リンク集自動生成システムの開発

知的情報収集を支援する新しい検索ツール

Development of a System for Automatically Generating Link Collections

An innovational information retrieval tool for intelligent information search

（電力技術研究所　お客さまネットワークG　情報通信T）

情報検索や知識発見で最も有用なのは専門家が作成したリンク集である。このような知見に基づき、本研究ではインターネット上の既存リンク集の自動収集と分析を行うことによって、特定分野の動向を反映したリンク集を再構成する手法を提案した。さらに自動生成されたリンク集を俯瞰するための視覚化ツールを開発した。

(Information and Communication Team, Customer Supply Network Group, Electric Power Research and Development Center)

Link collections are the most useful for information retrieval and knowledge search. Based on the concept, we have proposed a method for reorganizing link collections which reflect the trend of a category by collecting and analyzing process of link collections in the internet. Moreover, we have developed a visualization tool for birds-eye viewing the generated link collections.

1 研究の背景と目的

現在インターネットに蓄積されているWebページ数は数十億とも言われており、日々の情報収集に欠かせない社会インフラとなっている。本研究ではこれらの情報を体系化してユーザに提示する技術を開発し、社内の様々な専門業務（調査・研究・知財・マーケティングなど）における情報収集を支援する検索ツールの実現を目指した。

Webにおける情報検索や知識発見のプロセスを振り返ると、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して何らかの支持を表明したものである。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、網羅性について必ずしも保証されるものではない。ある程度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、目的とする分野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と考えられる。本研究ではこのような着想に基づいたリンク集の自動生成と視覚化の手法を提案する。

2 システムの概要

(1)コンセプト「リンク集によるコンテンツ評価」

提案手法では次の仮定に基づきサイトの評価を行う。

Ⅰ．より多くのリンク集から参照を受けるサイトほど情報源としての信頼性が高い。

Ⅱ．ある一つのリンク集から参照されているサイトの集合は互いに関連が深い。

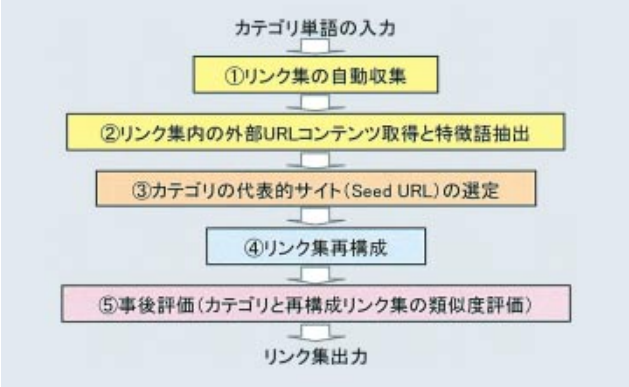
Ⅲ．特定のサイトを参照する複数のリンク集から参照されているサイトの集合は互いに関連が深い。

(2)リンク集再構成手法

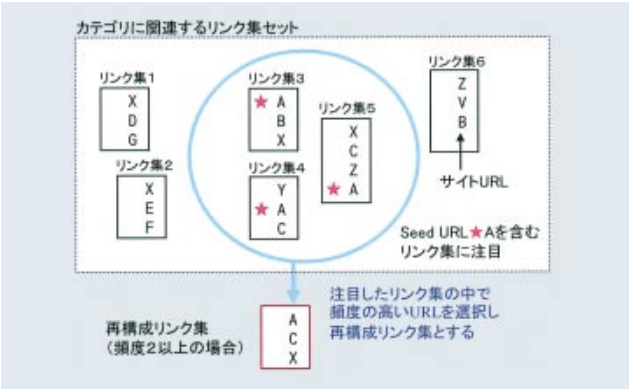
上記のコンセプトに基づくリンク集自動生成システムの処理概要を第1図に示す。また同図の手順④リンク集再構成の詳細を第2図に示す。

リンク集自動生成では、まず入力として生成したいカテゴリの単語を与える。自動生成処理は次の5つのプロセスから成る（第1図参照）。

①あらかじめ自動収集プログラムによってリンク集ページを収集しデータベース化しておく。このデータ



第1図　リンク集自動生成システムの処理概要



第2図　リンク集再構成アルゴリズム

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。

③リンク集再構成の前段階としてカテゴリ内の主要なサイト（Seed URL）を選定する。選定にはカテゴリ単語をページの特徴語に含むURLを選ぶ、あるいはリンク集セット内で頻度の高いURLを選ぶ、などの基準を用いる。

④リンク集セットの中でSeed URLを含むリンク集に着目し、その中で一定数以上で出現するURLを選択し、新たなリンク集を再構成する（第2図参照）。

⑤最初に入力したカテゴリ単語と再構成されたリンク集の特徴語の意味的な類似度を評価し、所定の類似度の基準を満たすものをリンク集として保存する。

（3）再構成リンク集の視覚化手法

上述の手法により自動生成した複数のリンク集を2次元マップで視覚化する技術を開発した。本手法ではリンク集の構成要素であるサイトと特徴語を平面上にノード（楕円のシンボル）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（線分）で表現するグラフ構造を生成する。各ノードはばねモデルとよばれる力学系を模擬したモデルによってレイアウト計算が行われ、意味的に近いノードがマップ上で近傍に配置される。

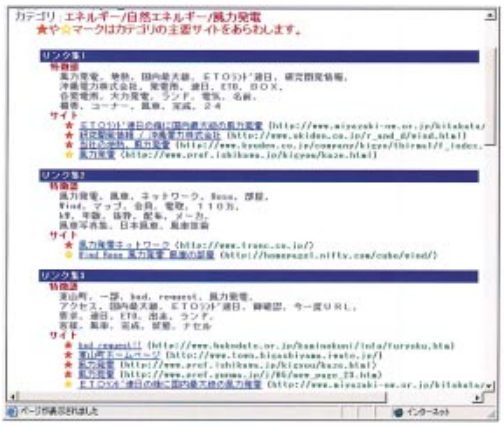
3 システムの実装

（1）検索ディレクトリの実装

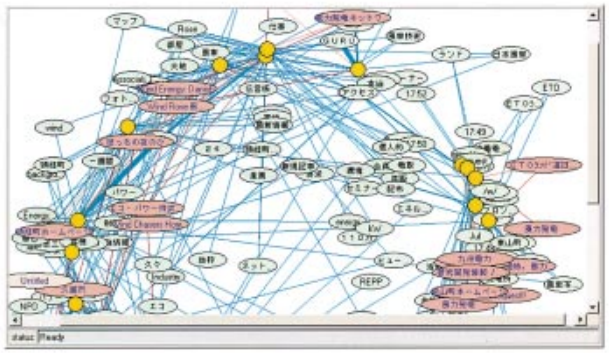
リンク集自動生成手法を実装した検索ディレクトリNetSurfer ﾄ Boardを開発した。自動生成したリンク集の例を第3図に示す。リンク集の★印はSeed URLを表す。

（2）リンク集視覚化ツールの実装

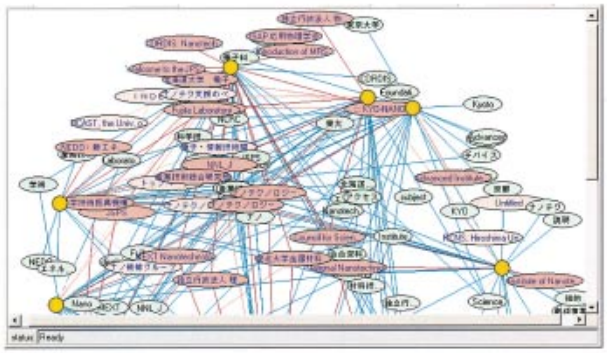
リンク集を視覚化するためのソフトウェアNetSurfer ﾄ Mapを開発した。動作イメージを第4図、第5図に示す。図は複数のリンク集を視覚化したものであるが、マップ上ではサイトと特徴語の各ノードは重複しないように表示されている。赤色の楕円ノードがサイトを表し（濃い赤はSeed URL）、緑色の楕円ノードが特徴語を表す。また黄色の円ノードはリンク集を表している。本ソフトでは表示されたマップ自体が検索インタフェースとなっており、サイトノード（赤色の楕円）をダブルクリックすることによって参照先のページを閲覧することができる。



第3図　自動生成リンク集の例（カテゴリ 風力発電）



第4図　リンク集視覚化の例（カテゴリ 風力発電 第3図のリンク集を視覚化したもの）



第5図　リンク集視覚化の例（カテゴリ ナノテクノロジー）

（3）システムの試験公開

平成15年末より社内において試作システムを試験公開し評価を進めている。また平成16年3月より下記URLでプレビュー版システムを一般公開し、視覚化ツール試用版の配布を行っている。

http://netsurfersboard.com/nsb/

4 今後の展開

ユーザからのフィードバックを得るために試験公開と評価を継続していく。またマーケティング業務などデータマイニング分野への適用も検討していく予定である。

（本システムはTIS株式会社との共同研究で開発した。）

執筆者／瀬川 修
Segawa.0samu@chuden.co.jp