話題範囲に着目した Web 閲覧履歴の空間的把握手法の提案

枝　隼也[†1]　　福原　知宏[†2]　　佐藤　哲司[†1]

これまで図書館や書籍を用いて調べていた授業やゼミでのレポート課題も，複数の Web ページやサイトを巡回することで，短時間に必要な情報を収集することが可能となった．本論文では、筆者らが提案している，閲覧したページの探索履歴を話題の遷移として可視化する手法を，自己組織化マップ (SOM: Self-Organizing Map) を用いて実装し，話題遷移の網羅度や，話題の拡がりを評価する．提案法を用いることで，提出されたレポートだけでは分からない，ページ閲覧の順序やレポート課題に対する調査範囲を空間的に把握することが出来ることを明らかにする．

A visualization method focused on range of topics from the Web exploration logs

Junya EDA,[†1] Tomohiro FUKUHARA[†2] and Tetsuji SATOH[†1]

The report in teaching and seminars were investigated using the library and books so far By visiting multiple Web pages and sites, Became possible to gather the necessary information quickly. In this paper, the authors have proposed, Visualization techniques as a transition to talk about the search history pages viewed, Self-organizing map (SOM: Self-Organizing Map) is implemented by using a Topics covered at the transition and to assess the extent of the topic. By using the proposed method Just do not know the report was submitted, The survey covers a sequence of page views and reports on the issues To show that we can grasp space.

†1 筑波大学図書館情報メディア研究科　〒 305–8550 茨城県つくば市春日 1–2
†2 産業技術総合研究所 サービス工学研究センター

1. はじめに

大量の情報が恒常的に生産・蓄積されているインターネット情報資源は，人々のあらゆる生活場面で利用されてきている．これまで図書館で調べていた授業やゼミでのレポート課題も，複数のページやサイトを巡回しそこに書かれた情報を集約することで短時間に作成することが可能となった．しかし，そこには以下に示す課題が残されている．

(1) レポートの出題者は，提出されたレポートを見ただけでは，様々な情報収集手段のいずれを使ってレポートを作成したのかが分からない．明らかに誤った情報が記述されていたとしても，どのような過程で誤りが混入したのかが分からないので，適切な指導を行うことが困難である．

(2) レポートの作成者は，訪問したページに記述された内容が，課題に対して十分な内容であるかを知ることができない．課題に関連するページは，ほとんど無限に検索されるので，どこまで調べたら十分に調べたことになるのかが分からない．また，自分の調べ方が効率的に行われていて無駄がないかを知りたい．

これらの課題を解決する手法として，筆者らは閲覧した Web ページをキーワード平面上で空間的に把握し共有することが有効であると考え提案している．本研究では，あらかじめ調査する話題の範囲が明らかである課題レポートの調査を対象とした履歴情報の空間的把握手法の確立を目的とする．このような課題レポートでは，どのような情報を調べるべきかの探索範囲があらかじめ決まっており，レポート作成者は，検索エンジンなどを使ってその範囲内を調査することから，レポート作成者毎の違いは，探索の順序と探索範囲をもれなく調査したかという，探索の過程に限定することができる．

そこで，本論文では，探索範囲となる話題空間全体を自己組織化マップ (SOM: Self-Organizing Map) を用いて二次元空間に展開し，個々のレポート作成者が探索する過程をマップ上の軌跡として表示する手法を実装し，話題遷移の網羅度や，話題の拡がりを評価する．その際に，探索過程で訪問した個々のページがカバーしている話題の広がりもマップ上に表示する．表示された軌跡と話題の広がりから，利用者であるレポート作成者が効率的に情報探索をしているか，また，課題として与えた探索範囲を網羅しているかの判断を支援する．探索範囲とする話題空間を二次元平面に展開する自己組織化マップは，T.Kohonen により提案された教師なしのニューラルネットワークアルゴリズムであり，高次元データを 2 次元平面上へ非線形写像するデータ解析方法である．このため，調査すべき全ての文書集合から抽出した，段落内での単語共起による特徴ベクトルを入力としてマップを作成する．

これは，同段落内に共起するキーワードは，関連性の高い概念を表していると考えられるためであり，そこから生成されるマップは，類似した概念を有する単語が距離的に近い位置に配置されたタグクラウドとなることから，単語の配置から話題の関連と広がりを把握することができる．一方，マップ上への履歴の表示は，履歴ページのキーワードとのマッチングによって行う．ユーザに提示する際には，各ページごとの話題がマップのどのあたりを閲覧していたのか時系列で示す必要があるため，時間調整スライダを操作し，その時点の閲覧済み話題空間がわかるように提示する．

以下 2 章で，関連研究について述べ，本研究の位置づけを示す．3 章では，筆者ら [1] が提案している，Web 閲覧履歴の空間的把握手法を説明し，4 章でその手法を実装したシステムの詳細について説明する．次に 5 章では，提案法によって作成されたマップについて話題の拡がり, キーワードのまとまりができていることを確認する．さらに，提案法に実データを適用し話題遷移の様子を考察する．最後に 6 章でまとめと今後の課題について述べる．

2. 関 連 研 究

Web 探索行動を支援する研究と Web 探索に SOM を応用しようとする研究について概観し，本研究の位置づけを明らかにする．

服部ら [3] は，複数のページを比較し，あるテーマにそって網羅的に閲覧する必要がある情報探索に対して，半自動的に関連情報を収集し，ユーザがまだ閲覧していない情報が多く含まれるページを Web 検索の上位に提示する手法の提案と実装評価を行っている．ページ話題の網羅度は，閲覧済みの単語空間と評価対象文書の，単語の共起度を基に算出される．本手法においても，SOM で作られた空間内で，ユーザ履歴内の単語の共起で話題の網羅度を算出しようと考えている．

陳ら [2] は，Web コンテンツ間に張られたリンク数によって，そのコンテンツ同士の関連度を計算し，関連度の高いコンテンツのみの探索空間を構成する手法を提案している．ユーザの意図に合った空間を提示することで，検索精度が向上することを示している．本提案法では,2 次元の探索話題の空間を SOM によって構築する．さらにユーザの閲覧履歴を用いて今までたどった話題の課程を示すことで，より空間的に話題の関連を把握することが出来るのではないかと考えている．

また，Web ページの話題を空間的に表現しようとする研究について村上ら [4] はユーザの時系列の興味空間上に外化記憶の想起を支援する，興味空間ブラウザを提案している．興味空間ブラウザは，過去に閲覧したページの閲覧日時，クリックしたアンカーテキストを保存し，数量化 3 類を用いてキーワード群の相関を計算し 2 次元空間上に表示している．相関の強いキーワード同士が近傍に配置されユーザに提示されることによって，自分が過去に閲覧したページの内容を理解できるとともに似ているキーワードと似ている内容のページを空間的に把握できるとしている．本手法では，空間的にキーワードを配置し，さらにそのキーワード空間上に閲覧したページキーワードをハイライトすることで閲覧したページのカバーする内容も把握する事ができる．

種市ら [5] は，ユーザの Web 探索行動における情報評価に着目し，短期大学生を対象として Web 探索行動の実験調査を行っている．その中でユーザは，Web ページの情報を必要なものであるかどうか評価する際，タイトルやページ構成見やすさなどで評価することも多く，コンテンツの詳細な内容や複数の情報源を比較検証したりする本来重要なコンテンツの質的な評価は欠落する傾向が示された．本研究では，意味的な関連性が距離に反映されているキーワードの 2 次元平面上で，ユーザの閲覧した Web ページの内容をキーワードのハイライトする．それにより，複数のページのカバーする内容を視覚的に把握できるようになる．

また，Web 探索への SOM の応用研究として中山 [6] は，SOM の手法に生体模倣技術である神経細胞移動の仕組みを持ち込み，新たに大規模疎行列のデータの解析を行う MIGSOM を提案している．また，MIGSOM を実際の Wikipedia データに適用し可視化することで，記事をいくつものクラスタに分割できることを示している．本手法では，ある特定の探索課題に関して文書や文書群を選定し，基礎的な SOM によってキーワードの関連性をもとにを空間的に配置している．

3. 自己組織化マップ:Self-Organizing Map

自己組織化マップ (SOM: Self-Organizing Map) は、コホネン (T. Kohonen) により提案された教師なしのニューラルネットワークアルゴリズムで、高次元データを 2 次元平面上へ非線形写像するデータ解析方法である。自己組織化マップは、入力層と出力層により構成された 2 層のニューラルネットワークである．今，入力層には分析対象となる個体 j の特徴ベクトルを xj(xj1,xj2,…,xjn)，出力層には 12 個のノードがあるとする．図 1 で示すように，出力層における任意の 1 つのノードは，入力層における特徴ベクトルとリンクしている．初期段階では乱数により各変数との間に重み mi(mi1,mi2,…,min) が付けられている．

(a) 入力層 と出力層におけるすべてのユニットの中から，最も類似しているユニット mc を探し出し，そのユニットを勝者とする．類似度は式 (1) によって計算される．

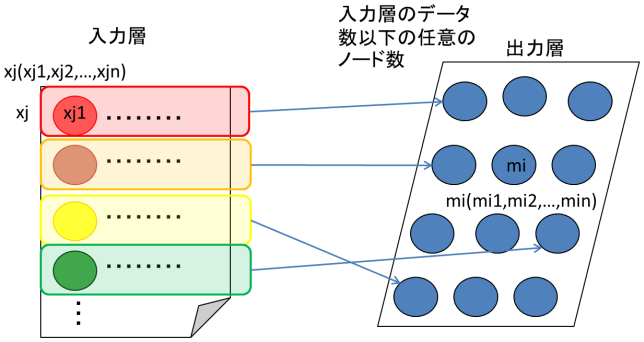

図 1　自己組織化マップの仕組み

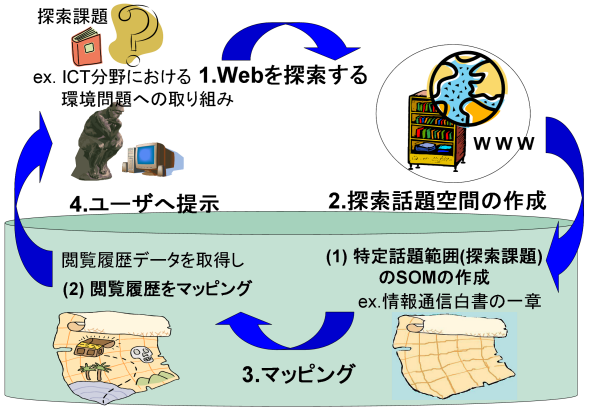

図 2　Web 閲覧履歴の空間的把握手法の処理の流れ

$$||xj - mc|| = min||xj - mi|| \tag{1}$$

(b) 勝者のユニットおよびその近傍のユニットの重みベクトル mi を更新する．更新は下記の式 (2) によって行う．

$$mi(t+1) = \begin{cases} mi(t) + hci(t)\lfloor xj(t) - mi(t) \rfloor & i \in Nc \\ mi(t) & i \notin Nc \end{cases} \tag{2}$$

$$hci(t) = \alpha(t) exp(\frac{-||rc - ri||^2}{2\sigma^2(t)}) \tag{3}$$

式 (3) の中の hci(t) は近傍関数で，ユニット c とその近傍のユニット i の近さによって xj の影響を調整する．式 hci(t) の中の α (t) は学習率係数で，rc と ri はユニット c と i の 2 次元上の座標ベクトルである．$\sigma^2(t)$ はユニット c の近傍領域 Nc の半径を調整する関数である．α (t)、$\sigma^2(t)$ は学習回数 (あるいは時間) を変数とする単調減少関数である．学習回数を変数とする最も簡単な単調減少関数は単調減少関数である．この t は学習回数，T は事前に設定した学習の総回数である．

(c) すべての入力特徴ベクトル (j=1,2,⋯,m) に対して (a)〜(c) を繰り返す

SOM の結果は，出力層の画面に図示される．個体の出力画面のユニットは，格子状 (正方形)，蜂の巣状 (六辺形) などが提案されているが，蜂の巣状が多く用いられている．蜂の巣状というのは，文字どおり蜂の巣のように正六角形のユニットを並べ，出力層の画面を構成する．出力層の画面は、上述のアルゴリズムにより、似ているもの同士を同じユニット、あるいはその近辺のユニットに配置する．

4. Web 閲覧履歴の空間的把握手法

提案する Web 閲覧履歴の空間的把握手法の流れを図 2 に示す．提案法は，その処理を大きく分けて

(1) 特定話題範囲 (探索課題) の SOM の作製段階

(2) 閲覧履歴をマッピングする段階

の 2 段階からなる．(1) では特定の探索話題空間を構成する文書を解析し，SOM を用いてキーワードを意味的な関係性を距離で表した 2 次元の平面に表現する．(2) では作製した SOM へ閲覧履歴をキーワードをハイライトする事でマッピングする．以下この手順に従って，各処理を説明する．

4.1 特定話題範囲の SOM の作成

SOM の計算をするためにキーワードの関連性を示した行列を作成する必要がある．そこで文書内の，各段落ごとにキーワードの共起回数を計算し行列で表す．これは，同段落内に共起するキーワードは，関連性の高い概念を表していると考えられるためである．つまり段落内共起関係で作成した行列に，SOM を適用した際，関連性の高い語が近くにマッピング

されるようになると考えられる．

4.2 閲覧履歴のマッピング

次に，特定話題範囲の SOM に閲覧履歴をマッピングしていく．マッピングしたイメージを図 4 に示す．ここでは，ユーザの閲覧したページ内のキーワードとマップのキーワードのマッチングを行い，そのキーワードをハイライトしていくことで，履歴のマッピングを行う．実際にユーザに提示する際には，各ページごとの話題がマップのどのあたりを閲覧していたのか時系列で示す必要があるため，時系列バーを操作し，その時間の閲覧済み話題空間を把握することができる．

5. 実装システム

前項で説明した提案法を実装したシステムについて各処理の実装を詳細に述べる．まず，(1) 特定話題範囲 (探索課題) の SOM の作製段階では，Web 閲覧履歴を話題の遷移として扱うため，各ページをキーワードの集合で表す必要がある．キーワード抽出のための形態素解析には MeCab[⋆1]を用いる．また，文書から特徴的なキーワードを抽出する手法として中川ら [7] の提案する連接頻度に基づく手法がある．本研究では文書からのキーワード抽出に，中川らの手法が実装された Perl モジュール TermExtract[⋆2] を用いる．

(2) 作成した SOM 上に閲覧履歴をマッピングする処理において SOM の計算にはオープンソースのデータ解析ソフト R[⋆3]を使用し，kohonen ライブラリを用いた．その計算結果より，キーワードのノード所属情報を基に，JavaScript を用いてキーワードを空間的に配置する．

6. 評 価 実 験

6.1 特定話題範囲 (探索課題) の SOM に関する評価

提案手法の処理のうち，(1) 特定話題範囲 (探索課題) の SOM の作製段階について，実データでの実験を行った．図 3 は，総務省 [8] の情報通信白書内の第 2 章グリーン ICT による環境負荷軽減と地域活性化の章を対象に SOM によってマップを作成したものである．ここでは，キーワードが 630 語抽出されさらに，段落数は 67 段落あり，630 行× 67 列の行列を抽出した．次に,15 × 15 で 225 ノードのマップを作成し，SOM の計算行った．一部

⋆1 http://mecab.sourceforge.net/
⋆2 http://gensen.dl.itc.u-tokyo.ac.jp/
⋆3 http://www.r-project.org/

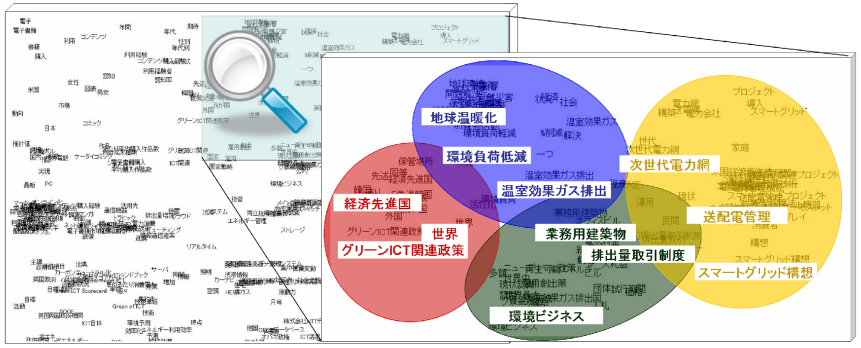

図 3 提案法を情報通信白書へ適用したマップ

を拡大し代表的なキーワードを見ると，関連する話題ごとに分割されている様子がわかる．「スマートグリッド構想」というキーワードの周りには,「次世代電力網」等といった関連するキーワードが並び，黄色で囲んだようなクラスタが形成されていることがわかる．その他も同じように,「環境ビジネス」に関連するクラスタ,「地球温暖化」に関するクラスタ等が抽出されていることがわかる．

6.2 閲覧履歴のマッピングに関する評価

次に提案手法の処理のうち，(2) 閲覧履歴をマッピングする段階，の処理について実データでの実験を行った．図 4 は 6.1 節の実験で用いたものと同様のデータから作製した特定話題範囲の SOM へ，ユーザ 1 名が「環境と ICT の取り組み」というテーマで閲覧した履歴ページのひとつをマッピングしキーワードをハイライトした様子である．ここでは，マップ上でハイライトされたキーワードが一部の固まったクラスタを形成していることがわかり，ユーザが閲覧したページ内の話題範囲がマップに表現できていることが分かる．表 1 は，同データでの SOM 上にハイライトされたキーワードと，閲覧した Web ページ中には存在したが，SOM 上に存在しなかったためハイライトされなかったキーワードの一例である．ハイライトされた語には「米国経済再生法」「送配電管理」等の主に米国での環境面と経済の取り組みを示すものがある．これらの語は SOM 上でも近い位置に配置されており，これらの語で話題範囲を示す事が出来ているといえる．しかし，閲覧履歴ページ内には存在したが，SOM 上には存在しなかったためハイライトされなかった語については「グリーン・ニューディール政策」「ピークシフト」などの特定の話題における具体的な政策名や専門用語など

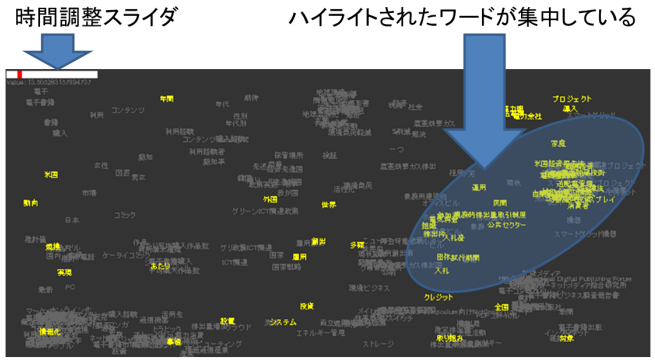

図 4　ページの閲覧順を考慮した提示

の語があった．これらの語は，今回用いたデータ内には出現していなかったため SOM 上に配置されていなかった．しかし，関連する話題の語が配置されているため話題の範囲を示すという点においては，実現できているといえる．

更に，あるテーマに沿った断片的な情報を閲覧していった場合，提案するマップ上でどのように履歴が話題範囲として提示されるのかを確認するための実験を行った．ここでは，特定話題範囲（探索課題）を「地球環境の現状」とし総務省の環境白書，第 1 章地球とわが国の環境の現状を用いて提案手法により SOM を作製した．そこに，朝日新聞記事データ集 2006 から記事分類：地球環境のページを収集し，各ページをそれぞれマッピングしていった．新聞記事とは，あるテーマに沿ってある時点で書かれた情報のまとめであり，全体のテーマからすると断片的な情報であると考えられるため，提案手法を適用し各記事の話題の範囲を表現できることを確認する．図 5 は，朝日新聞のある記事を提案法によって作成した，環境白書の SOM 上へマッピングしたものである．新聞記事は 1 つの記事にそれほど文量がないため，ハイライトされるキーワードが少なくなってしまっているが，ある程度空間的に偏ってハイライトされていることがわかり，その記事の話題範囲を表現できているのではないかといえる．表 2 は，同データでの SOM 上にハイライトされたキーワードと，ある記事ページ中には存在したが，SOM 上に存在しなかったためハイライトされなかったキーワードの一例である．こちらもハイライトとされなかった語は「米国立大気研究センター」「NCAR」等の具体的な組織の名称等の語が多く見られた．

表 1　情報通信白書から作成したマップへある Web ページをマッピングした際のキーワード

ハイライトされた語の例	ハイライトされなかったが履歴ページにあった語の例
米国経済再生法	グリーン・ニューディール政策
送配電管理	ピークシフト
排出量取引制限	スマートグリッド・シティ
公共セクター	コージェネ
電力網	ナトリウム硫黄電池

表 2　環境白書から作成したマップへのある朝日新聞記事のマッピングした際のキーワード

ハイライトされた語の例	ハイライトされなかったが履歴ページにあった語の例
地球温暖化	米国立大気研究センター
海洋	氷河地震
気候変動	ＮＣＡＲ
影響	効果
海水温	検証

6.3 考　　察

(1) 特定話題範囲 (探索課題) の SOM 作製についての評価実験では，情報通信白書の一部に提案法を適用し SOM を作製した．主観評価では，同じ話題を表すようなキーワードが抽出され，マップ上の近い位置に固まって配置されたと言えた．しかし,「日本」のような特定の話題以外でも広く使われる一般的な語が，ある特定のクラスタに固まって配置されるのではなく，複数のクラス他の間などマップ上で全体的に点在してしまっている事もわかった．

(2) 閲覧履歴のマッピングに関する評価実験について，図 5 は，環境白書の SOM へ朝日新聞記事データ集 2006 内の記事分類が「地球環境」である記事を提案法によってマッピングしたものである．新聞記事では，1 つの記事で文章量が少なくキーワード数も少ないため，作製した特定の話題空間の SOM 上へマッピングを行っても，多くのキーワードがハイライトされることは少なかった．しかし，少ないキーワードでもある程度空間的に偏ってハイライトされているページが見られたことから，提案法によって全体話題空間上である 1 つのページの話題範囲を表現することが出来ているといえた．

また，時間調整スライダを用いて閲覧履歴ページを時系列順に提示することによって，探索範囲の話題空間マップが充足していく様子を観察することが出来た．さらに，充足度については，キーワードのハイライトされる数で測定することを考えている．また，今後データ数を増やし追加実験を行う予定である．

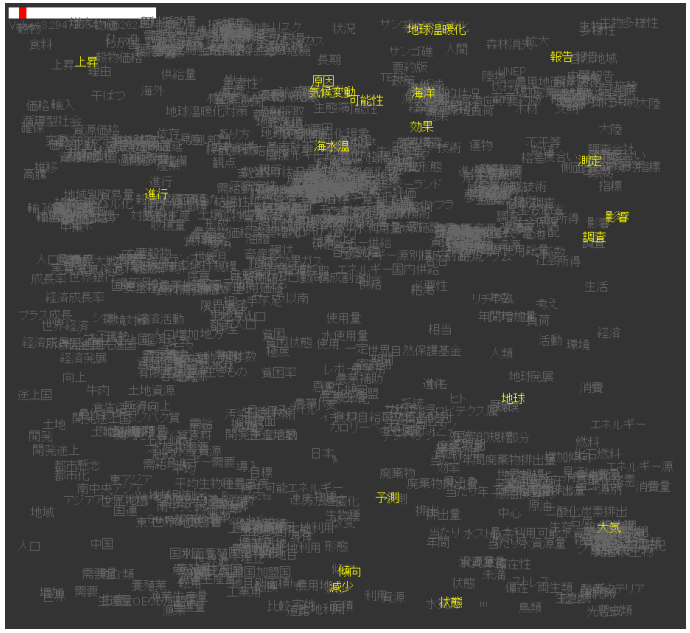

図 5 環境白書 SOM へある新聞記事ページのマッピング

7. ま と め

SOM を用いて，ユーザの探索課題に関するマップを作成し，そこに閲覧履歴をマッピングする手法について提案した．提案手法を実データに適用したところ，テーマごとにある程度，関連度の高いキーワードが近くにマッピングされることを確認した．また，あるテーマに沿った断片的な情報として新聞記事データを使用した実験を行った．環境白書から作製した SOM 上に，記事分類：地球環境の記事データをそれぞれハイライトしマッピングしていった結果，SOM 上でハイライトされたキーワードはある固まったクラスタを形成する様子が観察され，提案法によって，全体話題空間に対する断片的な情報の話題範囲を表現することが出来ると言えた．

今後の課題として，話題網羅度の空間的な計算手法の検討を行うこと，ユーザへの提示方法を検討したシステムの策定，複数人の使用を想定し，協調探索へ向けての具体的な仕様を検討することが挙げられる．

参 考 文 献

1) 枝隼也, 福原知宏, 佐藤哲司. Web 閲覧履歴の空間的把握手法の提案. 第 19 回 Web インテリジェンスとインタラクション研究会 WI2-2011-05, pp. 27–28, 2011.
2) 陳光敏, 小林亜樹, 山岡克式, 酒井善則. Web コンテンツ間類似度を用いた関連情報探索空間の構成法. 信学技法, No. 2004-02, pp. 19–24, 2004.
3) 服部元, 武吉朋也, 小野智弘, 滝嶋康弘. Web の半自動検索を利用した網羅的なテーマ関連知識習得支援方式. 第 8 回情報科学技術フォーラム,F-033, pp. 463–468, 2009.
4) 村上晴美, 平田高志. Www からの情報獲得・整理支援：思考・興味空間ブラウザ. 情報処理学会研究報告. 情報学基礎研究会報告, Vol. 2001, No.20, pp. 167–174, 2001-03-05.
5) 種市淳子, 逸村裕. エンドユーザーの web 探索行動：短期大学生の実験調査にもとづく情報評価モデルの構築. *Library and information science*, No.55, pp. 1–23, 2006.
6) 中山浩太郎. Migsom:神経細胞移動モデルに基づく自己組織化マップ〜大規模リンクドデータへの応用〜. Web とデータベースに関するフォーラム,5-3, 2010.
7) 中川裕志, 湯本紘彰, 森辰則. 出現頻度と連接頻度に基づく専門用語抽出. 自然言語, Vol.10, No.1, pp. 27–45, 2003.
8) 総務省. http://www.soumu.go.jp/.