

コミュニティの動向を反映したリンク集再構成手法

瀬川 修[†] 河井 淳^{††} 西俣 憲幸^{††} 坂内 和幸^{††}

[†] 中部電力株式会社 ^{††} TIS 株式会社

1 はじめに

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。本稿ではこのような着想のもとに、コミュニティの動向を反映したリンク集再構成手法を提案する。

2 リンク集の自動収集

リンク集の自動収集は検索エンジンを用いて行なう。カテゴリを表すキーワードと「リンク集」、「link」などの所定の単語を AND 条件で検索し、得られたページの中でよりリンク集らしいと判別されたものを保存する。ここではリンク集らしさの判別条件として、1) 外部サイトへのリンクが一定数以上存在する、2) タイトルタグに「リンク集」、「link」などの文字列が含まれる、3) ファイル名に link などの文字列が含まれる、などの条件を用いている。

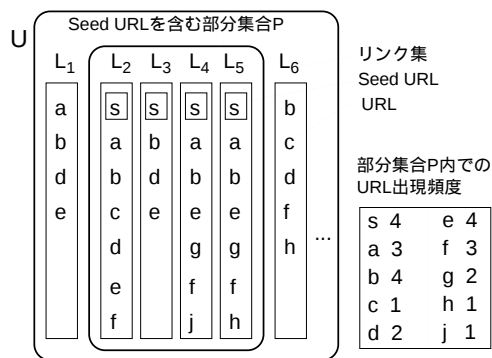
3 リンク集再構成手法

提案法は、1) 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、2) ある一つのリンク集から参照されているサイト群は互いに関連が深い、3) 共通のサイトを参照している複数のリンク集から参照されているサイト群は互いに関連が深い、などの仮定に基づくものである。

提案法では、まず前処理として前節で収集した特定カテゴリ分野のリンク集セットに含まれる外部リンクの全てについてコンテンツページを取得し、各 URL

ごとに特徴語を抽出する。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。リンク集再構成のアルゴリズムを以下に示す。

1. 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シード URL (当該カテゴリの代表的なサイト) として選択する。
2. シード 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。
3. 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 (コサイン測定) を用いる。



部分集合Pの中で出現頻度がn以上のURLを選択

$n=2$ の場合 再構成したURL集合 $L_r = \{s, a, b, d, e, f, g\}$

$n=3$ の場合 再構成したURL集合 $L_r = \{s, a, b, e, f\}$

図 1: リンク集再構成手法

なお手順 2. においては、イエローページ的な総合リンク集や一般的に引用されやすいマスコミ、ポータル、検索エンジンなどのサイトの影響を防ぐために、シード URL が含まれるリンク集で、かつシード URL の

A method for reorganizing link collection reflecting the trend of a community.

[†] Osamu SEGAWA (Segawa.Osamu@chuden.co.jp)

^{††} Jun KAWAI (kawai@karl.tis.co.jp)

Chubu Electric Power Co., Inc., Nagoya, 459-8522 Japan ([†])

TIS Inc., Osaka, 564-0051 Japan (^{††})

近傍に共起する URL のみを再構成の対象とするなどの選択基準も考えられる。

また手順 3 においてカテゴリ語の関連語を得るには、各 URL 集合のラベル語をパッセージとしカテゴリ語を検索質問としたパッセージ検索 [1] を行い、検索に適合したパッセージ全体の中で重要語を選択して関連語とするなどの方法を用いる。

4 評価実験と考察

実験システムに与える入力カテゴリ単語のみである。実験では手順 2 における n の値は 2 とした。カテゴリ “robotics” についてリンク集を再構成した結果例を表 1 と表 2 に示す。* 印はシード URL を表す。比較のため検索エンジン Google に対しキーワード “robotics” で検索した (言語指定なし) 上位 10 の URL を表 3 に示す。

表 1: 再構成リンク集 (1): カテゴリ robotics

<u>label words</u>	Robotics, Lab, Institute, Vrije, JANUS, University, Sources, Harvard, Internet, AI, Universiteit, AAAI, Columbia, Group, Brussel, Intelligence, Project, Index, Association, Laboratory
http://robotics.bu.edu/	(Boston University)
*http://www.ri.cmu.edu/	(The Robotics Institute)
http://robotics.eecs.berkeley.edu/	(UC Berkeley Robotics and Intelligent Machines Lab)
http://arti9.vub.ac.be/www/welcome.html	(AI Lab, Vrije Universiteit Brussel)
http://www.aaai.org/	(American Association for Artificial Intelligence)
http://hrl.harvard.edu/	(Harvard Robotics Laboratory Home Page)
http://www.cs.brandeis.edu/agents	(Brandeis University)
http://www.cs.indiana.edu/robotics/world.html	(Internet Robotics Sources)
http://www.cs.columbia.edu/robotics/	(Robotics Group at Columbia University)
http://robby.caltech.edu/	(Robotics home page)
http://borneo.gmd.de/AS/janus/pages/janus.html	(The JANUS Project)

表 3 の Google の検索結果では上位 1~3 に “robotics” を社名に含む無関係な企業サイトが入るなど本来意図したカテゴリ分野の情報とは関連ないものまで提示されてしまう。これに対し表 1 の提案法による再構成リンク集 (1) では大学の研究機関を中心に robotics 研究分野のコミュニティの代表的と思われるサイトが得られていることがわかる。また表 2 の再構成リンク集 (2) では (1) とは異なるホビーを主体とした robotics のコミュニティやキットパーツのメーカーなどのサイトが得られている。このように提案法によればシード URL の違いにより特定分野の異なるコミュ

表 2: 再構成リンク集 (2): カテゴリ robotics

<u>label words</u>	Robotics, Robot, Group, Society, Lynxmotion, Kits, Automation, Services, PARTS, Engineering, Amateur, Triangle, Area, Winnipeg, equipped, Seattle, DPRG, view, Atlanta, pages
http://www.robotics.com/	(Arrick Robotics - PC-Based Motion Control Products)
http://www.dprg.org/	(DPRG Dallas Personal Robotics Group)
*http://www.botlanta.org/	(The Atlanta Robot Club Web Site)
http://www.lynxmotion.com/	(Lynxmotion Robot Kits Menu)
http://www.seattlerobotics.org/	(Seattle Robotics Society)
http://www.robotgroup.org/	(The Robot Group)
http://www.robotstore.com/	(The Robot Store: Robots, Robot Kits, Robot Parts)
http://www.esit.com/	(Engineering Services Inc. Automation & Robotics)
http://www.portlandrobotics.org/	(PARTS Portland Area Robotics Society)
http://www4.ncsu.edu/unity/users/r/rnbowen/tar/	(The Home of Triangle Amateur Robotics)
http://www.winnipegrobotics.com/	(Winnipeg Area Robotics Society)

ニティを見出すことが可能であり動向分析などに効果的であると考えられる。

5 おわりに

特定分野のコミュニティ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。今後は本手法を用いたディレクトリの自動構築についても検討していきたい。

参考文献

- [1] G.Salton *et al.* “Approaches to passage retrieval in full text information systems”, in *Proc. of 16th ACM SIGIR international conference on research and development in Information Retrieval*, pp.49-56, 1993.

表 3: Google の検索結果上位 10: キーワード robotics

1.	http://www.3com.com/ (3Com.com)
2.	http://www.usr.com/ (US Robotics)
3.	http://www.usr.com/support/s-main-menu.asp (USR Support)
4.	http://www-robotics.cs.umass.edu/robotics.html (Robotics Internet Resources Page)
5.	http://www.ri.cmu.edu/ (The Robotics Institute)
6.	http://www.robotics.org/ (Robotic Industries Association Home Page)
7.	http://www.frc.ri.cmu.edu/robotics-faq/ (Robotics FAQ Table of Contents)
8.	http://www.irobot.com/ (iRobot Corporation)
9.	http://www.robotics.com/ (Arrick Robotics)
10.	http://www.cyber-robotics.com/ (Zeus 1000 Internet Marketing Robot)