リンク構造に基づくWebイジング検索モデルの提案

Proposal for Web Ising Retrieval Model Based on Link Structure

阿部 友一[1] 河野 浩之[2]

Tomokazu ABE Hiroyuki KAWANO

PageRank や HITS アルゴリズムをはじめとする多数の Web 構造マイニングの研究がなされている．しかし，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．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．

PageRank, HITS and other web mining algorithms have been developed for discovering web documents with usefulness and high quality. In web mining models, the significant issue is the scalability of web page groups and keeping a balance between the search time and search accuracy, which indicate the performance of the search engine. 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 reduces 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.

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 を起点に，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解析し，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や技術開発が活発に行われている [1, 3, 4, 6, 7, 9]. なお，情報抽出の際に用いられるデータは，Web ページ中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る．

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する 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，検索処理に関わる不要ページの排除や抑制効果を示すかの検証を行う．特に，実際の Web データを用いてモデル化と検索を行い，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 モデリングに用いるイジングモデルを説明する．3 章で，実際に提案する Web 検索モデルの提案を行い，4 章で 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．最後に，結論と今後の課題を 5 章で述べる．

2. 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．

2.1 イジングモデルの性質

2 次元イジングモデルは，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．

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) 式により，ある状態 S のハミルトニアン $H(S)$ を (2) 式で表す．

$$E_i = -J\sum_j S_i S_j - BS_i \quad (1)$$

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

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

[1]学生会員 京都大学大学院情報学研究科博士前期課程
[2]正会員 京都大学大学院情報学研究科システム科学専攻
kawano@i.kyoto-u.ac.jp

任意の状態 S の出現確率を，分布関数 $Z = \sum_S e^{-\frac{1}{k_B T}H(S)}$ を用いて，(3) 式で与える．ここで，k_B はボルツマン定数，T は温度である．

$$\text{ボルツマン分布} \quad \omega(S) = \frac{e^{-\beta H(S)}}{Z} \tag{3}$$

2.2 イジング探索法

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．

【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．

次に，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，(4) 式で与える．

$$B = H_d(m_d(a) - \theta_d) \tag{4}$$

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 [2] 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変化を示す性質を利用する点である．例えば，連続的な値を取る対象物として，画像データ内の顔検出 [2] などが対象となっており，イジング探索法を用いることによる検出回数の削減や計算速度の向上が研究されている．

3. Web イジング検索法の提案

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ため，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可能である．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できる．

3.1 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 スピンを，Web 上の 1 ページに対応させる．また，スピン間の結合定数 J を，Web ページ間の結びつきの強さとみなす．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 数は 4 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 に隣接しているモデルに拡張する．また，外部磁場定数 H_d，ページ S_a に対してユーザーが必要としているページらしさを $m_d(a)$，必要なページであるかどうかの閾値を θ_d とする．

まず，取り扱うスピン数を M 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．なぜならば，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からである．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，式 (5) と定義する．

$$E_i = -\frac{4J}{N(i)} \sum_j S_i S_j - H_d(m_d(a) - \theta_d) S_i \tag{5}$$

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 (6) 式で与える．

$$\Delta E_i = \frac{8J}{N(i)} \sum_j S_i S_j + 2H_d(m_d(a) - \theta_d) S_i \tag{6}$$

3.2 Web グラフに対する LIR アルゴリズム

本節では，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． □

4. 実験と考察

本章では，LIR アルゴリズムの検索性能を論ずる．

4.1 Web データを用いた実験

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，実験データとして使用した [8]．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）と，Intel(R) Xeon(TM) Processor 1.80GHz 等で構成する機器で行った．

最初に，実験データに対して，リンクがある時に 1，無い時に 0 を対応させた，疎行列となるリンク関係行列を作成する．そして，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) を用いる [5]．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．

4.2 検索範囲の縮小

本実験では，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，約 1500 万ページ群から抽出し，さらにそのページ群から 7,8 先リンク程度で辿ることのできるページ群を加えることによって検索対象ページ群を構成した．その結果，LIR アルゴリズムを適用する検索対象ページ総数は 34,423 ページとなった．

4.3 パラメータ変動による検索結果の挙動

初めに，式 (5) 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．

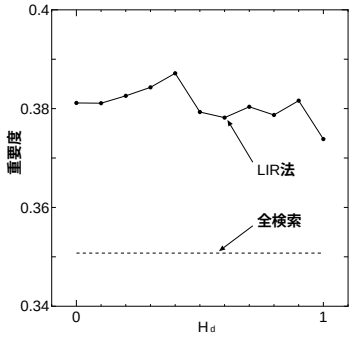

図 1: 検索されたページの重要度の平均値 (J, β) = (0.4,0.2)

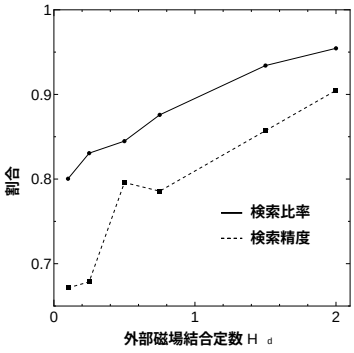

図 2: 外部磁場結合定数 H_d に対する検索比率・精度 ($J = 1.0, \beta = 0.5$)

図 1 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．図 1 からは，LIR アルゴリズムを適用することによって，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ことが分かる．なお，以下に示す幾つかのパラメータに対しても同様の結果を得た．

図 2 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す．ここで，実線は検索比率，破線は検索精度を示し，共に値が低い方が望ましい．すなわち，H_d の値は小さい方が良いと考えられる．

図 3 は，温度に反比例する β と検索比率の関係を示している．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振る舞い，検索比率が上昇する．今後，詳細な検討を必要とするが，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示すため，このような単調でない挙動を生じた可能性がある．

表 1 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，

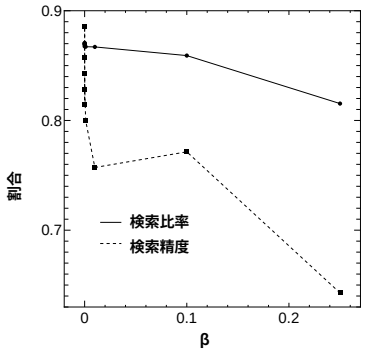

図 3: 温度に反比例する β に対する検索比率・精度 ($J = 1.0, H_d = 0.5$)

表 1: 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

検索比率 (II) は，その値を 3 倍したものである．この時，表 1 の全ての場合で，検索比率が減少している．なお，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) の場合と大きな変化はなかった．

5. 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，Web ページをはじめとして，単調増加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は解決ができないため，計算コストを削減する多様な技術を必要とする．そこで，本稿では，物理現象を扱うイジングモデルを用いて Web モデリングを行い，Web イジング検索モデルとして提案した．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，適切に候補削減できる可能性があることを実データを用いた実験により確かめた．さらに，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．

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課題である．また，イジングモデルにより臨界現象を適切に表現できることが良く知られているので，本実験により示した性質が理論的にどのような意味をもつかを追求する必要もある．

[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 [8] に感謝する．本稿の一部は，文部省科学研究費 (15017248, 13680482, 14019049, 14213101) の研究成果による．

[文献]

[1] S. Chakrabarti: “Mining the Web: Analysis of Hypertext and Semi Structured Data,” Morgan Kaufmann Publishers, (2002).

[2] 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.

[3] Hiroyuki Kawano, Minoru Kawahara: “Mondou: Information Navigator with Visual Interface,” Data Warehousing and Knowledge Discovery, Second International Conference, DaWaK 2000, pp.425-430, 2000.

[4] 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.

[5] 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.

[6] J. M. Kleinberg: “Authoritative Sources in a Hyperlinked Enviroment,” Journal of the ACM, pp.604–632, 1999.

[7] 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.

[8] http://research.nii.ac.jp/~ntcadm/index-ja.html

[9] 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.

阿部 友一 Tomokazu ABE
京都大学大学院情報学研究科博士前期課程 (2003 修了). 2001 京都大学工学部情報学科卒業．日本データベース学会学生会員.

河野 浩之 Hiroyuki KAWANO
1997 京都大学大学院情報学研究科システム科学専攻助教授．1990 京都大学大学院工学研究科数理工学専攻博士後期課程．工学博士．ネットワークシステム，データベース，データマイニングなど，情報システムに関わる範囲に興味を持つ．電子情報通信学会，情報処理学会，人工知能学会など所属.