定義できる意味的情報単位の概念を導入
- 利用者の巡行経路に応じて，動的に航行可能なリンクを有効にする状況依存リンクを提案
- 意味的情報単位の semantics を到達可能経路間の関連から等価性を導き形式化

している．

本論は，製作者が提供するハイパーテキスト文書の意味的なまとまりを製作者自身が定義し，製作者の意図どおりに利用者を巡行させることを目的とする．複数のハイパーテキスト文書から構成された，一連の情報を検索する研究が行われているが，それらは，WWW 上から意味的なまとまりを発見することを目的としている [3]．製作者の提供したい意図には焦点をあてておらず，既存のデータから情報単位を発掘するというアプローチである．

一方，本論は製作者が予め意味的なまとまりとして意味的情報単位を定義でき，状況依存リンクを用いて製作者の意図どおりに利用者が閲覧できる機構を提案している．それは，情報資源であるハイパーテキスト文書内にリンクを動的に生成するのではなく、利用者の巡行経路をもとに既存のリンクを隠蔽或は有効にする動的リンク機構である．

製作者が発信する情報パッケージとしてのハイパーテキスト文書群とそのリンク構造を製作者の意図に基づいた経路として維持することで，ハイパテキスト検索技術を製作者側から支援する情報資源構成の一つの解決法が明らかになる．これにより、製作者側から経路補完サービスが実現でき，利用者が検索に費やした労力を有効に反映するような情報提供が可能になる．このアプローチは利用者が膨大な経路候補を情報検索システムから得て，その梗概から次の情報資源を選択するだけではなく，製作者側から経路候補を提供するものである．

2 状況依存リンク

本論ではハイパーテキストコンテンツ作者が，ある意味をページ集合とその間のリンクによって表現しようとしている単位を意味的情報単位と呼ぶ。ハイパテキストコンテンツ作者は A から閲覧を開始し B, C を経て D を閲覧することに何らかの意味を与えている場合，AB, BC, CD なるハイパリンクを用意する。しかしながら、現行の WWW 環境では意味的情報単位に対して閲覧を始めるコンテンツを選択する手段は、他のサイトからのリンクや URL による直接アクセスなどによって実現されているので B から 閲覧を開始して C を経て D を閲覧することができ，利用者に作者の意図通りの閲覧をさせることが確実ではない。これはハイパテキスト文書は閲覧する順序をリンクの巡行によって実現しているが，閲覧を開始する文書が適切でなければ，伝達したい情報が欠如したり，冗長になってしまう問題の一因となっている．そこで、意味的情報単位に対する閲覧を，利用者の巡行経路に基づいて到達可能経路を提供する基本概念を考察し，コンテンツ作者の意図を利用者の巡行経路に反映させる状況依存リンク機構を提案する．

2.1 意味的情報単位

コンテンツ作者が A, B, C の順に閲覧を期待すれば図 1 の (i) のような AB, BCのリンクを定義する．また，B または C を A の閲覧の後に閲覧させたい場合には図 1 の (ii) のように AB, ACのリンクを定義する．

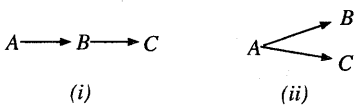

図 1: 既存のリンク

図 $1(i)$ では閲覧順序が決まっているが，図 $1(ii)$ では B が先に閲覧されても C が先に閲覧されてもよい．しかしながら，B, C の閲覧は A の閲覧の後でなければならない．意味的情報単位はこのような

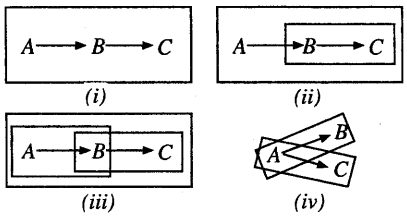

図 2: 意味的情報単位

性質を表現できる経路である．

意味的情報単位は利用者の巡行経路をもとにして，次に閲覧すべきコンテンツを示す状況依存リンクによって結ばれる．

図 $2(i)$ は A から閲覧を開始すれば B へのリンクが有効になり，その後 B を閲覧すれば C へのリンクが有効になるような意味的情報単位を矩形で示している．一方，B または C から閲覧を開始すれば，製作者は A から閲覧してほしいので状況依存リンクを用いて既存のリンクを隠蔽する．図 $2(ii)$ は A から閲覧を開始しても B から閲覧を開始してもよいが，C の閲覧は B の閲覧の後であることを示している．図 $2(iii)$ は三つの意味的情報単位から構成されているが，閲覧が可能な経路は図 $2(ii)$ の例と等しい．

図 $1(ii)$ の例では A から B, C に分岐するようなグラフ構造をもち，AB の巡行と AC の巡行は意味が異なる．したがって，A から閲覧を開始すれば B または C への既存のリンクが有効になる図 $2(iv)$ のような意味的情報単位となる．例えば，図 $2(i)$ は意味的情報単位 $p = ABC$，図 $2(ii)$ は二つの意味的情報単位 $q_1 = ABC$, $q_2 = BC$ と定義される．

図 3 では製作者がコンテンツ集合 $U = \{u_1, u_2, u_3, u_4\}$ を，他の製作者がコンテンツ集合 $V = \{v_1, v_2, ..., v_8\}$ をそれぞれ提供しリンクを定義している．また，コンテンツ集合 V の製作者は意味的情報単位 p_1, p_2, p_3 を定義しようとしている．

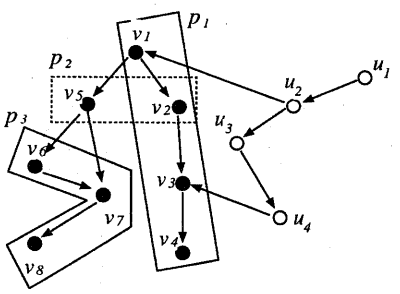

図 3: 意味的情報単位 2

u_4 から v_3 へのリンクは，製作者の意図に反するリンクであるが，q_2 から p_1 へのリンクは意味的情報単 p_1 の先頭コンテンツへのリンクとなっている．コンテンツ集合 P の製作者にとっては意味的情報単位 p_1 を閲覧してもらえることが期待できる優良なリンクであるが，利用者に必ず意味的情報単位 p_1 の全てのコンテンツを閲覧させることを強制するものではない．しかしながら，u_4 から v_3 へのリンクによって v_4 の閲覧を許しては製作者の意図どおりの閲覧とはいえない．

p_1, p_3 は各コンテンツが直接継っているので意味的情報単位として定義できるが，p_2 は直接のリンクがないため意味的情報単位とならない．

2.2 状況依存リンクの例

製作者が一連の意味を表現するために定義した意味的情報単位中のコンテンツに対して，外部から定義されるリンクが，製作者が期待するリンクの先頭コンテンツであるとは限らない．

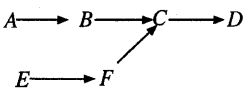

図 4: 外部からのリンクの例

図 4 のように製作者が閲覧してほしい順序にしたがって，AB, BC, CD のリンクを用意しても，外部からのリンク FC が定義でき，C から D への経路も存在してしまう．本来，製作者は $ABCD$ の順にコンテンツが閲覧されることで一連の意味を表現したいのだから，CD のみの閲覧は本位ではない．しかしながら，意味的情報単位外から C へのリンクを制限することは WWW がもつ柔軟性を損ない望ましくない．状況依存リンクは意味的情報単位中の個々のコンテンツへのアクセスを許すが，先頭コンテンツ以外へのアクセスに対して利用者が巡行した経路に基づいて，既存のリンクを有効にしたり隠蔽したりする動的リンク機構である．

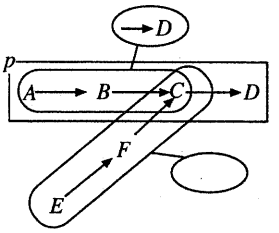

図 5: 状況依存リンクの例 1

図 5 において，利用者が巡行した意味的情報単位 p 上の経路 ABC は長円で表されている．これに対して状況依存リンクは次に閲覧すべきコンテンツ D への既存のリンクを楕円中に表されているように有効にして利用者を D に導く．一方，外部から定義されたリンクを辿って巡行した経路が EFC であれば C から D への既存のリンクを隠蔽する．

3 到達可能経路

コンテンツ作者が，提供する情報に利用者を導く手段は，

1) 検索サイトへ URL と共に梗概を登録

2) メーリングリスト等で概要と URL を周知

3) 広告バーナーによるリンク

等がある．一方，コンテンツ作者以外が，リンク集等により利用者をコンテンツ作者の意図する閲覧順序を無視して導くことも可能である．そこで，コンテンツ作者が利用者に対して閲覧順序を指定する機構が必要となる．

3.1 到達可能経路の意味

意味的情報単位 $p = \{v_1 v_2 ... v_n\}$ において意味的情報単位 p 中のコンテンツ v_i $(1 < i \leq n)$ に到達可能とは v_1 から順に辿り、v_{i-1} に到達していることをいう。このとき $\{v_1, v_1 v_2, ..., v_1 v_2 ... v_n\}$ を到達可能経路と呼び，$Path(p)$ と記述する．また、v_i に既に到達していれば v_i に到達可能であるとする。つまり，到達可能経路は利用者がコンテンツ作者の意図通りに v_i $(1 \leq i \leq n-1)$ に到達すれば v_{i+1} への経路 $v_i v_{i+1}$ が航行可能になるということを意味する．

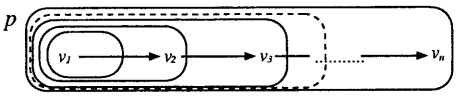

図 6: 到達可能経路

逆に，利用者の巡行経路に到達可能経路の部分集合が存在しなければ，状況依存リンクが既存のリンクを隠蔽し，製作者が定義した閲覧順序を維持する．

3.2 到達可能経路間の関係

意味的情報単位 $p_1 = \{v_1 v_2 ... v_n\}$ が他の意味的情報単位を部品として持つ場合，例えば他の意味的情報単位を $p_2 = \{v_k v_{k+1} ... v_{k+m}\}$ $(1 \leq k \leq n, m \leq n-k)$ とすると，p_2 の到達可能経路は $Path(p_2) = \{v_k, v_{k+1}, ..., v_{k+m}, v_k v_{k+1}, ..., v_k v_{k+1} ... v_{k_m}\}$ である．意味的情報単位 p_1 と p_2 の到達可能経路 $Path(p_1), Path(p_2)$ の間には $k = 1$ または $k = n$ において，

$$Path(p_2) \subset Path(p_1)$$

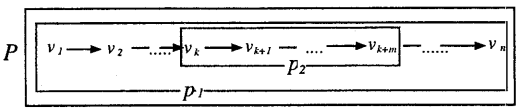

図 7: 二つの意味的情報単位

の関係があるが，$k \neq 1, k \neq n$ ではこの関係が成立しない．しかしながら，p_1 は p_2 上の経路であり，意味的情報単位 p_1 上の全ての意味的情報単位を P とすると，$p_1 \in P, p_2 \in P$ が成り立つ．ある意味的情報単位 p_1 中に他の意味的情報単位 p_2, p_3, p_4 が図??のように連結がなく含まれる場合，p_1 の到達可能経路上の全ての意味的情報単位集合 P の到達可能経路は

$$Path(p_1) \cup Path(p_2) \cup Path(p_3) \cup Path(p_4)$$

であり，

$$Path(P) = \cup_{p_i \in P} Path(p_i)$$

と記述することができる．

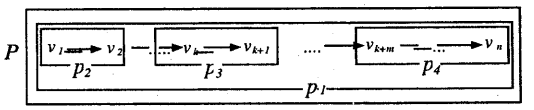

図 8: 複数の意味的情報単位

一方，二つの意味的情報単位において図 9 のような p_1 に p_2 の先頭コンテンツが含まれる場合，p_1 は p_2 と連結しているという．つまり，利用者は p_1 上を巡行している間に p_2 の先頭コンテンツを閲覧するので，p_2 上の経路を巡行できる．

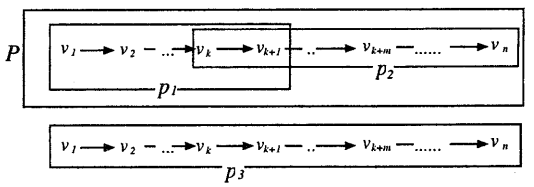

図 9: 意味的情報単位の連結

このとき，意味的情報単位集合 P の到達可能経路は $Path(p_1)$ と $Path(p_2)$ の和だけでなく，p_i と p_j 連結して得られる $v_1 v_2 ... v_k ... v_m$ という経路の到達可能経路も含んでいる．p_1, p_2 を連結してできる到達可能経路を $Path((p_1, p_2)^+)$ とかく．本来，記号 R^+ は閉包であるが，本論では連結によって導出される到達可能経路という意味で用いている．した

がって，意味的情報単位集合 P の到達可能経路は

$$Path(P) = Path(p_1) \cup Path(p_2) \cup Path((p_1, p2)^+)$$

と記述できる．

さらに一般化するために，意味的情報単位集合 P 中の意味的情報単位 $p_i, p_j, ..., p_m$ が連結しているとは，各意味的情報単位の先頭コンテンツを $p_{i_1}, p_{j_1}, ..., p_{m_1}$ としたとき，

$$\begin{aligned} & p_{j_1} \in p_i \wedge p_{j_1} \neq p_{i_1} \\ & \quad \wedge p_{(j+1)_1} \in p_j \wedge p_{(j+1)_1} \neq p_{j_1} \\ & \quad \wedge \ldots\ldots\ldots \\ & \quad \wedge p_{m_1} \in p_{m-1} \wedge p_{m_1} \neq p_{(m-1)_1} \end{aligned}$$

を満たすことをいう．また，意味的情報単位集合 P 中の全ての連結を P^+ と表す．これにより，意味的情報単位集合 P の到達可能経路は

$$Path(P) = \cup_{p_i \in P} Path(p_i) \cup Path(P^+)$$

と一般化できる．

3.3 等価な意味的情報単位

意味的情報単位は製作者の意図どおりに利用者を巡行させる概念である．すなわち，到達可能経路が等しい二つの意味的情報単位は等価である．

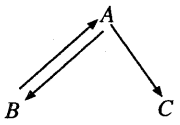

図 10: 複合到達可能経路

ここで図 10 のような反対方向のリンクを持つような意味的情報単位を考える．たとえば，意味的情報単位集合 P_1 が AB, BA という意味的情報単位を持ち，P_2 が ABA, BA という意味的情報単位を持っていたとしよう．すると，到達可能経路 $Path(P_1), Path(P_2)$ はそれぞれ

$Path(P_1)$=

$\{A, B, AB, ABA, BAB, ABAB, BABA, ...\}$

$Path(P_2)$=

$\{A, B, AB, ABA, BAB, ABAB, BABA, ...\}$

となり，

$$Path(P_1) = Path(P_2)$$

である．このとき，P_1 と P_2 は等価であるといい，

$$P_1 \approx P_2$$

と書く．逆に $P_1 \approx P_2$ であれば到達可能経路も等しく，必要十分条件

$$P_1 \approx P_2 \leftrightarrow Path(P_1) = Path(P_2)$$

である．

これにより，複雑に連結する多数の意味的情報単位を整理でき，$Path(P^+)$ の計算量を減少させることができる．

4 議論

本論では，意味的情報単位を経路として論じているが，より一般的にグラフとして扱う場合の問題点と実装手法について議論する．

4.1 グラフとしての意味的情報単位

WWW は巨大な有効グラフとして捉えることができるので，製作者が意味的情報単位をグラフとして定義したい要求は自然なものである．意味的情報単位を導入したのは，製作者の意図どおりに利用者を閲覧させるためであるから，意味的情報単位を経路から有効グラフに進化させなければならない．前述の議論までで述べたのは一筆書きにできる意味的情報単位集合であった．

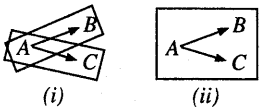

図 11: 経路とグラフ

例えば，図 11 の (i) は二つの経路，(ii) はグラフとして意味的情報単位が定義された例である．利用者が A を閲覧したならば次は B または C のどちらか一方のみを閲覧させたいという要求は確かに存在

する．この場合，図11の(ii)の分岐したリンク間にはXORの関連がある．図11の(i)もそのように捉えることができるように見えるが，製作者はABを巡行することとACを巡行することにそれぞれ意味を与えている．つまり，図11の(i)と(ii)は明らかに意味が異なるのである．

ここでは，意味的情報単位をグラフとして定義した場合，任意個の意味的情報単位の連結がうまく書けるかどうかを例に議論を進める．

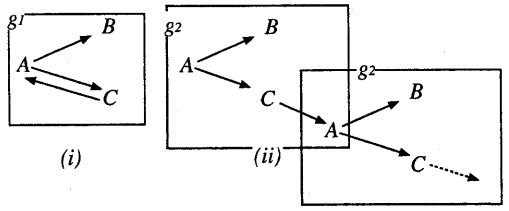

図12: 経路とグラフ

図12の(i)と(ii)は，果して等価なのであろうか．図12の(i)をグラフg_1，(ii)をグラフg_2としよう．g_1もg_2も共にグラフであるので比較可能である．違いは，g_1はCからAに戻る経路が存在するのであり，g_2はAからCを閲覧すればg_2に連結できることである．今，意味的情報単位をグラフとして扱っているのだから，利用者の巡行により到達可能になるのは経路ではなくグラフである．これを到達可能グラフと呼ぶことにしよう．g_1には次に到達可能なグラフがないことがわかる．一方，g_2はg_2を任意個連結しただけ到達可能である．経路としての意味的情報単位では観点が経路であったので単に連結できたが，グラフとしての意味的情報単位では製作者の意味付け次第で構造が変わってしまう．このような製作者の多様な要求を満たすグラフとしての意味的情報単位の形式化は今後の課題である．

4.2 状況依存リンクの実装

利用者の巡行経路から次のコンテンツへのリンクを動的に隠蔽または有効にする状況依存リンクの実装は多くの問題点を持っている．ある意味的情報単位を$p = v_1v_2...v_n$，利用者uの巡行経路を$navigationPath(u)$，次に到達可能なコンテンツへのリンクlを有効にする関数を$visualize(l)$としたとき，

if $navigationPath(u) \in Path(p)$
then $visualize(l)$

というsemanticsを実装できるハイパーテキスト言語が存在すれば実装可能である．

XML(eXtensible Markup Language)の行外リンクやリンクにプログラムを埋め込むことができるDynamicHTMLでは実装可能なように見える．また，HTMLでもフレームセットを用いることで実現できるように見える．しかしながら，状況依存リンクの目的は利用者を製作者の意図どおりの順序で閲覧させることであり，必ず意味的情報単位の最後尾のコンテンツまでの閲覧を強制するものではない．

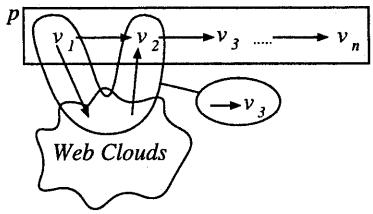

図13: 状況依存リンクの機能

また，状況依存リンクは利用者が巡行すべき意味的情報単位上の順序を規定するだけなので，v_1の次は必ずv_2を閲覧しなければならないのではなく，v_1を閲覧した後，図13のように意味的情報単位外のコンテンツを任意個閲覧してもv_2には到達可能で，状況依存リンクはv_3へのリンクを有効にしなければならないのである．したがって，次のコンテンツへのリンクを有効にするかどうかの判別は，次のコンテンツの記述と利用者のブラウザの機能によって実現されなければならない．

つまり，$navigationPath(u) \in Path(p)$は$navigationPath(u)$から$Path(p)$に含まれる順序列が存在するかどうかを判定するのであるから，$navigationPath(u)$を記憶しながら伝搬させるリンク機構を実現できなければならない．このため，利用者エージェントであるブラウザが利用者の巡行経路を記憶し，記憶された巡行経路を引数として動

的に到達可能経路を処理する機構を記述できるハイパーテキスト言語が必要である.

5 むすび

インターネット検索技術としてハイパーテキスト文書を発見する研究は内外で盛んに行われており,物理的な 1 ページを発見する技術は長足の進歩を遂げている. また, 複数のコンテンツから構成される論理的なまとまりを発見する研究も活発に行われている. それらが提供する情報資源の候補は, 元来ハイパーテキスト文書群への入り口でしかなく, 製作者の意図どおりの閲覧を利用者に促すものではない. 本論はハイパーテキスト文書群内の航路を製作者の作成意図どおりに保つことで発信する情報の意味を誘導し、閲覧によるコンテンツの意味を誤解しにくくする手法を提案した。

＜謝辞＞
本研究を進めるにあたり, 多くの御意見と助言を頂いた本学工学部ならびに, 自然科学研究科田中研究室の皆様に深く感謝致します. また, 研究環境において御支援頂きました, 本学経済学部 Keizo Nagatani 教授, 三谷直紀教授と玉岡助教授, 流通科学大学高橋秀行教授に深く感謝致します.

参考文献

[1] Keishi Tajima, "Querying Composite Object in Semistructured Data", *Proc. of 5th International Conference on Foundation of Data Organization*, Nov. 1998.

[2] Kenji Hatano, Ryouichi Sano, Yiwei Dan and Katsumi Tanaka, "An Interactive Classification of Web Documents by Self-Organization Maps and Search Engines", *Proc of DASFAA'99*, Apr. 1999. (To appear)

[3] Wen-Syan Li and Yi-Leh Wu, "Query Relaxation by Structure for Web Document Retrieval with Progressive Processing", *Proc. of Advanced Database Symposium '98*, pp. 19-25, Dec. 1998.