1T-9

頻出語の類似度を用いたＷｅｂコミュニティ抽出

樋上 真人　　志田 晃一郎　　横山 孝典
武蔵工業大学

1　はじめに

Web 上には様々な情報を掲載したページがあり，ユーザが目的にあったページを発見するために検索エンジンが用いられている．検索エンジンでは自分の求めているページを発見するためにキーワードを入力するが，求めているページを発見するキーワードを見つけることができないことがある．そこでWebページを話題の類似しているページごとに分類し，表示することができればユーザが目的のページを発見することが容易になると考えられる．

2　従来研究

一般にWebページは同じ話題のWebページ同士が互いにリンクで繋がっている．似た話題のWebページ集合をWebコミュニティといい.Webコミュニティを抽出する方法として「コミュニティの外のページへのリンクよりもコミュニティ内のページ同士のリンクを多く持つ」という条件を満たすWebページ集合をMax-Flowアルゴリズムを用いる方法がある．このようなコミュニティをMax-Flowコミュニティという[1].

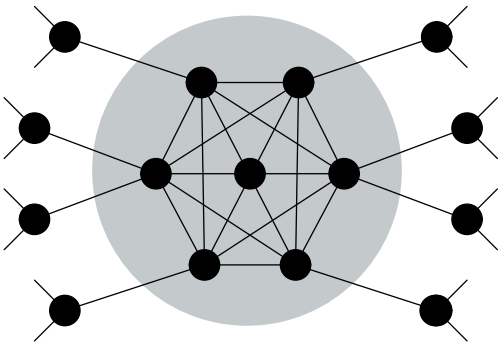

図１ Max-Flowコミュニティ

・Max-Flowコミュニティの抽出

Max-Flowコミュニティを抽出する手順を以下に示す．

1. コミュニティの種となるシードページを決定する．
2. シードページから２リンク辿りWebページを抽出する．
3. Max-Flowアルゴリズムを実行し，コミュニティにあたる部分を発見．
4. Webページにリンク関係を用いて点数付けを行う．
5. 点数の一番高いWebページをシードページに追加する．
6. １～５を結果が安定するまで繰り返す．

従来研究ではシードページに追加する方法にリンク関係を用いて算出された点数を使用しているため，内容の類似していない話題のページでもシードに選択されてしまう可能性があり，元々のシードページと関係ない話題のWebページがコミュニティに含まれてしまう問題がある．

本研究の目的はコミュニティに話題が類似していないページが混入することを防ぐことである．

3　提案手法

話題が類似しているWebページ同士ならばそこに現れる単語にも類似性があると考える．そこで，コミュニティの条件を満たすページに点数付けを行い新たなシードページを決定するが，このときの点数付けにシードページ群との頻出語の類似度を用いる．

シードページに新たに追加されたWebページの頻出語が初期シードページ群と類似していればそこからリンクを辿り抽出されるページも初期シードページと類似しているページであることが期待できる（図2）．

頻出語とはWebページから単語を切り出し，それらに対して単語の出現頻度TF(Term Frequency)を計算し，これが上位の語をそのWebページの頻出語とする．TFは〝単語の出現回数〟を〝文書中の総単語数〟で除算した値とする．

それぞれのWebページにおいて，頻出語のTFの２乗和が１になるよう正規化し，シードページ群とその他のページの頻出語のTFの内積を取り（表1），この値が最も高いWebページを新たに追加するシードページとする．

Extract Web Communities using Similarities of Appear Frequent Words.
Masato Hikami, Koichiro Shida and Takanori Yokoyama
Musashi Institute of Technology

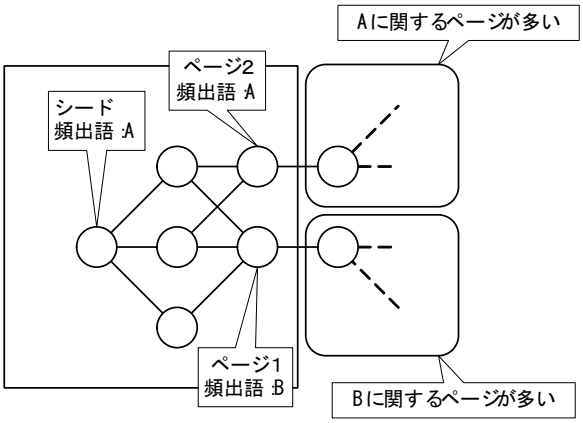

図２ 追加するシードページ

表１ 頻出語の内積

	シード	Page A	項ごとの積
PC	0.707	0	0
Windows	0.5	0.1	0.05
東京	0.5	0.8	0.4
名物	0	0.52	0
		内積	0.45

4 実装

対象とするＷｅｂページは.jp ドメインに属するページとし，Ｗｅｂページから単語を抽出するために形態素解析ツールである"Chasen"を用いて名詞のみを抽出する．

シードページからリンクを辿り Web ページを抽出する際に，Web ページの持っているリンク（リンク先）だけではなく，WEb ページへリンクしているページ（リンク元）も抽出する． これには Yahoo 検索 API を利用し，指定した Web ページへリンクしている Web ページを検索するクエリー（link:URL）を用いる．

実際にいくつかのコミュニティを抽出した結果，より類似しているページが抽出できた．

5 おわりに

本研究では頻出語を用いた Web コミュニティ抽出システムを作成した．

単語を Web ページから抽出する際に名詞のみを抽出したが，"日"などといった月日を表し，特定の話題を示さないような単語が頻出語となることがあった． 今後の予定として抽出する単語の選定を行う予定である．

参考文献

[1] 今藤紀子，喜連川優，"グラフ構造によるウェブコミュニティの特徴分析 :Max-Flow vs HITS"，電子情報通信学会技術研究報告 Vol.104，No.177，pp147-152 2004.