Vol.2012-ICS-165 No.6
2012/1/12

Web ブロック間のリンク構造に基づく閲覧者の観点の構造化システムの試作

佐野博之[†1]　白松俊[†1]
大囿忠親[†1]　新谷虎松[†1]

本研究では閲覧者の観点から構造化された Web の実現を目的としている．Web における既存のリンク構造は Web ページ制作者の観点に基づく Web の構造であると言える．既存の Web コンテンツ間に対して Web 閲覧者がリンクを作成することが可能な機構を実現することによって，Web 上に存在する既存のネットワーク構造とは異なる，新しいリンク構造の実現が期待できる．閲覧者の観点から構造化された Web を用いることで，閲覧者の目的に合った Web コンテンツの情報推薦技術が実現する．本稿では閲覧者の観点を Web ブロック間のリンク構造を用いてモデル化し，Web ページを Web ブロックへと分割するための手法，および閲覧者が Web ブロックに対して観点を記述するためのシステムについて述べる．

Implementing a Structuring System for Viewpoints of Web Users based on Hyperlinks between Web Blocks

Hiroyuki Sano,[†1] Shun Shiramatsu,[†1]
Tadachika Ozono[†1] and Toramatsu Shintani[†1]

Our goal is to create structured Web based on viewpoints of Web users. The hyperlink network on the Web is a structure based on viewpoints of Web creators. A system that allows Web users to add hyperlinks between existing Web contents will generate a new network, which differs from the existing one. The structured Web based on viewpoints of Web users enables to build a new recommendation system which serves user's purpose. In this paper, viewpoints of Web users' are represented by hyperlinks between Web blocks. We propose the method to segment a Web page into Web blocks and the system which enables users to add hyperlinks between Web blocks that have some relations based on their viewpoints.

図 1　Web ページ中に含まれる複数の Web コンテンツ
Fig. 1　Multiple Web contents in a Web page

1. はじめに

本研究では閲覧者の観点から構造化された Web の実現を目的としている．Web における既存のリンク構造は Web ページ制作者の観点に基づく Web の構造であると言える．既存の Web コンテンツ間に対して Web 閲覧者がリンクを作成することが可能な機構を実現することによって，Web 上に存在する既存のネットワーク構造とは異なる，Web 閲覧者同士の新しいリンク構造の実現が期待できる．閲覧者の観点から構造化された Web を用いることで，閲覧者の目的に合った Web コンテンツの情報推薦技術が実現する．

1 つの Web ページ中には複数の Web コンテンツが含まれている．**図 1** は Yahoo!ニュースのスクリーンショットである．この Web ページの中にはニュース記事本文の他に，サイトロゴや広告，サイトメニュー，関連記事などの複数の Web コンテンツが含まれている．我々は文献[1] において，Web ページの閲覧者が Web コンテンツを特定するための付箋アノテーションシステムを提案した．付箋アノテーションシステムによって閲覧者が任意のコンテンツに対してアノテーションを行い，アノテーション間に対してリンクを作成することを可能とした．本リンクは閲覧者の観点を表現していると考えられるが，付箋アノテーショ

†1 名古屋工業大学 大学院工学研究科 情報工学専攻
Dept. of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology

© 2012 Information Processing Society of Japan

ンの貼り付け対象となった Web ページの DOM 構造によっては，Web コンテンツの特定精度に問題があった．

本稿では Web ページを Web ブロックと呼ばれる意味的にまとまりのある単位へと分割し，Web ブロック間に対して閲覧者が自由に自身の観点を記述することを提案する．以降，第 2 章では閲覧者の観点を定義し，そのモデル化，および閲覧者の観点を収集しるためのシステムについて述べる．第 3 章では Web ページを Web ブロックへと分割するための手法について述べる．第 4 章にて考察を行い，最後に第 5 章にて本稿をまとめる．

2. 閲覧者の観点

閲覧者の立場から見た Web ブロックの役割や Web ブロック間の関連性を，本研究では閲覧者の観点として扱う．Web ブロックとは，Web ページ中に存在する意味的にまとまりのある単位のことである．Web 情報の構造化を試みた研究は多数存在するが，構造化に関して閲覧者の観点に着目した研究は少ない．

Web ページの制作者の目的は，自分の所有する Web サイトの訪問数を増加させる，コンバージョン率を増加させる，Web 広告のクリック数を増加させることが主に挙げられる．制作した Web ページが検索エンジンの検索結果において上位にランクされることは上記の目的を満たすための必要条件であり，Web ページ制作者は検索エンジン最適化 (SEO) を行う．SEO に影響を与える要素としてリンクが挙げられる．検索エンジンでは HITS[2] のようなリンクを用いたスコアリングアルゴリズムによって Web ページのスコアリングをすることが多い．Web ページの制作者は自身の Web ページのスコアの向上を意識したリンクを作成する傾向にある．

Web ページの閲覧者は制作者と異なるリンク構造を作成すると考えられる．第 2.1 節では Web ブロック間のリンク構造を用いた閲覧者の観点のモデル化について述べる．第 2.2 節では閲覧者の観点を収集し Web を構造化するためのシステムについて述べる．

2.1 提案モデル

閲覧者の観点をグラフ構造を用いてモデル化する．Web 閲覧者が Web ブロック間に対してリンクを作成することによって，閲覧者の観点に基づいた Web ブロック間の関連性を表現する．たとえば，ニュース記事 A とブログ記事 B の間に関連があると考えた場合には，ニュース記事 A からブログ記事 B へリンクを作成する．その結果，Web ブロックをノードとし，リンクをエッジとした有向グラフが形成される．形成されるグラフ構造を，本研究では閲覧者の観点として扱う．

グラフの各ノードに相当する Web ブロックに対して，それぞれの Web ブロックの役割を示すためのラベルを付与する．文献[3] では Web ページに含まれる要素をオブジェクトとして捉え，そのオブジェクトが担う役割を 6 種類に分類している．その 6 種類とは，Information オブジェクト，Navigation オブジェクト，Interaction オブジェクト，Decoration オブジェクト，Special オブジェクト，Page オブジェクトである．本研究では Web ブロックに対して役割ラベルを与えるが，文献[3] に倣い，Information ブロック (INF)，Navigation ブロック (NAV)，Interaction ブロック (INT)，Special ブロック (SP) の 4 種類を用意した．INF は，閲覧者に対して情報を伝達することを目的としたブロックに対して付与されるラベルである．例としてニュース記事やブログ記事などが挙げられる．NAV は，閲覧者を他の Web ページへと導くことを目的としたブロックに対して付与されるラベルである．例としてニュースのヘッドラインやサイトメニューが挙げられる．INT は，閲覧者が Web ページに対してアクションを起こすためのブロックに対して付与されるラベルである．例として検索フォームや，印刷・ブックマークするための JavaScript を実行するためのボタンが挙げられる．SP は，広告やサイトロゴ，コピーライトなどを記述したブロックに対して付与されるラベルである．

以上のことを形式的に述べる．閲覧者 u の観点が形成するグラフ G_u を $G_u = (B_u, R_u, \delta_0, \delta_1)$ と定義する．B_u は閲覧者 u によって関連性が表記された Web ブロックの集合であり，Web ブロックと役割ラベルのタプルで表現する．$B_u = \{(b_1, l_1), (b_2, l_2), \cdots, (b_n, l_2)\}$ である．R_u は閲覧者 u が Web ブロック B_u 間に対して記述した関連性であり，$R_u = \{r_{u1}, r_{u2}, \cdots, r_{um} | m \leq {}_nP_2\}$ である．$\delta_0 : WB_u \rightarrow R_u$，$\delta_1 : WB_u \rightarrow R_u$ は始点関数，終点関数である．これらはエッジからノードへの写像である．$\delta_0(r_{u1}) = b_i$，$\delta_1(r_{u1}) = b_j$ であった場合には，閲覧者 u が Web ブロック b_i から Web ブロック b_j に対して関連性があると記述したことを意味する．すなわち，r_{u1} は，閲覧者 u によって Web ブロック b_i から Web ブロック b_j に対して作成されたリンクを意味する．

図 2 に，閲覧者の観点の例を示す．この図の中では，3 つの Web ページ P_A, P_B, P_C を対象としている．それぞれの Web ページの Web ブロック構成は，$P_A = \{b_{A1}, b_{A2}\}, P_B = \{b_{B1}, b_{B2}, b_{B3}\}, P_C = \{b_{C1}, b_{C2}, b_{C2}\}$ である．閲覧者 u がこれらの Web ページ中に含まれる 4 つの Web ブロックに着目して図 2 のような関連性を記述した場合には，$B_u = \{(b_{A2}, INF), (b_{B1}, NAV), (b_{B3}, INF), (b_{C3}, INF)\}$，$R_u = \{r_{u1}, r_{u2}, r_{u3}, r_{u4}\}$ となる．写像 δ_0, δ_1 に関しては，**表 1** に示す通りである．

本研究では Decoration ブロック，Page ブロックの 2 種類の役割ラベルを用意しなかった．

© 2012 Information Processing Society of Japan

Vol.2012-ICS-165 No.6
2012/1/12

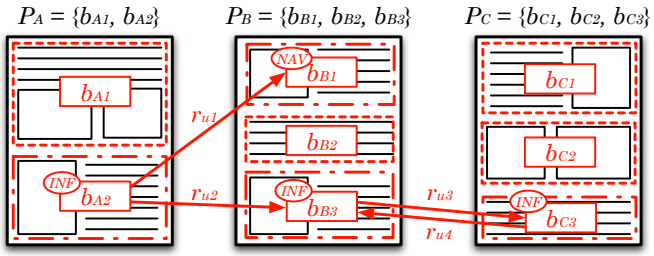

表 1 写像
Table 1 Map

r	δ_0	δ_1
r_1	b_{A2}	b_{B1}
r_2	b_{A2}	b_{B3}
r_3	b_{B3}	b_{C3}
r_4	b_{C3}	b_{B3}

図 2 閲覧者の観点の例
Fig. 2 Example of user's viewpoints

図 3 付箋アノテーションシステム
Fig. 3 Screenshot of the Annotation Stickies

その理由について述べる．Jinlin らは Decoration オブジェクトを，Web ページの一部分を装飾するためのパーツとして定義している．Decoration オブジェクトの例として，<hr>タグによる水平線や，箇条書きを行うときに各項目の頭につける記号などが存在する．これらのオブジェクト単体が，意味的にまとまりのある単位として抽出されることはないと予想される．また，Jinlin らが定義した Page オブジェクトとは Web ページ全体のことを意味している．上記の理由により，Decoration および Page は分割結果のブロックの役割としては相応しくないため，本研究ではこれらの役割ラベルを用意しないこととした．

2.2 観点収集システム

文献[1] では，閲覧者が Web ページ中のコンテンツを一意に特定することが可能な付箋アノテーションシステムを実現している．ユーザは**図 3** の左のように，コンテンツに対して

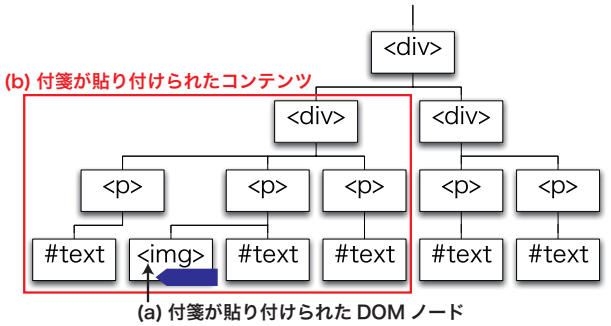

図 4 DOM ツリーに基づいて付箋が貼り付けられた Web コンテンツを特定
Fig. 4 Extract Web content based on DOM tree.

付箋を貼り付けることができる．提案システムでは付箋アノテーション間にリンクを作成する機能を実現しており，ユーザは本システムを用いることによって任意の Web コンテンツ間にリンクを作成可能となっている．すなわち，Web ページ製作者の観点とは別に，閲覧者が自身の観点に基づいたリンク構造を生成することが可能である．

付箋アノテーションシステムの実装は DOM ノードに対して付箋を貼り付ける仕様となっており，貼り付け対象となった DOM ノードの親の親のノードを，貼り付け対象となった Web コンテンツとして扱う．例を**図 4** に示す．図 4 では，(a) に示した画像に対して付箋が貼り付けられたとする．その際，(b) で示した が貼り付け対象となった Web コンテンツとして扱われる．付箋アノテーションシステムでは，上記に示した単純なヒューリスティクスによって貼り付け対象となった Web コンテンツを特定している．貼り付け対象となった Web ページの DOM 構造によっては Web コンテンツの特定精度に問題があった．

付箋アノテーションシステムにおいて，計算機が Web ページを自動で分割し，閲覧者にとって意味がある単位で貼り付け対象となった Web コンテンツを特定できることが好ましい．次章では閲覧者から観点を取得する前処理として，Web ページを Web ブロック単位へと分割する手法について述べる．

3. 観点取得のためのブロック分割

Web ページ分割に関して様々な手法が提案されている[4)–6)]．高精度な Web ページ分割により検索エンジンの精度向上[7)] など多くの利点が指摘されており，研究の余地がある．本研究での Web ページ分割には，文献[8)] で提案したタイトルブロックを用いた分割手法を用

© 2012 Information Processing Society of Japan

図 5 Web ページから抽出された細分化ブロックの例
Fig. 5 Example of twelve minimum-blocks extracted from a Web page

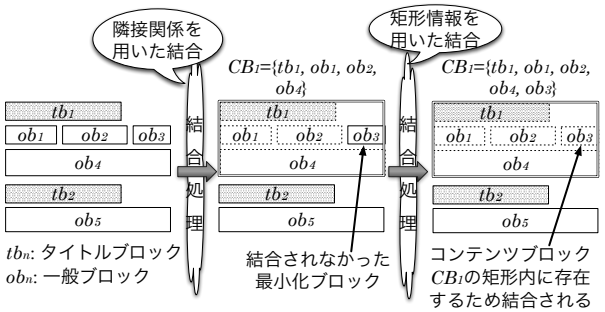

図 6 結合ステップ
Fig. 6 Assemble Steps

いる．

3.1 細分化ブロックへの分割

提案手法では Web ページ分割の第一ステップとして，Web ページを細分化ブロックという非常に細かい単位へと分割する．細分化ブロックとは “子ノードとしてブロックレベル要素を持たないブロックレベル要素” のことである．ただしインライン要素であっても，細分化ブロックの兄弟ノードである場合には，そのインライン要素も 1 つの細分化ブロックとして抽出する．これにより，Web ページ上にレンダリングされる全ての要素がいずれかの細分化ブロックに属することとなる．

図 5 は Google で “名古屋工業大学 新谷研究室” と検索した結果の Web ページから，検索結果の上位 3 件の部分を切り取ったスクリーンショットである．この図の中には実線で囲った 12 個の細分化ブロックが存在する．

3.2 タイトルブロックの抽出

各細分化ブロックに含まれているテキスト情報に着目して細分化ブロックの結合を行い，Web ページを意味的にまとまりのあるコンテンツブロック単位へと分割する．結合の際にはタイトルブロックに着目した結合を行う．タイトルブロックとは，細分化ブロックの中でも特に，直下の Web コンテンツの見出しとなる細分化ブロックのことである．Web コンテンツが多数配置されている Web ページには，人が閲覧したときに読解しやすいよう，Web コンテンツの上部にタイトルブロックが配置されていることが多いことが挙げられる．すなわち，タイトルブロックは複数の Web コンテンツ間の仕切りとして利用することが可能であると言える．

図 5 の中には 3 つのタイトルブロックが存在する．“新谷研究室” というテキストを持つ細分化ブロック，“新谷研への道順” というテキストを持つ細分化ブロック，“名古屋工業大学 情報工学科 新谷研究室所属 小山充智のサイト” というテキストを持つ細分化ブロックの 3 つである．これら 3 つの細分化ブロックは直下に存在するブロックのタイトルを表しているため，タイトルブロックとみなすことができる．また，図 5 のスクリーンショットは，これらのタイトルブロックを区切りとして意味的に 3 つに分割できる．我々は予備実験を行い，大半の Web ページに含まれる Web コンテンツが，タイトルブロックとそれに続く本文・画像から構成されていることを確認した．

決定木学習によって作成した判別器を用いて，細分化ブロックの中からタイトルブロックを抽出する．

3.3 タイトルブロックを用いた結合

細分化ブロックの隣接関係と矩形情報を用いて細分化ブロックの結合処理を行う．具体例を**図 6** に示す．図 6 の左のように，タイトルブロックが 2 個 (tb_1, tb_2)，一般ブロックが 5 個 $(ob_1, ob_2, \cdots, ob_5)$ の，合計 7 個の細分化ブロックが存在する場合を考える．タイトルブロック tb_1 に着目した場合，まずは tb_1 がコンテナへと格納される．すなわち，$Container = \{tb_1\}$ である．tb_1 の下に隣接する細分化ブロックは，ob_1 と ob_2 である．これらは 2 個とも一般ブロックであるため，結合処理を進める．ob_1 と ob_2 がコンテナへと格納され，$Container = \{tb_1, ob_1, ob_2\}$ となる．ob_1 の下に隣接する細分化ブロックは，ob_4 である．ob_4 は一般ブロックであるため，結合処理を進める．ob_4 がコンテナへと格納され，$Container = \{tb_1, ob_1, ob_2, ob_4\}$ となる．ob_4 の下に隣接する細分化ブロックは，tb_2 であ

© 2012 Information Processing Society of Japan

Vol.2012-ICS-165 No.6
2012/1/12

る．tb_2 はタイトルブロックであるため，隣接関係を用いた結合処理を終了する．この時点で tb_1, ob_1, ob_2, ob_4 の 4 つの細分化ブロックの結合を完了した（図 6 中央参照）．ob_3 は tb_1 と隣接していないため，上記の処理を行っただけでは ob_3 は結合されないという問題が発生する．この問題を解決するために，次のステップとして，結合された細分化ブロックの集合が形成する矩形内部に位置する細分化ブロックも，同一のコンテンツブロックへ結合する．図 6 では，コンテンツブロック CB_1 の矩形内に ob_3 が存在する．したがって ob_3 も CB_1 に結合した後に，結合処理を終了する（図 6 右参照）．

図 6 では簡略化のためにコンテンツブロックが 1 段組の図を示したが，本アルゴリズムでは，Web ページが 1 段組であることを仮定しない．コンテンツブロックが複数の段組でレイアウトされた Web ページには，タイトルブロックも複数の段組で存在する．それぞれのタイトルブロックに対して直下に存在する一般ブロックを結合していくという処理を繰り返し行っていくため，全てのタイトルブロックに対して処理を完了した時には，Web ページが複数の段組へと分割される．

3.4 分割結果

本分割手法の適用例を**図 7** に 4 つ示す．赤い矩形が分割結果を表している．矩形が存在しないところは，システムによって分割されなかったところである．図 7 の (a) は Yahoo! ニュース[⋆1]のニュース記事，(b) は Ameba ブログ[⋆2]のブログ記事，(c) は Amazon.co.jp のトップページ[⋆3]，(d) は新谷研究室のトップページ[⋆4]である．(a), (b) は分割が成功していると判断された結果，(c) はほぼ成功していると判断された結果，(d) は失敗と判断された結果である．

提案手法では，Web ページ中のタイトルブロックが存在しないところではコンテンツブロックへの結合処理が行われないため，Web ページ中に分割されない領域が発生する．(d) では，タイトルブロックが 2 つしか抽出されなかったため，コンテンツブロックも 2 つしか生成されず，Web ページのほぼ大半が分割されなかった．(d) の Web ページ右上には，画像を用いて表現されているタイトルブロックが 4 つ存在するが，本研究で生成した分類器ではこれらのタイトルブロックを一般ブロックと誤判定した．これは決定木学習で利用した訓練データの中に，画像を用いて表現されているタイトルブロックがあまり含まれていなかったことが原因である．訓練データを見直し，画像を用いてタイトルを表現しているタイトルブロックの抽出精度を上げることによって (d) のような Web ページの分割精度を上げることが可能である．

⋆1 http://headlines.yahoo.co.jp/hl
⋆2 http://ameblo.jp/
⋆3 http://www.amazon.co.jp/
⋆4 http://www-toralab.ics.nitech.ac.jp/index-j.html

図 7 提案手法による Web ページの分割結果
Fig. 7 Segmentation result by proposed method

4. 考　　察

提案モデルでは閲覧者の観点をグラフ構造として表現する．ユーザによって作成されたグラフ構造はユーザの好みや意見を反映していると考えられる．すなわち，ユーザプロファイルとして利用することも可能であり，情報推薦への応用が考えられる．

グラフの類似性を調査することで，人気のあるコンテンツの発見が容易になることが期待

© 2012 Information Processing Society of Japan

できる．グラフの類似性を比較する手法として，グラフ編集距離を用いる手法[9] や，最大共通部分グラフを用いる手法[10] がある．WWW のユーザは情報収集をする際に検索エンジンを利用する．ユーザは検索エンジンの検索結果の上位にランクされた Web ページを積極的に開き，下位にランクされた Web ページを開くことは少ない．検索結果のランキングは Web ページ制作者の観点によって作成されたリンク構造によってスコアリングされたものである．たとえ検索結果が下位にランクされる Web ページであっても，本稿で提案した閲覧者の観点から再評価を行うことにより，その Web ページは価値のあるものとなる可能性を秘めている．閲覧者の観点を多数の閲覧者間で共有することによって，検索エンジンに頼らない情報収集を行うことが可能になると考えられる．

また，類似する Web ページから類似コンテンツを抽出するための Web ラッパー構築を支援することも期待できる．Web ラッパーとは Web ページ中から特定のコンテンツを抽出するために作成されるプログラムのことである．Web ラッパーの開発に関しては DOM 構造に依存した記述をすることが多い．Web ラッパーの欠点として HTML 構造の変化に脆弱な点が挙げられる．したがって対象となる Web ページの DOM 構造が多少でも変化するたびに Web ラッパーを再構築する必要がある．Web ラッパーを自動で構築する研究も盛んに行われているが，適用できる Web ページに制限が多く，まだ精度も不十分である[11]．

5. おわりに

本稿では Web ページを Web ブロックと呼ばれる意味的にまとまりのある単位へと分割し，Web ブロック間に対して閲覧者が自由に自身の観点を記述することを提案した．提案モデルでは閲覧者の観点をグラフ構造として表現する．複数の閲覧者から収集した観点を用いてグラフ間の類似度や和集合グラフを考慮することによって，情報推薦への応用が可能になると考えられる．本研究の今後の課題として，本稿で提案した閲覧者の観点を収集し，有効性の検証を行っていく予定である．

謝辞 本研究の一部は科研費（22500128），および総務省 戦略的情報通信研究開発推進制度 (SCOPE) の支援を受けた．

参考文献

1) 佐野博之，浅見昌平，大囿忠親，新谷虎松: Web エージェントを用いた Web コンテンツへの付箋アノテーションシステムの実現, コンピュータソフトウェア, Vol.26, No.3, pp.69–77.
2) Kleinberg, J: Authoritative sources in a hyperlinked environment, in Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 604-632 (1998)
3) Jinlin, C., Baoyao, Z., Jin, S., Hongjiang, Z., and Qiu, F.: Function-Based Object Model Towards Website Adaptation, in Proceedings of the 10th international conference companion on World wide web, WWW ’01, pp. 587—596, New York, NY, USA (2001), ACM
4) Cai, D., Yu, S., Wen, J.-R., and Ma, W.-Y.: Extracting con- tent structure for web pages based on visual representation, in Proceedings of the 5th Asia-Pacific web conference on Web technolo- gies and applications, APWeb ’03, pp. 406—417, Berlin, Heidelberg (2003), Springer-Verlag
5) Cao, J., Mao, B., and Luo, J.: A segmentation method for web page analysis using shrinking and dividing, Int. J. Parallel Emerg. Distrib. Syst., Vol. 25, pp. 93—104 (2010)
6) Fernandes, D., Moura, de E. S., Silva, da A. S., Ribeiro-Neto, B., and Braga, E.: A site oriented method for segment- ing web pages, in Proceedings of the 34th international ACM SIGIR conference on Research and development in Information, SIGIR ’11, pp. 215—224, New York, NY, USA (2011), ACM
7) Vadrevu, S. and Velipasaoglu, E.: Identifying primary content from web pages and its application to web search ranking, in Proceedings of the 20th international conference companion on World wide web, WWW ’11, pp. 135—136, New York, NY, USA (2011), ACM
8) Hiroyuki Sano, Shun Shiramatsu, Tadachika Ozono and Toramatsu Shintani: A Web Page Segmentation Method based on Page Layouts and Title Blocks, Vol.11, No.10, pp.84–90 (2011)
9) Bunke, H.: On a relation between graph edit distance and maximum common subgraph, Pattern Recogn. Lett., Vol.18, No.9, pp.689-694 (1997)
10) John W. Raymond and Peter Willett: Maximum common subgraph isomorphism algorithms for the matching of chemical structures Journal of Computer-Aided Molecular Design, Vol.16, pp.521-533 (2002)
11) 山田泰寛，池田大輔，坂本比呂志，有村博紀: WWW からの情報抽出 : Web ラッパーの自動構築 (< 特集 >WWW 上の情報の知的アクセスのためのテキスト処理), 人工知能学会誌, Vol.19, No.3, pp.302–310.

© 2012 Information Processing Society of Japan