DEWS2003 2-P-02

リンク構造に基づくWebイジング検索モデルの提案

阿部　友一†　　河野　浩之†

† 京都大学 大学院 情報学研究科 システム科学専攻　〒606–8501 京都府京都市左京区吉田本町

あらまし　Web構造マイニングにより求めたページの重要度を情報検索に用いる有効性は高く，PageRankやHITSアルゴリズムをはじめとする多数のアルゴリズムの提案がなされている．また，商用Web検索エンジンでは，この種のアルゴリズムの応用や，情報フィルタリング手法を採用して，より良質の検索結果の提示が試みられている．しかしながら，検索エンジンの性能指標となる検索精度や検索時間などのバランスを確保した上で，どれだけの規模のWebページ群を対象としたWebマイニングを行うかは大きな課題である．特に，多数のWebページの内容やリンク関係を解析するための処理速度が問題になる．そこで，本稿では，Webリンク構造に注目したWeb構造マイニングを効率良く行うために，磁性体を表す基礎的な物理モデルであるイジングモデルを用いて，Web情報をモデル化する．そして，大規模なWebページ群を効率よく検索するリンク構造型イジング検索 (Link structural Ising Retrieval) を行うLIRアルゴリズムを提案し，NTCIR情報検索システム評価用テストコレクション構築プロジェクトで利用されているWebデータを用いて，その有効性を検証する．特に，提案するLIRアルゴリズムに関係する幾つかの熱力学的パラメータの変動が，検索速度や検索精度にどのような影響を与えるかを，不要ページの抑制効果の面から議論する．

キーワード　Webマイニング, イジングモデル, 情報検索, ハイパーリンク構造

Proposal for Web Ising Retrieval Model Based on Link Structure

Tomokazu ABE† and Hiroyuki KAWANO†

† Department of Systems Science, Graduate School of Informatics, Kyoto University

Abstract　Many web structure mining algorithms, such as PageRank, HITS and others, has been developed for discovering web documents with usefulness and high quality. In order to retrieve better search results, several commercial web search engines adopt this kind of algorithms and also use techniques of information filtering. In web mining models, the significant issue is the scalability of web page group, keeping balance between the search time and search accuracy, which indicate the performance of the search engine. The processing speed to analyze a lot of web contents and link relation is also significant problem. In this paper, we employed the Ising model, which is known as a fundamental physical model of magnetism to perform efficient web structure mining based on its link relation. We propose LIR algorithm (Link structural Ising Retrieval) which efficiently reduce the size of informative web pages in a large web community. We evaluated the validity of LIR algorithm using actual web data of the NTCIR (NII-NACSIS Test Collection for IR Systems) Project. We argue the characteristics of search speed and search accuracy by changing several thermodynamic parameters related to the LIR algorithm.

Key words　Web Mining, Ising model, Information Retrival, hyper-link structure

1. はじめに

インターネット上の情報流通量の増大には目覚しいものがあるが，その技術変化も劇的なものであるゆえに，大きな社会現象を起こし続けている．特に，Web システムの成長は著しく，非常に豊富で密度の濃い情報源になりつつある．しかしながら，ポータルサイトで得られる URL を元に，クリックして得られる Web 上のデータは部分的なものであるため，大量の Web ページの内容やリンク関係を計算機により解析し，そこから意味ある有用な情報を抽出する Web マイニングと呼ばれる研究・技術開発が活発化している [5], [18]．なお，情報抽出の際に用いられるデータは，Web ページに記述されたテキストやハイパーリンク，Web サーバ上のログ，さらに，利用者側のブラウジング履歴などが対象となっている．

また，Web マイニングの重要な応用として Web サーチエンジンがあり，数多くの Web ページからユーザーが必要とするページを効率よく発見する技術が開発され続けている．実際，Web の構想が提案された 1989 年以後，着実に成長を続けていたが，1995 年あたりを境に Web ページ数が急増した．その結果，Web 検索エンジンで，すべての Web ページを収集することが困難になり，より性能の良いサーチエンジンを提供するために，良質な Web ページを優先した検索結果提示や，新しい Web ページを中心に収集するなどが試みられた．さらに，サーチエンジンの性能を示す指標として検索対象ページ数があげられるが，2002 年 6 月に AlltheWeb [1] が 21 億ページを超え，Google [8] を上回ったと宣言をした．なお，2003 年現在，Google では 3,083,324,652 ページと提示されており，Web ページ数は単調増加傾向にある．

ところで，このような大規模なページ数を操作する上で多くの問題が存在するが，リンク関係行列を数値計算する場合には主記憶領域の問題がある [15]．例えば，100 万ページのリンク関係行列の各要素を 2byte int 型を利用して構成するための記憶領域は，$2 \cdot 10^6 \cdot 10^6 \approx 2$TB 必要である．従って，例えば 20 億ものページに対しても操作可能な技術開発が必要である．実際，近年の Web ページ数の増加により，Web サーチエンジンはクラスタ構成され，良質の検索結果を提供するために計算コストを削減した実装可能なアルゴリズムが必要となっている．

そこで，本稿では，新たな Web 検索アプローチとして，磁性体の臨界現象を良く表すイジングモデルを Web のリンク構造に対して応用した Web 検索モデルを提案する．そして，提案した Web 検索モデルが不要なページを排除し，どの程度の抑制効果を示すかを検証する．また，実際の Web データを用いてモデル化，検索を行い，磁性体の特性を表す幾つかのパラメータを変化させることで，検索対象領域の大きさを，どの程度小さくすることができるかを示す．

以下，2 章では Web マイニングの概要，3 章では Web モデリングに用いるイジングモデルを説明する．4 章では，実際に提案する Web 検索モデルの提案を行い，5 章では Web データを用いて実験を行った際の環境説明，パラメータ変化による振る舞いの結果を示す．そして 6 章で結論と今後の課題を述べる．

表 1　Web mining の分類

名称	対象データ
Web contents mining	Web documents, picture, $\cdots$
Web structure mining	link structure, graph structure
Web usage mining	access log, proxy log, cash, $\cdots$

2. Web マイニング

Web マイニング研究は，データベース，情報検索，人工知能，機械学習や自然言語処理等の多くの研究と関連する．なお，Web マイニングが対象とする領域のおおよその分類として，Web 上の注目する Web ソースの種類により表 1 のような Web content mining，Web structure mining，Web usage mining の 3 つのカテゴリに分類されている．加えて，検索結果をユーザーが理解する支援となる情報可視化 (information visualization) など様々な側面から Web システムに対するアプローチが行われているので，本章では，Web Mining について簡単に整理する．

2.1 Web コンテンツ・マイニング

Web ページはテキスト，画像，オーディオ，ビデオ，メタデータ，ハイパーリンクといったような幾つかのデータ型から構成される．これらのデータ型を用いて Web 上から情報を抽出する技術を Web content mining と呼ぶ．これらのデータ型の中でもテキスト，ハイパーリンクといったデータ型単一に関する研究は活発に行われているが，画像，オーディオなどのマルチメディアデータに関する研究は未だ始まったばかりである．

我々の研究を含め，テキストベースの情報検索は盛んに行われてきた [12], [14], [20] が，テキストマイニングを行う際に用いられる Web ドキュメントを 3 分類している．すなわち，非構造化ドキュメント，半構造化ドキュメント，構造化ドキュメントである．

- 非構造化ドキュメント

非構造化ドキュメントとはいわゆるテキストの事である．このテキストの中から精度の高い特徴ベクトルを抽出するために，さまざまな研究が行われている．例えば，重要な特徴ベクトルを抽出するために，頻度の低い単語，stop word といったものは取り除いたりする前処理を必要とする．また LSI(Latent Semantic Indexing) により Web ドキュメントの特徴を記述する特徴的な単語を少ない語数で構成し，元のドキュメントの特徴ベクトルを低い次元のベクトルへと変形するアプローチもある．

- 半構造化ドキュメント

Web ページ内のハイパーリンクを用いて，Web ドキュメントから特徴ベクトルの抽出を行うため多くのデータマイニング手法が利用される [24]．例えば，ハイパーテキストによる分類やカテゴリ化，クラスタリング，Web ドキュメント間の関連性学習，パターンや規則の抽出学習，半構造化データ内でのパターン発見，ラッパー抽出など多岐に渡る．

- 構造化ドキュメント

抽出される構造を，関係データベースのテーブルに相当するように，メタデータや辞書などにより制限できる Web ドキュメントを構造化ドキュメントという．規則正しく制限された値域内でしか特徴抽出ができないため，大量に氾濫している多くの Web ドキュメントを対象とした検索には向かない．

2.2 Web 構造マイニング

従来，ページ群検索手法として，テキストの類似性を用いた方法が一般的に用いられてきた．しかし，Web 文章を対象にした場合には，さらにハイパーリンク構造を考慮してドキュメント間の関係を調べる手法が考えられる．すなわち，Web structure mining は，ディレクトリ構造やハイパーリンクのグラフ構造に注目するアプローチである [19]．リンク構造を用いた技術として,「多くの良質なページからリンクされているページは，やはり良質なページである」という考えを用いた Google における PageRank アルゴリズムや HITS アルゴリズム [16]，Web Trawling などは，その代表例である．

例えば，HITS アルゴリズムでは，ユーザーから入力されたトピック周辺のページの集合に対して，情報源として有用なページ（Authority）と，リンク集として有用なページ（Hub）を，Authority ページと Hub ページの相互関係を利用して求める．すなわち，Authority 値と Hub 値を求めることで重要なページを探し出す．この手法が対象とする Web グラフは，あるトピックに関連したページ群であり，そのページ間のリンクのほとんどは意味のあるリンクであると考えている．このアプローチを発展させたものとして，Web グラフの隣接行列の特異値分解を行いさらにコミュニティの特徴的キーワードを抽出する研究がある [9]．これらの技術は，関連する Web ページコミュニティの発見，重要ページのランク付けに有効であることが分かっている．

また，Web Trawling は大規模な Web スナップショットデータから特定の構造を高速に検索するための研究である．Web コミュニティ発見を，ネットワークフロー問題に帰着させ，最大流・最小カット定理に基づいて，内部と外部とのリンクが少なくなるように Web コミュニティを見出す研究等がある [7]．

その他，Web グラフ構造として，巨視的には蝶ネクタイ構造であるとの研究 [3]，微視的には関連ページが 2 部グラフになっているとの指摘 [4] などがある．また，中間構造の候補として，興味を共有する Web ページ集合を用いて，Web コミュニティに特有な構造をモデル化することができれば，より強力なシステム実現も可能とされる．例えば，ディレクトリ構造とリンク構造を組み合わせて，Web 上のドキュメントを取り扱いやすくするグループ化技術 [17] などが役立つと思われる．

2.3 Web 利用マイニング

Web usage mining は Web ページの直接的なデータを用いる以外に，Web サーバ上のログデータなどを用いてマイニングを行う技術である．対象となるデータは広範であるが，一般的には Web 内のクライアント，プロキシサーバー，ログデータ，Cookie，クリックストリームなどが相当する．なお，ユーザープロフィールの学習，ユーザーナビゲーションパターンの学習の 2 種のカテゴリがある．

2.4 情報視覚化

多くの検索エンジンは，検索された URL を列挙するだけで，必要とするページかどうかは，ユーザーがアクセスして判断しなければならない．中には，検索結果の要約出力を行う場合もあるが，多数の検索結果の関係の判断には困難が伴う．そこで，検索結果データの理解を支援するために，情報可視化 (information visulization) が広く応用されている．我々の研究 [13], [21] 以外にも，Web のリンク構造とアンカーテキストをラベルに利用し Web コミュニティチャートをばねモデルを用いたグラフ構造に可視化する Browser [6] など数多くの研究がある．

3. イジングモデル

物質は低温では秩序的な状態を取り，高温では無秩序な状態を取る．例えば，磁性体は低温では磁力を保っているが，ある温度を超えると，磁力を突然失ってしまう臨界現象を起こす．この臨界現象を適切に再現するモデルとしてイジングモデルがある．そこで，Webモデルの新たな可能性を求めて，本稿ではWebのモデル化を行う際にイジングモデルを用いる．なお，本章では，2次元イジングモデルを中心に，イジングモデルに関係する典型的な熱力学量を説明し，Webイジング検索に必要なアルゴリズムを紹介する．

3.1 イジングモデルの性質

2次元イジングモデルは，図1に示すような各格子点にスピンが配置されたモデルである．磁性を持つ原子に対してスピンが与えられ，各スピンはスピン変数 S_i で表現される．ここで，$S_i = 1$ の時にスピン S_i が上向きの磁力を持ち，$S_i = -1$ の時スピン S_i が下向きの磁力を持つとする．

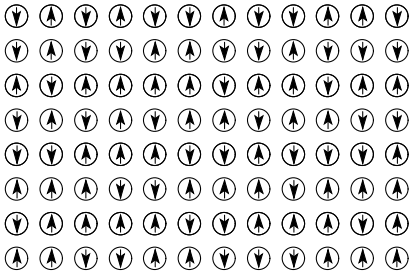

図1 イジングモデル

次に，イジングモデルに用いる幾つかの熱力学量を導入する．スピン S_i の相互作用エネルギー E_i を(1)式で，あるモデル状態 S のハミルトニアン $H(S)$ を(2)式で表す．

$$E_i = -J\sum_j S_iS_j - BS_i \tag{1}$$

$$H(S) = -J\sum_{<i,j>} S_iS_j - B\sum_i S_i \tag{2}$$

J は結合定数，B は外部磁場，S_j はスピン S_i の隣接スピン，$< i, j >$ は隣接する2つのスピンの組み合わせを表すものとする．一般に，$J > 0$ のモデルは強磁性体，$J < 0$ のモデルは反強磁性体である．統計物理学において，このモデルの任意の状態 S の出現確率を，分布関数 Z である(3)式を用いて，(4)式で与える．

$$Z = \sum_S e^{-\frac{1}{k_BT}H(S)} \tag{3}$$

$$\text{ボルツマン分布}\quad \omega(S) = \frac{e^{-\beta H(S)}}{Z} \tag{4}$$

ただし，k_B はボルツマン定数，T は温度である．そして，典型的な熱力学量が次のように与えられる．

$$\text{磁化}\quad M = \sum_S \omega(S)\left(\sum_i S_i\right) \tag{5}$$

$$\text{磁気感受率}\quad \chi = \sum_S \omega(S)\left(\sum_i S_i\right)^2 - M^2 \tag{6}$$

$$\text{エネルギー}\quad E = \sum_S \omega(S)H(S) \tag{7}$$

$$\text{比熱}\quad C_B = \sum_S \omega(S)H^2(S) - E^2 \tag{8}$$

3.2 スピン・フリップ・ダイナミクス

ランダムに選択されたスピン S_i に対して，スピン S_i の周辺スピンと外部磁場と関連のある相互作用エネルギー E_i を最小にするようにスピンの向きを変える操作のことをスピン・フリップ・ダイナミクス(spin flip dynamics)という．なお，スピン・フリップ・ダイナミクスを行う手順は，次の通りである．

【スピン・フリップ・ダイナミクス】

（1） スピン S_i を選択する．

（2） スピン S_i をフリップした時の変動エネルギー ΔE_i を計算する．

（3） $e^{-\beta\Delta E_i}$ に比例した確率に応じてスピン S_i をフリップする． □

ここで，相互作用エネルギー E_i が小さくなるようにスピンをフリップするということは，スピン S_i の周辺に上向きのスピンが多数あれば上向きに，下向きのスピンが多数あれば下向きにフリップすると考えられる．

3.3 イジング探索法 [10]

イジングモデルで用いていた相互作用エネルギーに含まれる外部磁場を

$$B = H_d(m_d(a) - \theta_d) \tag{9}$$

で与える．ここで，H_d は外部磁場との結合定数，$m_d(a)$ はスピン S_a の探索物らしさ，θ_d は探索物であるかどうかの閾値である．この時，イジング探索アルゴリズムは次のようになる．

【イジング探索アルゴリズム】

（1） 全領域のスピンを探索物状態（$S_i = -1$）とし，探索物候補だけを格納したリストを作る．

（2） 探索物候補のリストの中からランダムに1つのスピン S_a を選ぶ．

（3） スピン S_a の探索物らしさ $m_d(a)$ を求める．

（4） スピン S_a の周辺で状態を確定していないスピンに対して spin flip dynamics を数回行う．

（5） 探索物である状態から探索物でない状態にフリップしたスピンをリストから取り除き，逆に探索物でない状態から探索物の状態にフリップしたスピンは探索物候補に加える．

（6） 探索物候補がなくなるまで，2 から 5 の操作を繰り返す探索を進める． □

上述したイジング探索法の特徴は探索の全領域内において探索物らしさが連続的な値を取る事を利用する点である．連続的な値を取る対象物として，画像データ内の顔検出 [10] などが対象となっており，イジング探索法を用いた検出回数の削減や計算速度の向上が研究されている．

4. リンク構造を用いた Web イジング検索法

計算コストの高い検索評価を，膨大な数の Web ページに対して実行することは困難である．もっとも，すべてのドキュメントは不規則にリンクされているわけではなく,「多くの良質なページからリンクされているページは，やはり良質なページである」という状況であることが知られている．したがって，ユーザーが必要としている Web ページにリンクしている，もしくはリンクされているページは，必要性の高いページである可能性が高い．そこで，すべての対象ページを検索評価するのではなく，既に処理した Web ページから得られた情報を有効に利用する手法として，磁性体の基本的モデルを表現するイジングモデルの応用を考える．

すなわち，本稿では，Web structure mining の領域に関係するハイパーリンク構造に注目しながら，Web ページ検索において既に得られた情報を有効に利用するために，Web グラフに対するイジングモデルによるアプローチを提案する．実際，Web 検索で扱う問題では，ユーザーが必要とするページかそうでないかの 2 状態にページ群が分類されることから，各 Web ページをイジングモデルのスピンで表現することができる．加えて，一連の検索過程において必要なページかどうかの判定情報を外部磁場として組み込むことで，スピンフリップ・ダイナミクスを利用し，ハイパーリンクによる関連 Web ページの状態を確率的に推定する．以下，我々の提案するリンク構造型イジング検索 (Link structural Ising Retrieval, LIR と略す) アルゴリズムを述べる．

4.1 Web グラフに対するイジングモデルの適用

イジングモデルの 1 スピンは，Web 上の 1 ページに対応する．また，スピン間の結合定数 J は，Web ページ間の結びつきの強さに相当する．ここで，Web ページ群を強磁性体と考え，$J > 0$ とする．なお，二次元イジングモデルではスピン S_i に隣接するスピン S_j の数は 4 つであるが，これを Web ページ S_i がリンクしている Web ページ S_j していると拡張する．また，外部磁場定数 H_d，ページ S_a がユーザーが必要としているページらしさ $m_d(a)$，必要なページであるかどうかの閾値 θ_d を用いる．

まず，取り扱うスピンの数を M 個とし，対象となる Web ページに対してスピン $S_1, S_2, \cdots, S_M$ を割り当てる．また，あるスピン S_i に隣接するスピン S_j の数を $N(i)$ とする．すなわち，2 次元格子モデルで一般の磁性体構造を近似するのに対して，Web 構造は 2 次元格子モデルに制約されない．そのため，スピン S_i のリンク数 $N(i)$ は常に 4 とはならない．しかしながら，イジングモデルにおいて相互作用エネルギー E_a を計算するためのスピン S_j の個数は 4 である．よって，本稿では，イジングモデルにおける相互作用エネルギー E_i を以下のように変形する．

$$E_i = -\frac{4J}{N(i)} \sum_j S_i S_j - H_d(m_d(a) - \theta_d) S_i \quad (10)$$

また，スピン S_i をフリップした時の変動エネルギー ΔE_i を次式で与える．

$$\Delta E_i = \frac{8J}{N(i)} \sum_j S_i S_j + 2H_d(m_d(a) - \theta_d) S_i \quad (11)$$

ここで，次の LIR アルゴリズムを提案する．

【LIR アルゴリズム】

（1） まず，全 Web ページを検索物である状態（下向きスピン：$S_j = -1$）とし，検索物候補だけを格納したリストを生成する．

（2） 検索物候補のリストの中からランダムに 1 つのページ S_a を選択する．

（3） ページ S_a の検索物らしさ $m_d(a)$ を求める．

（4） 以下の操作を数回繰り返す．

a S_a からリンクされるページ S_i を選択する．ただし，S_i は検索物らしさを測定していないものとし，このような S_i が存在しない時は (5) に移る．

b $e^{-\beta \Delta E_i}$ に比例する確率で，S_i をフリップする．

c S_i の値が -1 から 1 に変わった時，S_i を検索物候補リストに入れ，逆に S_i の値が 1 から -1 に変

わった時，S_i を検索物候補リストから取り除く．

（5） 検索物候補がなくなるまで，2 から 4 の操作を繰り返し検索する． □

なお，LIR アルゴリズムを用いた実験内容の詳細は，次章で説明する．

4.2 相互作用エネルギー正規化の理由

前節において，スピン S_i へのリンクの強度がすべて 4 になるように正規化した．例えば，$N(x) = 100$ をもつスピン S_x へ張っている一本のリンクと $N(y) = 2$ であるスピン S_y へ張っているリンクでは，リンクの強さは異なっていると考えるべきである．ここで，スピン S_x へリンクしているスピンを $S_{x1}, S_{x2}, \cdots, S_{x100}$，スピン S_y へリンクしているスピンを S_{y1}, S_{y2} とする．一例として，$S_{x1}, \cdots, S_{x8} = 1$，$S_{y1}, S_{y2} = 1$，$S_{x9}, \cdots, S_{x100} = 0$ を与える．この時，スピン S_x へリンクしているスピンのうち，下向きのスピンは 8 個あり，スピン S_y へリンクしているスピンのうち下向きのスピンは 2 個である．従って，リンク数だけで考えると，スピン S_x は $S_x = 1$ に変化すると考えられる．

しかしながら，リンク数比率を考えると，スピン S_y へ張っている全てのスピンが上向きであるのに対して，スピン S_x へ張っているスピンのうち上向きであるのはリンク数の 0.8 %しかない．すなわち，スピン S_x へのリンクのうち不要なリンクが 99.2 %を占めている状況は，S_x を表すドキュメントが，Web サイトへの入口ページ，著名 Web サイトのトップページやリンク集サイトなどの可能性が高い．なお，この種の Web ページ発見には，既存の検索エンジン (yahoo, excite 等) で十分であろう．

また，一般に，Web マイニングの対象となる検索は，このようなメジャーな Web ページやサイトの発見が目的ではなく，興味のある情報を含む割には，埋もれてしまいがちな Web ページやサイトを探していると考えられる．したがって，上述のスピン S_x の Web ページは，ユーザーの要求に合致しないことが多い．したがって，LIR アルゴリズムではリンク数でなく，リンク数比率によってリンクの強さを正規化すべきであると考えた．

5. 実験と考察

5.1 Web データを用いた実験

実験データとして，国立情報学研究所 NTCTR-3 Web 用に収集された jp ドメイン中心の約 1500 万ページの Web ドキュメントのリンク関係データ，およびアンカーテキストに関するデータを使用した．処理は，1.2TB (CDS3-8RS-8x160) ハードディスクに格納し，Sun Cobalt LX50（Intel Pentium III 1.5GHz 512MB SDRAM）や，Intel(R) Xeon(TM) Processor 1.80GHz 等から構成される複数台で行った．

まず，実験データから，リンクがある時は 1，ない時は 0 を対応させたリンク関係行列を作成するが，行列は疎行列である．そこで，非ゼロ要素（この場合は 1 のみ）の要素情報のみを 2 つのベクトル col-ind と row-ptr によって格納する圧縮手法の一つである CRS 法 (Compressed Raw Storage) を用いた [2]．なお，col-ind ベクトルはどこにリンクしているかの情報を格納するベクトル，row-ptr ベクトルは対応するページのリンク先情報が col-ind ベクトルのどの成分に格納されているかの情報を格納するベクトルである．

5.2 検索範囲の縮小

本実験では約 1500 万のページ群から，キーワード「mining」をアンカー部とアンカー部の前後に含むページ群を抽出し，さらにそのページ群から 7,8 程度で辿ることのできるページ群を加えて検索対象ページ群とした．その結果 LIR アルゴリズムを適用する検索対象ページの総数は 34,423 ページであった（但し，表 4 を除く）．

5.3 パラメータ変動による検索結果の挙動

まず初めに，式 (10) 中に含まれるイジングモデルを記述する熱力学的パラメータ，外部磁場結合定数 H_d，β，結合定数 J に対する検索結果の挙動を調べる．検索結果は検索比率 (retrieval rate) と検索精度で評価する．検索比率とは検索対象ページ総数に対して LIR アルゴリズムで評価処理が必要となったページ数の比，検索精度とは全検索によって得られるページ数に対して LIR アルゴリズムで得られたページ数の比である．

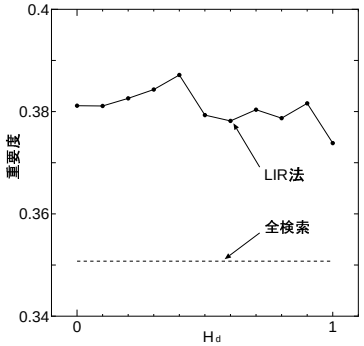

図 2 検索されたページの重要度の平均値 $(J, \beta) = (0.4, 0.2)$

図 2 は点線が全検索を行った時にヒットしたページの重要度の平均値であり，その値は 0.350756 である．また，実線は LIR アルゴリズムによってヒットしたページの重要度の平均値である．この図から LIR を行うことによって，重要度の低いページが排除され，重要度の平均値が大きくなるのは一目瞭然である．なお，以下で紹介するパラメータに関して同様の結果が得られた．

図 3 は外部磁場の結合定数 H_d に対する検索結果の挙動を示している．ここで実線は検索比率，破線は検索精度を示しており，共に値が低い方が望ましい．従って H_d の値は小さい方が良いとわかる．

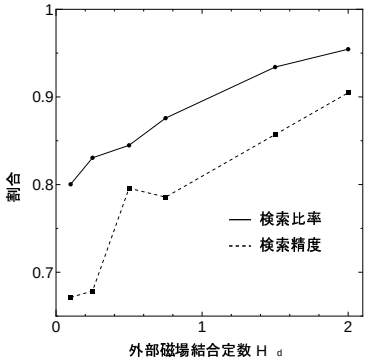

図 3　外部磁場結合定数 H_d に対する検索比率・精度 ($J = 1.0, \beta = 0.5$)

図 4 は，温度に反比例する β と検索比率の関係を示す．温度低下とともに評価処理回数が減少する．逆に，温度を上げると，高温下の分子運動がリンク関係を切断するかのように検索比率が上昇する．今後の検討を要するが，イジングモデルは臨界温度に到達すると，周りの振る舞いとは極めて異なる性質を持つ事ため，単調でない挙動を生じた可能性がある．

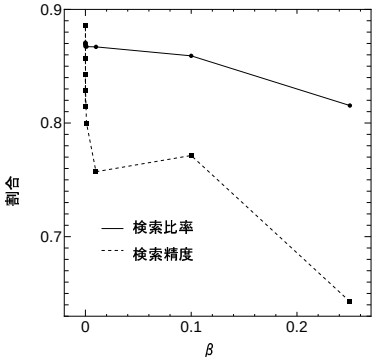

図 4　温度に反比例する β に対する検索比率・精度 ($J = 1.0, H_d = 0.5$)

次に，図 5 は，結合定数の変化による振る舞いである．$J = 0.5$ 近傍では評価比率が減少する．また $J = 0.5$ の点において，検索精度が下降し，評価処理回数が減少しており，望ましいパラメータ値を与える可能性がある．

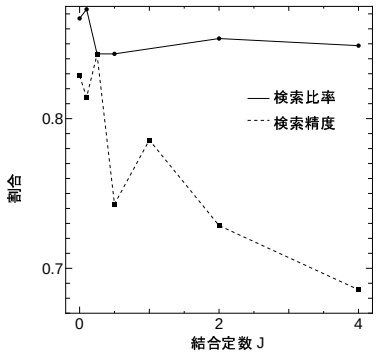

図 5　結合定数 J に対する検索比率・精度 ($H_d = 0.5, \beta = 0.5$)

次に，表 2 の項目にある隣接リンク段数が n であるとは，あるスピン S_a から n 回リンクを辿っていくことでスピン S_b にアクセスできる時，スピン S_b はスピン S_a の周辺スピンである事と考える．検索比率に関して，隣接リンク段数 n の変化による大きな差異は観察されなかったが，検索精度に関しては $n = 2$ の時に向上した．ここで，$n = 1$ の時は，リンク先がある Web サイトのトップの index ページにしかないという多数の Web ページが対象となるため，適切な値とは考えられない．従って，表 2 から，$n = 2, 3$ 程度が望ましいと思われる．

表 3 では，検索比率 (I) と検索比率 (II) において，LIR アルゴリズムの 4 ステップ目の処理回数の影響を調べた．検索比率 (I) は，対象となるスピンが張っているリンクの数，検索比率 (II) は，その値を 3 倍した．この時，表 3 の全ての場合において，検索比率は減少た．また，検索比率 (II) を，さらに数倍以上大きくした実験も行ったが，検索比率 (II) の場合と大きく変化しなかった．

最後の表 4 は，対象コミュニティサイズの増加に

表 2　隣接リンク段数 (n) による影響 (I)

n	検索比率	検索精度 (%)
1	87.11	83.8
2	87.00	81.8
3	86.97	81.2

表 3　隣接リンク段数 (n) による影響 (II)

n	(J,H_d,β)	検索比率 (I)	検索比率 (II)
1	(1.0 , 0.1 , 0.5)	0.799529	0.756965
1	(1.0 , 0.5 , 0.5)	0.848270	0.790315
3	(1.0 , 0.1 , 0.5)	0.804665	0.744264
3	(1.0 , 0.5 , 0.5)	0.847538	0.772083

表 4　総データ数の影響

総ページ数	検索比率 (%)
31	80.00
34423	78.12
213737	76.93

対する検索比率を調べたものである．データ数増加に伴い抽出比率が減少するが，特に大きな変化は観察されなかった．

6. 結論と今後の課題

近年の計算処理能力の向上により，大規模問題を扱うアルゴリズムや手法の実装が進んでいる．しかしながら，単調増加する Web ページをはじめとするネットワークコンテンツを適切に処理するには，演算能力の向上のみで解決ができないため，計算コストを削減する多様な技術が必要となる．そこで，本稿では，物理現象を扱うイジングモデルに基づく Web モデリングを行い，Web イジング検索モデルを提案した．また，提案した LIR アルゴリズムによって，検索対象となる Web ページ群から，適切に候補削減ができる可能性のあることを実データを用いた実験から確認した．なお，イジングモデルで扱われる温度に対する臨界現象などの性質に基づき，温度パラメータに対して Web イジング検索モデルにおいても特徴的な振る舞いを与えることを示した．

なお，本稿では，磁性体を表すモデルとして最も簡単で解析が十分に成されているイジングモデルを用いたが，ハイゼンベルグモデルのような連続的なスピン状態を用いたモデルへの拡張なども検討すべきである．また，イジングモデルにおいて臨界現象を適切に表現できる性質が良く知られているので，本モデルに現れる不規則な性質が理論的にどのような意味をもつかを深く追求する必要がある．

謝　　辞

Web リンク情報を提供して頂いた国立情報学研究所 NTCIR 情報検索システム評価用テストコレクション構築プロジェクト [22] に感謝する．本稿の一部は，文部省科学研究費 (13680482, 14019049, 14213101) の研究成果による．

文　　献

[1] http://www.alltheweb.com/
[2] R. Barrett, T. Chan, J. Donato, M. Berry and J. Demel: “反復法 Templates 応用数値計算ライブラリ,”(訳者: 長谷川里美, 藤野清次, 長谷川秀彦), 朝倉書店, (1996).
[3] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener: “Graph structure in the Web,” Proc. of the 9th International WWW conference, pp.247–256, 2000.
[4] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson and J. Keinberg: “Mining the Web’s Link Structure,” IEEE Computer, Vol.32, No.8, pp.60–67, 1999.
[5] S. Chakrabarti: “Mining the Web: Analysis of Hypertext and Semi Structured Data,” Morgan Kaufmann Publishers, (2002).
[6] 福地健太郎, 豊田正史, 喜連川優: “Web Community Browser における探索機構の実装と評価,” 電子情報通信学会技術研究報告, Vol.102, No.209, pp.473–479, 2002.
[7] G. W. Glake, S. Lawrence, C. L. Giles and F. M. Coetzee: “Self-Organization and Identification of Web Communities,” IEEE Computer, Vol.35, No.3, pp.66–71, 2002.
[8] http://www.google.com/
[9] 廣川佐千男, 池田大輔: “Web グラフの構造解析,” 人工知能学会誌, Vol.16, No.4, pp.525–529, 2001.
[10] K. Hotta, M. Tanaka and T. Mishima: “Multilevel Ising Search for Human Face Detection,” SPIE-Applications of Digital Image Processing XXI, pp.202–213, 1998.
[11] H. Kautz, B. Selman and M. Shah: “The hidden web,” AI magazine, Vol., No.2, pp.27–36, 1997.
[12] 川原稔, 河野浩之: “文献データベース情報検索ナビゲータの構築と評価,” 人工知能学会, pp.19–24, 1997.
[13] Hiroyuki Kawano, Minoru Kawahara: “Mondou: Information Navigator with Visual Interface,” Data Warehousing and Knowledge Discovery, Second International Conference, DaWaK 2000, pp.425-430, 2000.
[14] 河野浩之, 川原稔: “Web 検索におけるテキストマイニング,” 人工知能学会誌, Vol.16, No.2, pp.212–218, 2001.
[15] 河瀬基公子，川原稔，岩下武史，河野浩之，金澤正憲: “Web コミュニティ発見のための大規模 Web グラフに対するデータ圧縮計算手法,” データベースと Web 情報システムに関するシンポジウム (DBWeb2002), pp.423-430，2002.
[16] J. M. Kleinberg: “Authoritative Sources in a Hyperlinked Enviroment,” Journal of the ACM, pp.604–632, 1999.
[17] 小島 秀一, 高須 淳宏, 安達 淳: “Web ページ群の構造解析とグループ化,” NII Journal, No.4, pp.23–35, 2002.
[18] R. Kosaia and H. Blockeel: “Web Mining Research: A Survey,” ACM SIGKDD, Vol.2, pp.1–15, July 2000.
[19] 村田剛志: “ハイパーリンクの結合関係に基づく Web コミュニティの構造分析,” 人工知能学会（第 16 回）全国大会, Vol.17, No.3, pp.322–329, 2002.
[20] 那須川哲哉, 河野浩之, 有村博紀: “テキストマイニング基盤技術,” 人工知能学会誌, Vol.16, No.2, pp.201–211, 2001.
[21] 野村賢, 河野浩之, 川原稔: “ROC 距離に基づく先読み検索手法の提案と性能評価,” 情報処理学会論文誌:データベース, Vol.42, SIG3(TOD10), pp.56-65, 2001.
[22] http://research.nii.ac.jp/ ntcadm/index-ja.html
[23] http://physicsweb.org/article/world/14/7/09
[24] 坂本比呂志, 有村博紀: “ウェブ・マイニング,” 人工知能学会, Vol16, No.2, pp.233–238, 2001.