D-042

リンク集の自動生成と視覚化
Automatic Genaration of Link Collections and its Visualization

瀬川 修† 河井 淳‡ 坂内 和幸‡
Osamu Segawa Jun Kawai Kazuyuki Sakauchi

1. まえがき

WWW 上の情報検索や知識発見のプロセスにおいては整理・分類されたリンク集が有用であることが多い。第 3 者が作成したリンク集は参照先のコンテンツに対する何らかの支持を表明したものであると言える。ただし、ある分野のリンク集一つだけに注目しても参照しているサイトの信頼性、網羅性などについて必ずしも保証されるものではない。何らかのコミュニティが形成されている分野であれば、整理・分類された複数のリンク集が存在することが予想される。そこで分野内のリンク集をできるだけ網羅的に収集し分析を行うことによってコミュニティにおける有用なサイトを見出し、その分野の動向を大まかに把握することも可能であると考えられる。著者らはこのような着想に基づくリンク集の再構成手法を提案している [6]。本稿では、さらに自動生成したリンク集を 2 次元マップ上に視覚化表示し、特定分野の主要サイトやキーワードの傾向を直感的に把握可能な新しい検索インタフェースについて述べる。

2. リンク集の自動生成

本手法では、まず入力として生成したいカテゴリの単語を与える。自動生成の処理は次の 5 つのプロセスから成る。

1. あらかじめ自動収集プログラムによって既存のリンク集ページを収集しデータベース化しておく。自動収集ではリンク集らしさの判別に、1) 外部リンクの数、2) タイトルタグやファイル名に「リンク集」や「link」などの文字列が含まれる、などの条件を用いている。このデータベースより目的のカテゴリに関連するリンク集セットを抽出する。

2. リンク集に記載された URL のページ本体を取得し、前処理としてページ内の特徴語抽出を行う。特徴語抽出は各 URL コンテンツのタイトルタグおよび本文に含まれる単語のうち TF·IDF 値が上位のものを選択することによって行なう。

3. 収集したリンク集セットに含まれる外部 URL のリストを U とし、U の中で出現頻度 (リンク集セットの中での参照数) が一定数以上のもの、あるいは URL の特徴語にカテゴリ語を含むものを Seed URL(当該カテゴリの代表的なサイト) として選択する。

4. Seed URL を含むリンク集 (図 1 の U の部分集合 P) の中で出現頻度が n 以上 ($n > 1$) の URL を選択し新たな URL 集合を生成する。この時各集合について構成要素の URL の特徴語をもとに集合の主題を表すラベル語を決定する。

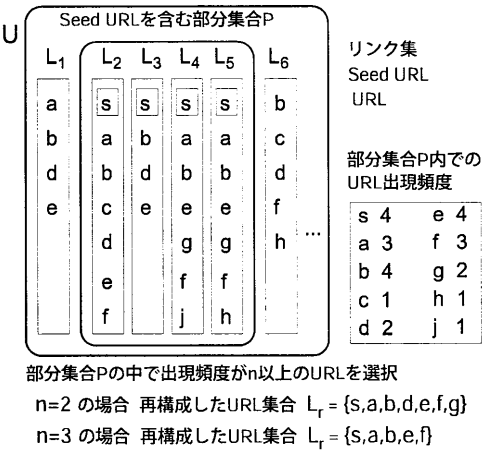

図 1: リンク集再構成手法

5. カテゴリと再構成された各 URL 集合のラベル語との意味的な関連性を評価し、評価値の高いものをリンク集として採用する。ここで意味的な関連性の評価にはカテゴリの特徴ベクトル (カテゴリ語とその関連語) と再構成された各 URL 集合の特徴ベクトル (ラベル語) のベクトル空間モデルによる類似度を用いる。

3. リンク集の視覚化

上述の手法では Seed URL ごとに主題が異なる複数のリンク集を生成するが、静的な HTML による結果の提示では情報量が膨大な上に、構成要素の類似したリンク集が数多く表示されるなど全体の見通しがよくない。そこで、自動生成した複数のリンク集を 2 次元マップで表示し、特定分野の動向を俯瞰するための視覚化手法を開発した。

本手法ではリンク集の構成要素であるサイトと特徴語 (ラベル語) を平面上にノードとして表現し、これらのノードが属しているリンク集との関係をアークで接続したグラフ構造を生成する。アークで結ばれたノードはバネモデルに類似した力学系によって互いに張力を受ける。また同時に各ノードはマップの中心より弱い膨張圧を受けることによって、リンク集としての接続関係を保ちながら外周方向に分散していく。このようなレイアウト配置によって、より多くのリンク集に属しているノードはマップの中心付近に滞留し、特定のリンク集に属しているノードは外周付近に集まってクラスタ (コミュニティ) を形成する。マップではアークの密集しているノードほどカテゴリ内における重要度が高く、注目に値するサイトまたはキーワードである。

†中部電力 (株) 電力技術研究所
‡TIS(株)

図 2: リンク集自動生成の例 (カテゴリ: 自動車)

4. システムの実装

4.1 検索ディレクトリの実装

リンク集自動生成手法を実装した検索ディレクトリを開発した。本システムでは 2004 年 7 月時点で 558 カテゴリのリンク集を自動生成し閲覧可能にしている。自動生成したリンク集の例を図 2 に示す。

4.2 リンク集視覚化ツールの実装と動作例

前述の手法に基づくリンク集の視覚化ツールを開発した。動作例を図 3 に示す。これは複数の自動生成リンク集を視覚化したものであるが、マップ上ではサイトと特徴語の各ノードは重複しないように表示されている。赤色の楕円ノードがサイトを表し (濃い赤は Seed URL)、緑色の楕円ノードが特徴語を表す。また、数字の付いた黄色の円ノードはサイトや特徴語の各ノードと接続したアークによってリンク集としてのまとまりを表している。本ツールではマップ自体が検索インタフェースとなっており、サイトノードをダブルクリックすることによって参照先のページを閲覧することができる。また、特徴語ノードをクリックすると、その単語を特徴語に含むサイトをハイライト表示する検索機能を備えている。

図 3 の「自動車」の例は比較的成熟した分野であり、業界の主要サイトとこれらを取り巻く関連サイトが概観できる。マップの右上にはメーカ系のコミュニティが形成されており、左上には関連メーカ系のコミュニティを見い出すことができる。またマップの中央付近には業界の中心となるサイト (JAF) が現れている。

5. 関連研究

Web からのコミュニティ発見の研究としては Kumar らによる Web Trawling[1]、村田の手法 [2] などがある。

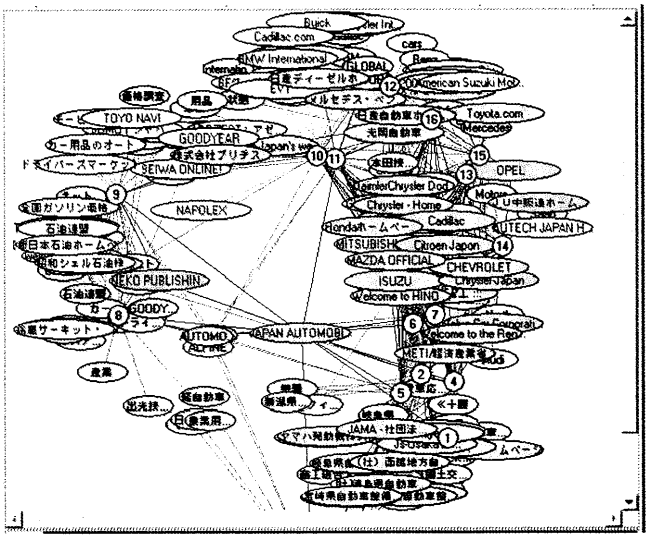

図 3: リンク集視覚化の例 (図 2「自動車」の視覚化)

提案手法は参照の共起性を利用する点でこれらの手法と類似しているが、サイト集合の発見では完全 2 部グラフによるリンク構造を前提としていない。また解析の対象を明示的にリンク集ページに限定している点が異なる。一方、Web 情報の視覚化の研究としては、サイト (集合) の関連性やキーワードの関連性を表示する手法の開発が行われている。前者の例としては、福地らの Web Community Browser[3] などが提案されている。また、TouchGraph は Google API を用いた検索結果の視覚化ツール Google Browser[5] を開発している。後者の例としては、高間らの Keyword Map[4] などが開発されている。これらに対し、提案手法では商用検索エンジンに依存せず、サイトとキーワードの関連性を同時に視覚化する検索インタフェースを実現している点が異なる。

6. むすび

本稿では特定分野の動向を反映したリンク集の自動生成と視覚化の手法を提案した。今後は本手法を発展させた知識発見のためのツールなどの開発も検討していく。

本稿で紹介した検索ディレクトリ及びソフトは試作版を http://netsurfersboard.com/nsb/ で公開している。

参考文献

[1] R.Kumar *et al.*, "Trawling the web for emerging cyber-communities", 8th WWW Conf., pp.403-416, 1999.

[2] 村田, "参照の共起性に基づく Web コミュニティの発見", 人工知能学会論文誌, Vol.16, No.3, pp.316-323, 2001.

[3] 福地 他, "Web Community Browser: Web コミュニティ構造の可視化と探策機構の実現", FIT2002 論文集, pp.205-206, 2002.

[4] 高間 他, "キーワードマップに基づく Web インタラクションと適合性フィードバックへの適用に関する考察", 人工知能学会研究会, SIG-KBS-A304-02, pp.7-12, 2004.

[5] TouchGraph, http://www.touchgraph.com/

[6] 瀬川 他, "コミュニティの動向を反映したリンク集再構成手法", 情処学会第 65 回全国大会 (3), pp.49-50, 2003.