データベースシステム 128－42
（２００２．７．１８）

ウェブグラフにおける密サブグラフの抽出とウェブコミュニティ

今藤 紀子† 喜連川 優†

† 東京大学生産技術研究所 〒 153-8505 東京都目黒区駒場 4-6-1
Email; {imafuji, kitsure}@tkl.iis.u-tokyo.ac.jp

要 旨

本論文では、ウェブグラフから構造の異なる二種類の密サブグラフを抽出し、そのウェブコミュニティとして性質を検証する。ウェブグラフは、ウェブページ、ハイパーリンクを意味するノード、エッジから構成される。ウェブコミュニティは同じトピックを扱うウェブページの集合を意味し、それを効率よく抽出する種々のグラフ論的アプローチが提案されている。これらのアプローチでは、ウェブコミュニティはウェブグラフにおいてグラフ構造的に密な部分とされるがその構造は手法によって異なる。我々は、完全 2 部グラフに基づく手法、最大フローアルゴリズムを利用する手法においてウェブコミュニティとして用いられる密サブグラフの構造の違いが、得られたウェブコミュニティの性質にどのような影響を与えるかを分析した。

Extracting Web Communities as a Dense Subgraph of Web Graph

Noriko IMAFUJI† Masaru KITSUREGAWA†

†Institute of Industrial Science, University of Tokyo
4-6-1 Komaba Meguro-ku, Tokyo, 153-8505 Japan
Email; {imafuji, kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract

In this paper, we describe the properties of the web communities derived as two distinct types of dense subgraphs in web graph. Web graph consists of nodes and edges, which represent web pages and hyperlinks respectively. A web community is a set of web pages having a common topic and so far various graph theoretical approaches have been proposed to extract web communities from web graph. Although all these approaches regard a web community as the dense part of web graph, the graph structres are different from each other. We analyze the effects of structural difference in the dense subgraph on the characteristic of web communities obtained by methods based on complete bipartite graphs and maximum flow algorithm.

1　はじめに

近年の WWW(World Wide Web) の急激な成長は、あらゆる分野にわたる非常に多くの情報の取得を可能にした。それと共に、莫大な情報量の中から自分が本当に必要な情報を得ることが益々困難になってきた。今後も WWW が成長し続けるとすればその傾向が強くなることは容易に予想される。こうした背景のなかで、効果的に情報の検索を行える様、莫大な数のウェブページの中から必要な情報に対してトピック的に関連の深いウェブページのみを抽出することが重要視されている。

近年ではこのような同じトピックを扱うウェブページの集合は一般的にウェブコミュニティと呼ばれ、これまでにウェブコミュニティを抽出するための様々な手法が提案されてきた。Kleinberg は [4] においてリンク解析に基づくアルゴリズムの HITS を提案した。これは、多くの良いオーソリティをリンクしているノードをハブ、逆に、多くの良いハブからリンクされているノードをオーソリティとし、これらオーソリティ、ハブノードの集合をウェブコミュニティとして抽出するというものである。また、ウェブページをノード、その間にはられているハイパーリンクを辺ととらえたグラフ構造を解析することによってウェブコミュニティを抽出する手法も [1, 2, 3] などによって提案されている。これらの手法では、ウェブコミュニティは [1] においては、完全 2 部グラフであり、[2, 3] では、コミュニティ内のノード同士の繋がりが外のノードよりも強いノード集合を指し、いずれにおいてもウェブグラフにおいてグラフ構造的に密な部分とされている。しかし、その密サブグラフの構造は手法によって異なり、その構造の違いが得られたウェブコミュニティの性質にどのような影響を与えているかは知られていない。

そこで我々は、すでにクロールしたウェブスナップショットを用いた実験により、完全 2 部グラフ及び最大フローアルゴリズムを利用した手法で得られるウェブコミュニティを比較し、その特徴を検証した。その結果、抽出母体が同じであってもそれぞれの手法によって得られるウェブコミュニティが異なることがわかった。本論文では、これらの実験の結果を示し、それぞれの手法によって得られるウェブコミュニティの傾向を分析する。そして、得たい情報のトピックによってどのように手法を選択すれば効率的にウェブコミュニティが抽出できるか述べる。

以下、2、3 節において、それぞれ完全 2 部グラフ及び最大フローアルゴリズムを利用したウェブコミュニティ抽出手法について詳しく説明し、4 節では、これらの手法を用いてウェブコミュニティを抽出した実験結果を示し、その結果から明らかになった二つの手法によって得られたウェブコミュニティの性質について述べる。また、5 節でまとめと今後の課題について述べる。

2　ウェブコミュニティ抽出手法１

本節では完全 2 部グラフを利用したウェブコミュニティ抽出手法について述べる。ウェブコミュニティ抽出における完全 2 部グラフの利用は初めに [1] によって提案された。Kumar らは [1] でウェブコミュニティにおけるコアとして完全 2 部グラフを利用し、ウェブグラフ内に含まれる完全 2 部グラフの数から、ウェブコミュニティがウェブ上にいくつあるかを示した。ここで述べる手法は、その手法を発展させたものである。

互いに素な 2 つのノード集合を X, Y とすると、X のどのノードも Y の全てのノードに結ばれているグラフを完全 2 部グラフという。また、$|X| = m, |Y| = n$ の時、この完全 2 部グラフを $K_{m,n}$ と表わす (図 1)。完全 2 部グラフがウェブコミュニティを得るた

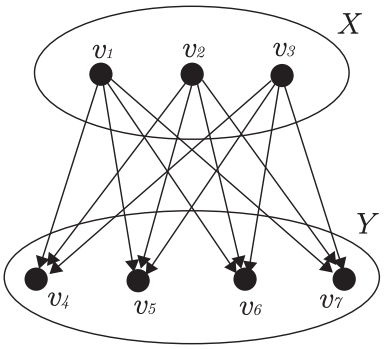

図 1: $K_{3,4}$; $X = v_1, v_2, v_3$, $Y = v_4, v_5, v_6, v_7$

めのグラフ構造に利用されるのは、ウェブ上に非常によく見られる現象である「互いに関連が強いページは、同じページから一緒にリンクされている」という概念に基づく。例えば、日産、トヨタ、ホンダなどのホームページは互いにリンクされていないものの、車会社へのリンク集などを持つページから同時にリンクされている。この場合、リンク集を持つ

ページが X のノード、それぞれの会社のホームページが Y のノードとなる。本手法は、シードを与えその周辺のグラフから全ての完全2部グラフ ($K_{3,3}$) を発見し、ノード集合 X, Y の重なり合いによってそれらのノード集合をマージしていき、最終的に得られたノード集合 Y をウェブコミュニティとして提示する。以下でその手順を示す。

input: 集合 S；シードノードの集合
output: いくつかのウェブコミュニティ C
$G = (V, E)$; 各 $v \in S$ を中心とした深さ 2 以内の周辺グラフ
r ; 集合 Y のマージのための閾値
1. G 内の全て $K_{3,3}$ を探索
2. 集合 Y ノードが全て同じ K についてそれらの集合 X をマージ
3. 同様に上で得られた集合 X に着目し集合 Y をマージ
$\mathbf{K}$=$K_{(1)}, K_{(2)}, \ldots, K_{(N)}$; 得られた完全2部グラフの集合
4. $K_{(i)}, K_{(j)} \in \mathbf{K}$ における集合 Y_i, Y_j について
$Y_i \cap Y_j \geq r$ の時、$C = Y_i \cup Y_j$

1 における $K_{3,3}$ の探索は深さ優先探索による。最終的に C がいくつか得られた場合は、何回のマージによって得られたかによってスコアをつけスコアの高いものを優位度の高いウェブコミュニティとして出力する。

3　ウェブコミュニティ抽出手法2

本節では [2] で提案された最大フローアルゴリズムを利用したウェブコミュニティ抽出手法について述べる。[2] では、ウェブコミュニティは「メンバーとなるページはコミュニティの外のページへの（又は、からの）リンクよりもコミュニティ内でのリンクを多くもつ」という条件を満たすウェブページの集合であると定義されている。図3はこの定義に基づいた簡単なウェブコミュニティの例である。

最大フローアルゴリズム [8, 10] とは、[9] によって定義された $s-t$ 最大フロー問題を解くためのアルゴリズムである。$s-t$ 最大フロー問題とは、有向グラフ $G = (V, E)$、辺容量 $c(u, v) \in \mathbf{Z}^+$、$s, t \in V$ を与えた時、全ての辺において容量を超えることなくソース s からシンク t への可能なフローの最大量を求めるものである。この時、最大フローと同時にそのグラフにおいて s, t を切り離すことができる切断辺の最小集合、つまり最小切断が求まることが [9] において示されている。たとえば、図3で示されたグラフが与えられた時、左側のノード集合内に s、右側のノード集合内に t が存在したとすると、最大フローアルゴリズムによって s から t への最大フローと最小切断（この場合、上下のノード集合を結ぶ5つの辺）が求まる。直観的には、シードの周辺グラフから粗な部分（辺）を切っていき最終的にシードと連結なノードがウェブコミュニティのノードであると見なされる。[2] では、定義されたウェブコミュニティがこの最大フローアルゴリズムによって求まることが証明されている。最大フローアルゴリズムを利用したウェブコミュニティ抽出の具体的な手順は以下のようになる。

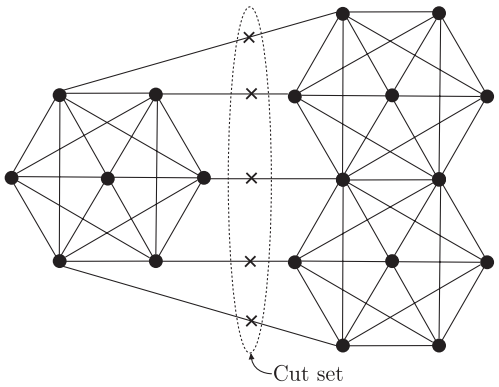

図 2: ウェブコミュニティ（左側のノード集合）の例

input: 集合 S；シードノードの集合
output: ウェブコミュニティ C
1. $G = (V, E)$; 各 $v \in S$ を中心とした深さ 2 以内の周辺グラフ
$k = |S|$ とする
2. s,t を V に加える
3. s から全ての $v \in S$ への辺 (s, v) を E に加える
このとき 辺容量 $c(s, v) = \infty$
4. 全ての $(u, v) \in E$ において辺容量 $c(u, v) = k$
$(v, u) \notin E$ の時
辺 (v, u)（辺容量 $c(v, u) = k$）を E に加える
5. 全ての $v \in V$ (但し、$v \notin S \cup s, t$) について
辺 (v, t) (辺容量 $c(v, t) = 1$) を E に加える
6. $s-t$ 最大フローアルゴリズムを実行

C=s と連結な全ての $v \in V$

7. C 内での辺数により各ノードをランクづけ
このうち高いランクのノードを S に加る

8. **1**~**7.** を数回繰り返す

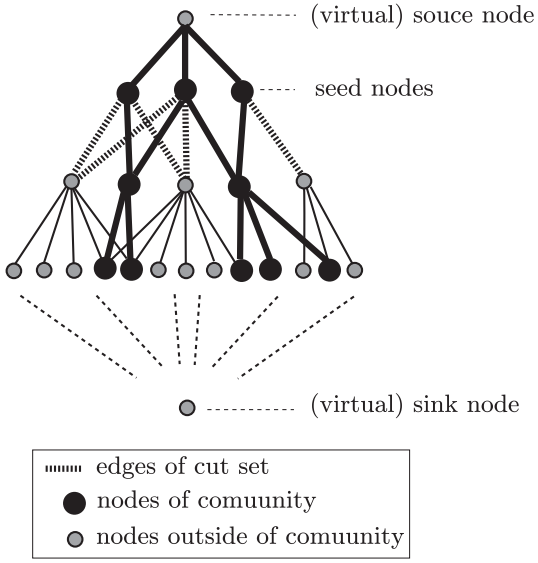

図 3: 最大フローアルゴリズムの適用例

図 3 は、この手順がどのように適用されるかを模式的に示している。図 3 において最上段及び最下段のノードが仮想的にソース及びシンクノードとしてグラフに加えられる。ソースノードからシードの各ノード、また、図中では省略しているが下二段の各ノードからシンクノードへそれぞれ辺が加えられる。手順 **6** によって最大フローが求まり、この時フローが辺容量に対して飽和状態になっている辺（図中の点線で示された辺）が切断辺となる。これらの辺によって切断されてもなおソースノードから到達可能なノードがウェブコミュニティのノードとなる。

4 実験結果と考察

4.1 実験の概要

実験で用いたシードノードは以下の二つである。

シードノード 1：www.orgel.com/index-j.html
サイトのタイトルは「パイプオルガンと音楽」で、そのコンテンツは、パイプオルガンに関する一般的な情報やコンサート情報、CD リリースの情報などである。

シードノード 2：www.alive-net.net/
サイトのタイトルは「地球生物会議」で、動物保護を目的とした活動や各種情報が得られるサイトである。

表 1 は、これらのシードノードにおける、周辺グラフのノード数、周辺グラフ内に存在する $K_{3,3}$、それらをマージして得られた $K_{n,m}$ の個数及び、両手法で得られたウェブコミュニティのノード数を示す。表における **Web Comm.(1)** 及び **Web Comm.(2)** はそれぞれ、完全 2 部グラフ、最大フローアルゴリズムを利用した手法によって得られたウェブコミュニティを意味する。但し **Web Comm.(1)** におけるデータは、得られたウェブコミュニティのうち最もノード数の大きいものについてである。

全ての実験を通して 2000 年に収集したウェブスナップショットを利用した。実験では、最初に与えるシードノードには極めて多くインリンク又はアウトリンクを持つものは避け、そのページで扱うトピックとしてあまりメジャーすぎないものを選んだ。最大フローアルゴリズムを利用した実験では、前節で述べた手順 **8** における繰り返しを 4 回行った。但し、様々なノードを与えて実験した結果、全ての実

表 1: シードノードと得られたコミュニティのサイズ

	www.orgel.com/index-j.html				www.alive-net.net/			
(繰り返し回数)	1	2	3	4	1	2	3	4
周辺グラフ (ノード数)	237	3266	3278	3499	35	717	4817	4840
含まれる $K_{3,3}$ の個数	31	162	162	166	0	62	194	194
含まれる $K_{n,m}$ の個数	2	9	9	10	0	5	20	20
Web Comm.(1) (ノード数)	14	8	8	8	0	5	12	12
Web Comm.(2) (ノード数)	4	13	62	88	3	15	17	63

験を通して1回目の処理の後得られるウェブコミュニティのメンバーはシードノードのみとなった。よって1回目のウェブコミュニティは、シードノードを中心にインリンク・アウトリンク両方向に深さ1のノードの中からリンク数の多いノードを新たにシードとして加える。

4.2 実験結果

表2,3は、シードノード1を与えたときの両手法で得られたウェブコミュニティのURLとそのサイトのタイトル又は内容を示している。完全2部グラフによる手法では複数のウェブコミュニティが得られた。得られたもののうち上位ウェブコミュニティには、音楽教育や一般的な音楽に関する情報が得られるウェブページが含まれている。二つ目のウェブコミュニティには、シードとなったウェブページと共に音楽関係のページが含まれている。シードのトピックであるパイプオルガンに関するウェブコミュニティはこの手法では存在しなかった。一方、最大フローアルゴリズムを利用した手法では、88のノードを含むウェブコミュニティが得られた。このうち、約7割がサイトのトピックとしてパイプオルガンを扱うもので、約9割が音楽というトピックに含まれ得るサイトであった。

表4,5は、シードノード2を与えたときの両手法で得られたウェブコミュニティのURLとそのサイトのタイトル又は内容を示している。完全2部グラフによる手法では、動物園のサイトの集合がウェブコミュニティとして得られた。シードのトピックである動物保護を扱ったサイトを含むウェブコミュニティはこの手法では得られなかった。一方、最大フローアルゴリズムを利用した手法では、63のノードを含むウェブコミュニティが得られた。このうち、約6割がサイトのトピックとして動物保護や絶滅危機に瀕している動物をトピックとして扱ったもので、約9割が動物に関連したトピックを扱うサイトであった。

4.3 考察

実験結果よりどちらの手法を利用しても得られたウェブページの集合は、ウェブコミュニティの概念「似たようなトピックを扱うウェブページの集合」に適っているといえる。得られたウェブコミュニティを比較すると、概して、完全2部グラフによる手法では、シードのトピックを包括するようなよりメジャーで一般的なトピックを扱うページ集合となっている。これは、シードノード2で得られた結果がその傾向を顕著に示している。シードノード2におけるトピックは動物保護に関するものであり、得られたウェブコミュニティは動物園のサイト集合となっている。動物園が動物関係で最も一般的なトピックであることは容易に推測できる。

逆に最大フローアルゴリズムによる手法では、シードのトピックに極めて近いかあるいはそれよりもさらにマイナーなある種限定されたトピックを扱うページ集合となっている。これは、特にシードノード1で得られた結果がその傾向を顕著に示している。シードノード1はパイプオルガンに関して一般的な情報が得られるサイトであった。最大フローアルゴリズムで得られたウェブコミュニティのページは、例えば、オルガン奏者のページや、オルガン制作会社、世界中に存在するパイプオルガンの詳細な情報が得られるサイトなどで、ほとんどがパイプオルガンというトピックに対してさらに専門的で非常に限定された情報が得られるサイトであった。

上述のような違いは、抽出する密サブグラフの構造の違いに起因すると考えられる。例えば、非常にメジャーなトピックに関しては、そのようなトピックを扱うサイトやそれらのサイトへのリンク集を持つようなサイトが多数存在するので容易に構造として完全2部グラフとなっている部分が抽出できる。一方、マイナーなトピックに関しては、マイナーなトピックを扱うサイトや、そのようなサイトへのリンク集が多く存在するとはあまり考えられず、このような状況の中、ウェブコミュニティになり得るような完全2部グラフが抽出できるとは考えにくい。実際、最大フローアルゴリズムによって得られたウェブコミュニティ内には、シードノード1の結果では、$K_{3,3}$ が一つ存在するが、シードノード2による結果では、$K_{3,3}$ より大きな完全2部グラフは存在していなかった。

以上のことから、あるトピックに関してより総合的、一般的な情報を得たい場合は、完全2部グラフによる手法、逆により深い、専門的な情報を得たい場合は、最大フローアルゴリズムによる手法を利用すると効果的にウェブコミュニティを得ることができると言える。

表 2: 完全 2 部グラフを利用したウェブコミュニティ 1

ウェブコミュニティ 1:	
www.asahi-net.or.jp/~bs2h-kir/	(音楽教育と MIDI)
www.yamaha.co.jp/edu/	(学校音楽教育支援のページ)
ss.jeugia.co.jp/~se/	(リコーダーと吹奏楽の部屋)
www.bekkoame.ne.jp/~shiyop/	(音楽教育リンク集)
www.fukuoka-edu.ac.jp/~kimurat/	(福岡教育大音楽教育)
www2s.biglobe.ne.jp/~taka-y/	(弘前教育大音楽教育)
www.nttl-net.ne.jp/onuki/	(養護学校の音楽)
www.wnn.or.jp/wnn-s/music/	(コネットミュージックワールド)
ウェブコミュニティ 2：	
www.wnn.or.jp/wnn-s/music/	(コネットミュージックワールド)
www.orgel.com/index-j.html	(パイプオルガンと音楽)
www.yamaha.co.jp/edu/	(学校音楽教育支援のページ)
bandits.aist-nara.ac.jp/~ryuuji-f/classicx/	(クラッシック大鑑)

表 3: 最大フローアルゴリズムによるウェブコミュニティ 1

スコア	URL 及び サイトのタイトル又は内容（全 88 ノード）	
19	www.bekknet.ad.jp/~sakazaki/eki/orglinks.html	(初期鍵盤楽器)
17	www.manaorg.co.jp/link.html	(オルガニスト、オルガンリンク集)
15	www2u.biglobe.ne.jp/~sorg/Suto-orgelbau.html	(須藤オルガン工房)
10	member.nifty.ne.jp/TsujiOrgan/	(辻オルガン)
10	member.nifty.ne.jp/Yamano/	(山野オルガン)
⋮	⋮	
1	rhic.sci.hokudai.ac.jp/orghist.html	(北海道大学のオルガン)
1	www.japan-net.ne.jp/~a-ueda/organ/	(パイプオルガン全般とリンク集)
1	www.hamamatsu-szo.ed.jp/irino-j/study/ongaku.htm	(音楽関連リンク集)
1	www.bach.or.jp/buecher.html	(バッハ関連書籍)
1	plaza27.mbn.or.jp/~matsui/org-joho.htm	(世界のオルガン情報)

表 4: 完全 2 部グラフを利用したウェブコミュニティ 2

ウェブコミュニティ 1:	
www.safari.co.jp/	(群馬サファリパーク)
www.hokkai.or.jp/kushiro-zoo/	(釧路動物園)
www.jin.ne.jp/kobe/ojizoo/	(王子動物園)
www.city.sapporo.jp/zoo/	(札幌市丸山動物園)
forum.coara.or.jp/cam/saru/	(高崎山自然動物園)
www.urban.ne.jp/home/tomoyuki/zoo.html	(徳山市立動物園)
www.aya.or.jp/~sczoo/	(埼玉県子供動物自然公園)
www.seagaia.co.jp/inf/infl/infl.htm	(フェニックス自然動物園)
www.arc-net.co.jp/kodoka/shoukou2/sc02.html	(旭山動物園)
www.islands.ne.jp/tobezoo/	(愛媛県立とべ動物園)
www1.999.com/uenozoo/	(上野動物園)
www.city.yokohama.jp/me/ygf/zoorasia/	(横浜市緑の協会)
www2.net-kochi.gr.jp/~nzp/	(高知市のいち動物公園)

表 5: 最大フローアルゴリズムによるウェブコミュニティ 2

スコア	URL 及び サイトのタイトル又は内容（全 63 ノード）	
20	www.orions.ad.jp/c/urls/word/a/n/imal+/wild-jp.html	(動物、野生動物リンク集)
12	www.nona.dti.ne.jp/~minmei/rink.htm	(動物関係リンク集)
11	www.sfc.keio.ac.jp/~t94122sk/extinct/new.html	(絶滅動物データベース)
4	www1e.mesh.ne.jp/edu-momo/chi-faq/2/kisyo.htm	(絶滅危機動物リンク集)
4	www.enjoy.ne.jp/~isshindo/kyouiku.htm	(教育関係リンク集)
⋮	⋮	
1	www.dive-one.com/	(絶滅動物データベース)
1	www.jca.apc.org/~kuzunoha/	(オオカミの仲間達)
1	homepage1.nifty.com/kiku-m/top.html	(小鳥のお宿)
1	www.wnn.or.jp/wnn-asia/	(アジアの自然と文化)
1	www.gran.ne.jp/~jorei/	(動物の条例法令集)

5 まとめと今後の課題

本論文では、完全2部グラフに基づく手法、最大フローアルゴリズムを利用する手法を用いて実際にウェブスナップショットからウェブコミュニティを抽出し、それぞれの手法において密サブグラフとして利用したグラフ構造がウェブコミュニティの特徴にどのような影響を与えているかを検証した。その結果、トピックとしてメジャーなものは完全2部グラフによる手法が、逆にマイナーなトピックに関しては、最大フローアルゴリズムを利用する手法のほうがより目的に合ったウェブコミュニティが得られることがわかった。

ウェブコミュニティはウェブ全体の一つの指針であり、ウェブコミュニティを正確に把握できればウェブ全体のトポロジを理解することができる。それ故、ウェブコミュニティを正確に抽出することは非常に重要な課題である。既存のグラフ理論的なアプローチによる手法を個々に利用しても、本当に必要とする情報が得られるようなウェブコミュニティを抽出できる場合とできない場合があり、実利用するにはまだ不十分であると言える。よって今後は、これらの手法をうまく統合させ、それぞれ個々の手法では抽出し得ない部分をも抽出することが可能な新たな手法を開発することを課題としている。

参考文献

[1] R.Kumar, P.Raghavan, S.Rajagopalan, and A.Tomkins, "Trawling the web for emerging cyber-communities," In Proc. 8th WWW Conference, 1999.

[2] G.W.Flake, S.Lawrence, and C.L.Giles, "Efficient Identification of Web Communities," In Proc. KDD 2000, 2000.

[3] G.W.Flake, S.Lawrence, C.L.Giles, and F.M.Coetzee, "Self-Organization and Identification of Web Communities," IEEE Computer,35(3), 66—71,2002.

[4] J.M.Kleinberg, "Authoritative Sources in a Hyperlinked Environment," In Proc. ACM-SIAM Symposium on Discrete Algorithms, 1998.

[5] S.Brin and L.Page, "The anatomy of a large-scale hypertextual Web search engine," In Proc. 7th WWW Conference, 1998.

[6] R.Lempel and S.Moran, "The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect," In Proc. 9th WWW Conference, 2000.

[7] J.Dean and M.R.Henzinger, "Finding related pages in the World Wide Web," In Proc. 8th WWW Conference, 1999.

[8] R.K.Ahuja, T.L.Magnanti, and J.B.Orlin, "Network Flows : Theory, Algorithms, and Applications," Prentice Hall, Englewood Cliffs, NJ, 1993.

[9] L.R.Ford Jr. and D.R.Fulkerson, "Maximal flow through a network," Canadian J.Math.,8:399—404, 1956.

[10] A.V.Goldberg and R.E.Tarjan, "A new approach to the maximal flow problem," In Proc. 18th Ann. ACM Symposium on Theory of Computing, 1986.