様式 C-19

科学研究費補助金研究成果報告書

平成　24 年　9 月　12 日現在

機関番号：12612
研究種目：基盤研究（C）
研究期間：2009　～　2011
課題番号：21500096
研究課題名（和文）多角的な Web 空間マイニングを実現するデータベース処理機構の研究

研究課題名（英文）A Study on Database Systems for Multi-Dimensional Web-Structure Mining

研究代表者
大森　匡　（OHMORI TADASHI）
電気通信大学・大学院情報システム学研究科・教授
研究者番号：30233274

研究成果の概要（和文）：　本研究では、Web データ集合上で入力としてユーザが自分の興味を表す多様な制約条件を与えた時に、その制約を満たす Web コミュニティの構造情報を効率的に計算するデータベースシステムを実現した。本システムは、Web ページのリンク構造を表すレコードについて、その始点ノードに制約を与える場合（FROM 型制約）と終点ノードに制約を与える場合（TO 型制約）の 2 種類の制約を考え、時間軸も入れて合計 3 次元の制約に応じた多様なコミュニティ構造を計算し、重要度でランクづけして出力する。ユーザが定義した任意の制約条件に対応したコミュニティ構造を効率的に計算できることも示した。

研究成果の概要（英文）：Web-Community Mining is a significant issue in today's cyberspace technology. One problem is how to find outstanding communities in a big data space, and the other is how to support user-given personalization in the web-community mining function. This research is aimed at solving the latter issue. The proposed solution is a new database system which provides a data-cube query model on a target web-space dataset. Namely, under a given multi-dimensional constraint, the system computes web-community structures under the constraint and returns a ranked list of communities. The proposed data-cube model has three dimensional constraints of FROM-type, TO-type, and the time of a data snapshot. Efficient set-oriented data operations on the data-cube model and their algorithms are also proposed.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	700,000	210,000	910,000
2010 年度	700,000	210,000	910,000
2011 年度	500,000	150,000	650,000
年度			
年度			
総　計	1,900,000	570,000	2,470,000

研究分野：総合領域
科研費の分科・細目：情報学　メディア情報学・データベース
キーワード：Web マイニング，データベースシステム，コミュニティ発見，問い合わせ処理

１．研究開始当初の背景

近年の Web 情報社会においては、対象とする Web データ集合においてどのようなコミュニティ構造が存在するのかを分析する研究は重要な課題である。とりわけ 2000 年代当初に大規模クローリングと完全 2 部グラフ構造（いわゆるコア）によるコミュニティ計算法が IBM から提案されて以来、コア検

出に基づいた Web コミュニティ構造分析は重要なテーマである。2009 年時点では、このような Web コミュニティ構造マイニングの研究では、巨大 Web データ集合上でサイト間リンクだけを使って完全 2 部グラフ（コア）構造を計算列挙するか、hub-authority スコアに代表されるスコア計算法しかなかった。これらの研究は、大規模な Web 空間全体を広く俯瞰して目立つコミュニティを発見することには適していた。しかし、一方で、ユーザが知りたいコミュニティ構造をパーソナライズしてより細かく調べるという能力は考慮されていなかった。そこで、本研究では、分析したい大学などの仮想組織の全 Web データを対象にして、入力としてユーザが自分の興味を表す多様な制約条件を与え、その制約を満たす Web コミュニティの構造情報を効率的に計算し、どのようなコミュニティが当該制約の下で重要かを計算するデータベースシステムを実現することとした。位置づけとしては、本システムは、研究開始時点では、Web コミュニティマイニングの研究において、ユーザが与えた制約条件に応じた personalization 機能を重視した Web データ上のコミュニティ分析を行うデータベースシステムの研究として位置づけられる。

２．研究の目的

本研究の目的は､Web 空間から集めた Web データの集合に対し、入力としてユーザが自分の興味を表す多様な制約条件を与え、当該制約を満たすような Web コミュニティの構造情報を出力として返すデータベースシステムを実現することである。

本システムは､Web ページのリンク構造を表すレコードについて、その始点ノードに制約を与える場合（FROM 型制約）と終点ノードに制約を与える場合（TO 型制約）の 2 種類の制約を考え、時間軸も入れて合計 3 次元の制約の組み合わせを考え、これによって多様なコミュニティ構造への問い合わせを記述する。

用いる制約条件は、最初は、大学内の学科別サブドメインに応じたものを考えた。例えば、大学ドメインを情報系、電気系、その他（機械・センター系）の分野別サブドメインに分けたとき、

- 「情報系ドメインから見て重要なコミュニティ構造を求めよ」（FROM 型制約）
- 「大学全体から見て電気系の中で重要なコミュニティを求めよ」(TO 型制約)
- 「情報系とその他系ドメイン間の相互関係において重要なコミュニティを求めよ」（FROM 型と TO 型制約の AND）

などの制約問い合わせが考えられる

具体的には、本システムは、利用者の意図に応じて、FROM 型・TO 型制約と時間軸によって 3 次元データキューブモデルを考え、このモデルに沿ったデータ集合演算体系を持たせることになる。

以上の枠組みは研究代表者らが 2006 年から提案し、2008 年度末にはデータキューブモデルやコアコミュニティグラフと呼ぶコミュニティ構造を表すグラフをコアから求める技法、当該グラフ上のランク計算法などの基礎を固めていた。これを受けて、2009 年度から本基盤研究で主に明らかにすべき点は以下の 3 点であった：

(1) 当該 3 次元制約に応じたデータキューブモデルに沿ったデータ集合演算と高効率な実行アルゴリズムを提案すること。すなわち、制約に対応したコアを計算した後にコミュニティを表すノード間の関係をグラフ構造で表すという方法は 2008 年度末には確立していたが、分析に用いるリンクの詳細度や制約のアドホックさに伴って、効率的な計算体系の実現は全く自明ではなかった。これを解決できなければ、高効率な問い合わせ処理は実現不能であった。

(2) あらかじめどの範囲までのデータキューブモデルを計算して用意しておくべきかを決めること。上の（1）と相補する課題であった。

(3) ユーザが定義した任意の FROM 型・TO 型制約に応じて Web コミュニティ構造を差分的に計算できるようにすること。すなわち、研究開始当初から、学科別ドメインでは対応できていたが､「キーワード K にヒットするページから前方リンク 5 ホップ以内にいる Web ページを対象とする」などの一般的な FROM 型制約や TO 型制約述語に対処して効率的に計算できる必要があった。

この他に、計算された Web コミュニティ構造を使った有用なコミュニティ情報の検出が実際に得られるのか、どのような Web コミュニティ分析用の問い合わせが本システムで実現できるのか、を示す必要があった。システム全体としては、与えられた問い合わせを処理する適当なデータ集合演算の処理木を構築して効率良く各演算を実行できるデータベース演算体系を確立することが目標であった。

３．研究の方法

ここでは、上述した研究目的を実現するために本研究で考案したシステムの諸概念を説明する。詳細な実現内容は４．で述べる。

[FROM 型･TO 型制約による制約下のコア計算と意義]

本研究では、Web ページをノード、リンクをエッジとおいてグラフ構造で Web データを表す。ノード i1, i2, …, in からノード v へリンクがあるとき、((i1,i2,…, in),v)を 1 つのリンクレコードと呼び、リンクレコード集合から Pruning 処理を経て apriori 法でコア計算を行う。以下、簡単のため、リンクレコードのことを単にレコードと呼ぶ。

図 1 に、FROM 型制約と TO 型制約に応じたコアの説明を示す。例として、制約としては、大学全体 uec.ac.jp ドメインを情報系学科ドメイン(ISCJ. IS と C と J 学科のドメインという意味で、全て情報系学科)と電気系ドメイン(EE)、その他(OTHER)の 3 つのサブドメインに分けた場合で考える。（図 1 の A,B,C が各サブドメインに対応している)。

この時、「FROM(A)制約を満たすコア」とは、A 学科のドメインを少なくとも 1 つ始点に持つようなリンクレコードの集合から計算されたコアのことである。(A は分野別学科を表す変数で、上の例では ISCJ か EE か OTHER)。このコアは、「A 分野から大学全体の Web 空間を見たときに重要なコア」を情報として表すと考えられる。このタイプの制約を FROM 型制約と呼ぶ。

一方、「TO(A)制約を満たすコア」とは、A 学科ドメインのページを終点としたリンクレコードの集合から計算されたコアのことである。（図 1 下)。このコアは、「大学全体から A 分野の Web 空間を見たときに重要なコア」を情報として表している。このタイプの制約を TO 型制約と呼ぶ。

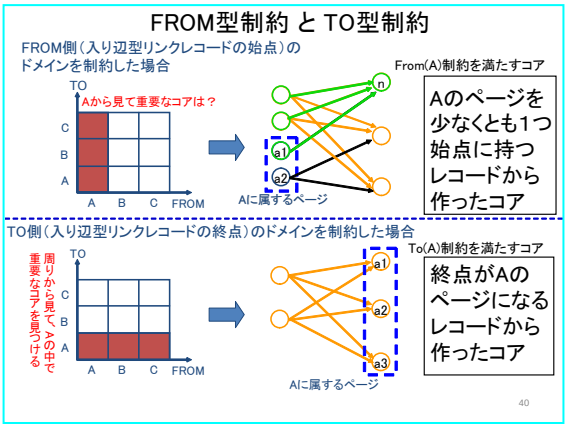

図 1 ：FROM 型制約・TO 型制約

この時、制約として FROM(A or C) AND TO(A or C) を考えると、この制約を満たすコアは、始点として A か C ドメインを（少なくとも 1 つ）持ち、かつ、終点として A か C ドメインを持つようなレコード集合から求まるコアである。従って、当該制約を満たすコアは、「A 分野と C 分野の相互関係において重要なコア」を表している。

本研究では、以上の考えに基づいて、FROM 型･TO 型制約に応じたコア集合からコミュニティ構造を表すグラフを作り、そのグラフ上でスコア計算を行って、当該制約の下での重要なコミュニティを計算する。

[FROM/TO 型制約を満たすコアからコミュニティ構造グラフを作る方法]

FROM(A)制約が与えられた場合を例にして本研究で提案するデータベースシステムの動作を述べる。

まず、FROM(A)制約を満たすコア集合を求め、そのうち強く関連するコア同士を 1 ノードに集約し、このノードが 1 つのコミュニティを表すと考え、ノード間の参照関係をエッジとしてグラフを作る。このグラフが、FROM(A)制約下のコミュニティ構造を表すグラフである。提案システムは、このグラフ上で PageRank 式を改良したスコア計算を行い、スコアの高い順にコミュニティを表すノードを答えとして出力する。本システムは、以上の手順で FROM(A)制約の下で重要なコミュニティを計算する。与えられた制約が FROM(A or C) and TO(A or C)なら、以上の手順によって、分野 A と C の相互関係において重要なコミュニティを計算することができる。本提案システムは、以上の考えに基づいて、多様な制約下でのコミュニティ構造を Web データ空間上で計算するデータベースシステムである。

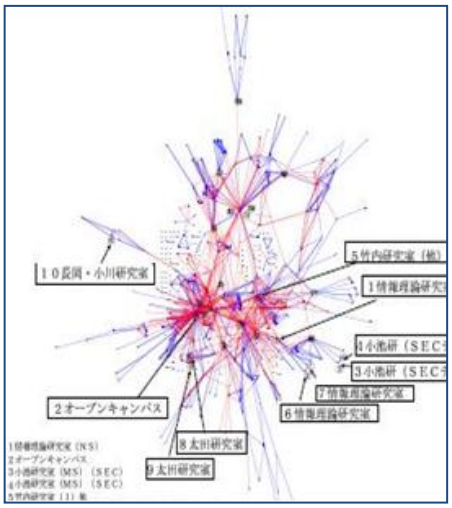

図 2 ：大学ドメイン全体のコミュニティ構造グラフとスコア計算結果

例として、図２は、2005 年度の uec.ac.jp ドメインの Web データ 10 万ページ・80 万リンクについて、サイト内リンクの枝刈り指数 b=8, コアの終点（center のこと。本研究では authority ノードと呼ぶ）の数 s=4 以上として求めた全コアから作ったコミュニティ構造を表すグラフである。(FROM 型・TO 型制約の両方を大学全体(ALL)に設定して求めたもの)。当該グラフ（コミュニティ構造グラフと呼んでいる）のノードは、終点を一定数以上共有するコアを１ノードへ集約したものであり、エッジは、ノードに含まれるコア間に参照リンクが存在するか、または、同一 Web ページが含まれたときに生成される。エッジ重みは同一サイト間：他サイト間で 1:10 に設定し、PageRank 式に似せたスコア計算式を採用している。

これに対して、b や s などのパラメタを同一にして FROM(OTHER)制約下で計算したコミュニティ構造グラフを図３に示す。

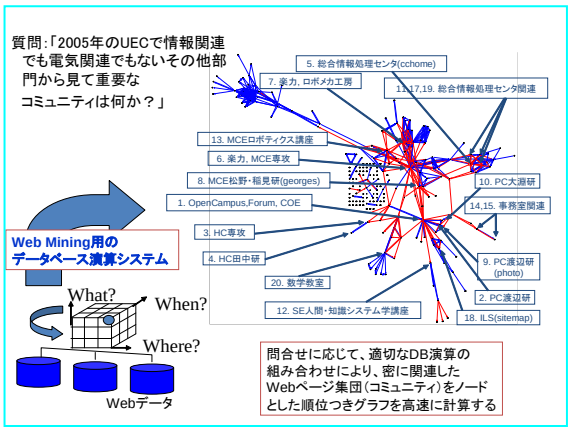

図３：FROM(OTHER)制約下のコミュニティ構造グラフ

この図では、制約なしの時の図２とは異なり、FROM(OTHER)制約の下で初めて順位が高くなるコミュニティが検出されていることを確認している。

このように、ユーザが知りたい FROM 型・TO 型制約の下で本研究が提案するコミュニティ構造グラフを計算すれば、当該制約下で有用なコミュニティ情報を得ることができる。しかし、効率的な計算を行うためには、関係データベース演算のような適切なデータ集合演算体系を考案し、それを使って、与えられた問い合わせを処理演算木で表して、効率的に実行する必要がある。そこで、本研究では、この演算体系と処理アルゴリズムの考案を主眼とし、その効率的実装を通して提案システム全体を実現することとした。

４．研究成果

<u>主な成果と位置づけ</u>

(1) FROM 型・TO 型制約に応じたフィルタリング演算、マージ演算の提案

上記の要求に応えるため、本研究では、既にある制約下で計算されたコア集合から差分的に FROM(A)制約や TO(A)制約（に応じたコア集合）を計算する技法として、フィルタリング演算とマージ演算を提案した。

フィルタリング演算は、コアの始点（ハブノード）と終点（オーソリティノード）が制約 A を満たすか否かで宣言的に計算する手法である。対象となるコア集合のスキャン１回で処理可能である。

一方、FROM(A)を満たすコア集合 a と FROM(B)制約を満たすコア集合 b を入力として FROM(A or B)を満たすコア集合を計算する演算が(FROM 型制約上の)マージ演算である。この演算結果は、a∪b では求まらないコアを含む。そこで、本研究では、a と b を計算するために使ったリンクレコード集合の始点を枝刈りすることで、差分となるコア集合を計算する算法を提案した。

マージ演算は TO(A)制約と TO(B)制約を考えたとき、TO(A or B)制約を満たすコア集合を差分計算するときにも使われる（TO 型制約上のマージ演算）。この計算も、単純な和演算では必要なコアが欠落する。本研究では、対象となるレコード集合を TO(A or B)のコア計算に必要なものだけに Pruning する方法を考案した。その結果、対象レコード集合の定数回スキャンと残ったレコード上のコア計算により、5 秒以内で差分計算できることを示した。

いずれの技法も、毎回完全に再計算をする場合には 100 秒以上かかる状況に対して、差分計算により 10 秒以下で済むことがわかった。そのため、コア計算における差分計算技法として新規性と有用性のある技法である。

(2) あらかじめ計算しておくべきデータキューブ制約の範囲について　及び、
(3) 任意の制約述語への対応について

学科別ドメインの制約下で効率的な差分計算を確立できたため、ユーザが与えた制約条件の階層の下でどこまでの範囲のコア集合を事前に計算しておくべきか、がデザイン上の課題となった。最終的には、サイト内リンクをどこまでコア計算の対象に含めるかを表す詳細度パラメタ b（b が大きいほど多くのサイト内リンクを考慮に入れてコア計

算する)、求めたいコアの終点数ｓ（ｓが小さいほど細かく調べる)、ユーザが与えた制約述語が対象全体（ALL）に対して直和分割か否か、の 3 点に応じて事前に計算するコア集合の範囲を決定することとした。これが項目（2）に応じた成果である。

さらに、上記（3）の成果として、ユーザが調べたい制約条件の和が空間全体の被覆にならない場合でも、上記（1）で考案した各演算処理アルゴリズムが適用可能であることを証明した。ここで、特に意義がある点は、FROM(A or B) and TO(A or B)のような複合問い合わせが、FROM(A)と FROM(B)を各々満たすコア集合から TO 制約上のフィルタリング演算と FROM 制約上のマージ演算の組み合わせで実行可能であることを示したことである。これによって、複合問い合わせであっても、提案したデータ処理演算を組み合わせて関係代数演算のように実行できることを示すことができた。

(4) 有用なコミュニティ情報の検出

以上の結果を受けて、本研究では最終的に学科別ドメイン述語やキーワード述語を使った FROM 型・TO 型制約下のコミュニティ構造計算を行った。例として、次のような分析を行った：

Q1:「IS 学科と M 学科の相互関係で見て重要な上位 20 コミュニティを求めよ。そのうち、IS だけ、または M だけから見て重要な上位 20 と異なるものは何か」

Q2:「キーワード X または Y から見て重要な上位 40 コミュニティのうちロボット関連のコミュニティはどれか」

などである。その結果、分野 A と B の相互関係でみると初めて高順位になるコミュニティなどを効率良く見つけることができた。

本研究では、コミュニティ構造グラフからの情報分析は各ノードのランクに基づいて行い、ランク上位 K 個のノード集合間で Jaccard 類似度による比較演算を用いて分析を行った。このように、コミュニティ分析において必要なデータ集合演算として、指定制約下のコア計算処理（従前のデータキューブ理論における実体化演算）や FROM 型・TO 型のマージ演算、フィルタリング演算、類似度結合演算、が明確になった点はデータベース研究の上で重要な成果といえる。

(5) その他の関連する成果について

本研究でコミュニティ構造グラフの分析を行なう過程で副次的に問題となった点の一つに、類似のコミュニティノードの間にどのようなグラフ構造があるかを調べたいという潜在的な要求があった。本研究では上位ランクのノード集合自体を答えとして扱ったが、ノード間の関連を調べる手段も必要という考えである。そこで、グラフで表現されたデータベースにおけるキーワード入力による部分グラフ列挙技法 MDPBF を開発した。

MDPBF は、グラフデータベースの解として、入力キーワード制約を満たす steiner 木をコストの小さい順に正確に上位 K 個列挙する算法である。本質的に NP-hard の列挙問題であるため、実装では 1 万ノード級のグラフで大きさを限定した解上位 10 個程度を 10 秒程度の計算時間で得るにとどまった。また、出力となるグラフ構造自体の理解の容易さにも問題があり、計算効率・情報品質の双方で実用的とするには課題を残した。

研究成果全体の位置づけと今後の展望

本研究で提案したデータベースシステムは、大規模検索エンジンから分析対象とする Web データを供給してもらって個別ユーザ対応でコミュニティ分析を繰り返すシステムとして位置づけられる。2012 年現在においても、Web コミュニティ分析の研究はできるだけ広範囲の Web 空間を調べて目立つコミュニティを求めることを主としており、本研究のようにデータ空間をユーザの意図に応じて多様な角度から多次元制約問い合わせとして分析するというデータキューブ特有の構想はない。しかし、ビッグデータ分析で見られるように、実際には本システムで提案したようなパーソナリゼーション対応のコミュニティ分析機能は重要である。この点で、本提案システムは独創性の高いものである。

また、Web データに代表されるリンクデータ集合からの情報構造計算において、関係データベースのような集合演算体系を導入して効率的計算システムの構築に成功したことは、本研究成果として大きな意義がある。

一方、本研究で補助的に開発したグラフデータベース探索技法 MDPBF は、アルゴリズム研究としては有意義と言えるが、本データベースシステムの出力となる巨大なコミュニティ構造グラフへの効率的適用では課題を残した。しかし、2012 年現在の巨大データ処理ではグラフ構造によって最終的に全ての情報をモデル化して利用することも多いと考えられるため、そのような状況での適切な利用法が今後の課題である。

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 0 件）

〔学会発表〕（計 7 件）

① M.Wang, L.Jiang, L.Zhang, T.Ohmori: “Exact Top-k Keyword Search on Graph Databases,” ACM Symposium on Applied Computing, Data Mining track, pp.985-986, 2011 March, Tunghai Univ. (Taiwan) http://dl.acm.org/citation.cfm?doid=1982185.1982400

② 王美蓉, 張麗茹, 金銀実, 大森匡: “グラフデータベースにおける冗長解抑制を伴った Top-k キーワード検索アルゴリズムに関する研究,” DEIM Forum 2011, B10-1, 電子情報通信学会(IEICE)等, 2011 年 3 月, ラフォーレ伊豆 (国内) http://db-event.jpn.org/deim2011/proceedings/pdf/b10-1.pdf
（2011 年 7 月公開）

③ 齋藤太陽, 大森匡, 星守 : “多次元的な Web 空間マイニングを行うデータベースシステムの実現 : 制約条件の一般化,” 第 2 回データ工学と情報マネジメントに関するフォーラム DEIM Forum 2010, F5-3, 電子情報通信学会(IEICE)等, 淡路国際会議場（国内）, 2010 年 3 月. http://db-event.jpn.org/deim2010/proceedings/files/F5-3.pdf
（2010 年 5 月公開）

④ 齋藤太陽,大森匡,星守 : “多次元的な Web 空間マイニングを行うデータベースシステムの実現, “ 第 72 回情報処理学会全国大会 3R-8, 東大（国内）, 2010 年 3 月.

⑤ 張洪鋒,大森匡,星守 : “Web 構造分析を目的とした多次元データマイニング機構の効率化 : To 型制約問い合わせの処理方法, ” DEIM Forum 2009, E7-3, ヤマハリゾートつま恋（国内）, IEICE 等, 2009 年 3 月. http://db-event.jpn.org/deim2009/proceedings/files/E7-3.pdf
(2009 年 5 月公開)

６．研究組織

(1)研究代表者

大森　匡（OHMORI Tadashi）

電気通信大学・大学院情報システム学研究科・教授

研究者番号：30233274

(2)研究分担者

（なし）

(3)連携研究者

（なし）

以上