社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－MPS－62 (24)
2006－BIO－7 (24)
2006／12／22

進化的計算手法を用いたWeb検索キーワードのクラスタリング手法の提案

丸山 崇† 北 栄輔†

膨大な情報源であるWeb上において，検索対象分野に関する知識が乏しい場合，検索者の知識から導く検索語では検索要求を具体的に表現できず，目的の情報に到達しにくい．本研究では，キーワードとクラスタの関連度，キーワード同士の類似度を用いたクラスタリングモデルを構築し，進化的計算手法を用いてWeb検索キーワードをクラスタリングする手法を提案する．実験の結果，提案手法は，キーワードを関連のあるクラスタに分類し，また，類似性があるキーワードを同じクラスタに分類できていることがわかった．

Web Keyword Clustring System by using Evolutionary Computation

TAKASHI MARUYAMA† and EISUKE KITA†

If the Web user has no knowledge, he cannot express search words from his knowledge and it 's difficult to get information of purpose. In this paper, we propose the Web Keyword Clustring System by using Evolutionary Computation. The clustering model uses a similar level of the keyword and a related level and the keyword of the cluster. As a result, this system classifies the keyword into a related cluster and the keyword with the similarity into the same cluster.

1. はじめに

膨大なWeb情報の中から，インターネット利用者が求める情報にたどり着くために最も使用するサービスが検索エンジンである．検索者は，自身の検索対象を表す検索語を検索エンジンに入力して，検索対象の情報を得ようとする．検索者が入力する検索語は，検索対象分野に対する知識と経験によって異なる．専門家と一般の人が用いる検索式のキーワード数，検索精度を比較した結果，検索者が検索対象分野に関する知識が乏しい場合，検索者の知識から導く検索語で検索要求を具体的に表現できず，目的の情報に到達しにくいと報告されている[1]．また，ポータルサイトgooを運営するエヌ・ティ・ティ レゾナント株式会社は，検索に対するユーザの不満を以下の様にまとめている[2]．

(1) 適当な検索キーワードが思い浮かばない
(2) 検索結果をもっと多く表示してほしい
(3) 解決しないことが多く，結局，家族や知人に聞くことが多い
(4) 広告があって，画面が見づらい
(5) 絞り込み条件が足りない
(6) 検索精度をあげてほしい

そこで，本研究では，進化的計算手法を用いてWeb検索キーワードをクラスタリングする手法を提案する．本手法は，検索エンジンが提示するURLと検索キーワードを関連付けることで，キーワード間の関連付け（エッジを生成）する．そして，進化的計算手法を用いて，クラスタとキーワードの相関値，および，クラスタ間をまたがるエッジを考慮したクラスタリングの最適化を行う．提案手法は，検索目的を実現するような有益なキーワードを提示することで(1)，(3)の問題点を解決し，それにより検索エンジンから検索目的に適合した検索結果を得ることで(5)，(6)の解決を目指す．

本論文の構成は以下のようになっている．まず，第2章において，キーワードの抽出方法，キーワードとクラスタの関連度，キーワード同士の類似度を説明し，それらの値により構築したクラスタリングモデルについて述べる．次に，第3章において，進化的計算手法を用いたWeb検索キーワードのクラスタリング手法の手順について述べる．第4章では，実験結果を述べる．最後に，第5章において，まとめを述べる．

† 名古屋大学大学院 情報科学研究科 複雑系科学専攻
Graduate School of Information Science, Nagoya University

2. キーワードのクラスタリングの概要とクラスタリングモデルの提案

提案手法は，検索語と検索エンジンが提示したURLを関連付け（エッジを生成）する．それぞれのURLに着目すると，複数のキーワードが関連付けられており，そこからキーワード同士を関連付け（エッジを生成）することで，キーワードグラフが生成できる．キーワード同士にまたがるエッジは，同じURLを得るのに用いられた検索語であり，関連があることを表す．ある検索語に関連のあるキーワードを得たい場合は，その検索語にエッジがあるキーワードを列挙すればよい．そこで，提案手法では，以上のようにして得られたキーワードをクラスタリングする．

提案するキーワードのクラスタリングモデルは，キーワードとクラスタの関連度とキーワード同士の類似度を考慮したクラスタリングとして定義される．また，本クラスタリングモデルは，各キーワードが複数のクラスタに属する場合を考慮したクラスタリングである．

ここで，クラスタに属する，属さないの関係は0/1で表すことができる．よって，キーワードがN個，クラスタがM個であるとき，キーワードのクラスタリングはNM個の0/1からなる0/1組合せ最適化問題となる．そこで，進化的計算手法を用いることで，複雑で大規模な本クラスタリングモデルを組合せ最適化問題として単純化し，効率良く最適化する．

2.1 キーワードの抽出方法

提案手法は，検索語とURLのエッジを生成して，そこからキーワード間のエッジを生成することで，検索語と関連のあるキーワードを抽出するためのキーワードグラフを生成する．キーワードグラフは，ノードである各キーワードにエッジを生成することで，キーワードの関連付けを表現する．

まず，検索語を用いて得たWebページを，その検索語とエッジを生成する．十分な検索事例を経ることで，URLとそれを検索したキーワードのエッジのデータが得られる．次に，URLとキーワードのエッジから，同じURLを検索したキーワード間のエッジを生成する．エッジに同じURLを多く持つキーワード同士は類似度が高いと判断できる．

本システムでは，以上のように，各キーワードの関連付けからエッジを生成しキーワードグラフを構成する．検索者が用いた検索語と関連のあるキーワード（同じURLを検索したキーワード）は，キーワードグラフのエッジを調べることで，容易に提示できる．

2.2 キーワードとクラスタの関連度

キーワードをクラスタに分類するには，キーワードとクラスタの関連度を定義する必要がある．また，キーワードグラフにおいてエッジの重みを考慮してクラスタリングするには，キーワード同士の類似度を定義する必要がある．そこで，キーワードとクラスタの関連度を定義し，更に，関連度から生成できるキーワードの特徴ベクトルと，特徴ベクトルを用いたキーワード同士の類似度を定義する．

本クラスタリングモデルでは，TF/IDF法[3]によりキーワードとクラスタの関連度を求める．関連度を測定するキーワードをx，クラスタとなるキーワードをqとする．“xとqにエッジがあるURLの総数”を$URL_x_and_q$と置き，“xとx以外のキーワードにエッジがあるURLの総数”を$URL_x_and_\bar{x}$，“URLの総数”をURL_totalと置く．キーワードとクラスタの関連度W_t^dを，以下の式1で示す．

$$W_x^q = tf(x,q) \cdot idf(x) = \frac{URL\_x\_and\_q}{URL\_x\_and\_\bar{x}} \cdot log\frac{URL\_total}{URL\_x\_and\_\bar{x}} \quad (1)$$

$tf(x,q)$は，xとqにエッジがあるURLの出現頻度を表し，$idf(x)$は，xとx以外のキーワードにエッジがあるURLの希少度を表す．

2.3 キーワードの特徴ベクトルとキーワード同士の類似度

本システムでは，キーワードとクラスタの関連度を用いてキーワードの特徴ベクトルを定義する．キーワードをx，分類できるクラスタを$q_1, q_2, \ldots, q_M$とおくと，キーワードxの特徴ベクトルは以下の式2で示される．

$$\mathbf{x} = (W_x^{q_1}, W_x^{q_2}, \ldots, W_x^{q_M}) \quad (2)$$

また，コサイン相関値[4]を用いて，キーワード同士の類似度を定義する．キーワードxとキーワードyの類似度は，以下の式3の$sim(\mathbf{x},\mathbf{y})$で示される．

$$sim(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{|\mathbf{x}||\mathbf{y}|} = \frac{\sum_{i=1}^{M} W_x^{q_i} W_y^{q_i}}{\sqrt{\sum_{i=1}^{M} W_x^{q_i}}\sqrt{\sum_{i=1}^{M} W_y^{q_i}}} \quad (3)$$

2.4 提案するクラスタリングモデル

本システムで用いるクラスタリングモデルでは，キーワードとクラスタの関連度からキーワードの特徴を表す特徴ベクトルを生成し，キーワード同士の類似性を表すコサイン相関値を求める．キーワードのクラスタリングは，キーワードとクラスタの関連度を大きくするような分類が望ましい．また，クラスタ間をまた

ぐキーワード間のエッジにおいて，そのエッジを持つキーワード同士の類似度が低いような分類が望ましい．そこで，提案するクラスタリングモデルは，キーワード間のエッジの重みをコストとして扱う．更に，本クラスタリングモデルは，キーワードとクラスタの類似度をコストに追加する．

ここで，クラスタリングするキーワードを N 個，クラスタを M 個とする．また，キーワード x とクラスタ q の関係（属するときは 0，属さないときは 1）を $cutq(x,q)$ とし，関連度を $W_x^{q_i}$ と表す．更に，キーワード x とキーワード y のエッジがまたぐクラスタの総数を $cutkw(x,y)$ と表し，特徴ベクトルを $\mathbf{x}$，類似度を $sim(\mathbf{x},\mathbf{y})$ と表す．本クラスタリングモデルは，以下の式 4 を最小化する問題として定義する．

$$KWcut = \alpha \sum_{x=1}^{N} \sum_{i=1}^{M} \{cutq(x,q_i) \cdot W_x^{q_i}\} + \beta \sum_{x=1}^{N-1} \sum_{y=x+1}^{N} \{cutkw(x,y) \cdot sim(\mathbf{x},\mathbf{y})\} \quad (4)$$

右辺の第 1 項は，キーワードとクラスタのコストであり，右辺の第 2 項は，キーワードのエッジの重みのコストである．よって，$KWcut$ は，キーワードが関連度の高いクラスタに属し，類似度が高いキーワード同士が同じクラスタに属することで小さくなる．また，α，β は，それぞのコストの重み係数である．これらを調節することで，それぞれのコストが $KWcut$ に与える重みを調節する．

3. 進化的計算手法を用いた Web 検索キーワードクラスタリング手法の手順

本研究で提案するクラスタリングモデルは，キーワードとクラスタの関連度とキーワード同士の類似度を考慮したクラスタリングとして定義されるが，クラスタに属する，属さないの関係を 0/1 で表すことで，0/1 組合せ最適化問題に置き換えることができる．

そこで，本節では，進化的計算手法を用いて，提案したクラスタリングモデルに従ってキーワードのクラスタリングを最適化する手順を述べる．本研究では，進化的計算手法に遺伝的アルゴリズム（Genetic Algorithm: GA）[5] を用いる．

以下に，進化的計算手法を用いた Web 検索キーワードのクラスタリング手法の手順を述べる．

(1) キーワード収集
事前に，検索語を用いて検索エンジンから検索結果（上位 L 個の URL）を得て，キーワードと URL のエッジを生成する．いくつかの検索語を用いることで，各キーワードと URL のエッジのデータを作り，エッジに同じ URL を持つキーワード間にエッジを生成する．

(2) クラスタとキーワードグラフの生成
検索者が用いた M 個の検索語からなるクエリーから，検索語の組合から 2^M-1 個のクラスタを定義する．また，検索語と関連のあるキーワードを抜き出し，キーワードのエッジからキーワードグラフを作成する．

(3) キーワードとクラスタの関連度，キーワード同士の類似度の計算
キーワードとクラスタの関連度を TF/IDF 法により計算し，キーワードの特徴ベクトルを生成する．更に，キーワードグラフからエッジのあるキーワード同士の類似度をそれぞれの特徴ベクトルのコサイン相関値により計算する．

(4) クラスタリング
クラスタリングするキーワードが N 個，クラスタが M 個のクラスタリング問題を，クラスタに属する，属さないの関係を 0/1 で表し，NM 個の 0/1 からなる 0/1 組合せ最適化問題とする．クラスタリングモデルの評価式を適合度関数とし，進化的計算手法を用いて 0/1 組合せ最適化問題を最適化する．GA により最適化された解より，キーワードのクラスタリングを得る．

(5) キーワード提示
各クラスタにおいて，クラスタリングされたキーワードをそのクラスタにおける関連度（TF/IDF 法）によりソーティングし，提示する．

4. 実験結果

本節では，検索語に “java プログラミング エディタ” を用いたときの実験結果と考察を述べる．まず，実験をする前に，Java，プログラミング，Linux の GUI 関連の語を検索語とた Google の検索結果を得る．そして，Google の上位 100 ページ検索結果を用いて，URL とキーワードの関連付けしている．そして，そこからキーワード間のエッジを生成することで，キーワードグラフを構築した．検索語に用いたキーワードは計 95 である．また，クラスタリングモデルのパラメータにおいて，α は 10.0，β は 1.0 とした．

本システムにおいて，“java プログラミング エディタ” を検索語とし，キーワードをクラスタリングした結果が図 1 である．

まず，“java エディタ” クラスタについて述べる．

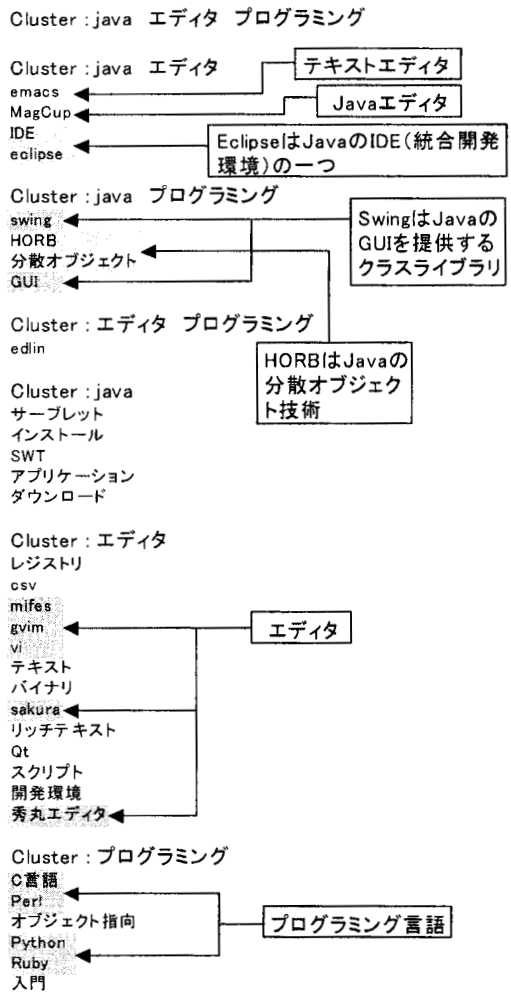

図 1 キーワードとクラスタリングの結果

'emacs' は，様々な言語のプログラミングエディタとして用いられており，Java の開発環境もある．'MagCup' は，個人が作成したフリーの java エディタである．'eclipse' は，Java のプログラミングを行う IDE（統合開発環境）の一つである．以上より，'MagCup'，'eclipse' は，java のエディタと関連があるキーワードであることがわかる．更に，'eclipse' と 'IDE' が同じクラスタに分類されていることがわかる．ただし，'開発環境' は，"エディタ" クラスタに分類されている．

次に，"java プログラミング" クラスタについて述べる．'swing' は，java の GUI を提供するクラスライブラリであり，'HORB' は，java の分散オブジェクト技術である．よって，'Swing'，'HORB' ともに，java のプログラミングに関連があるキーワードである．更に，'swaing' と 'GUI'，'HORB' と '分散オブジェクト' ともに同じクラスタに分類されていることがわかる．ただし，Java 用の GUI ツールキットである 'SWT' は，'java' に分類されている．

最後に，"エディタ",, "プログラミング" クラスタについて述べる．"エディタ" クラスタには，エディタである 'mifes'，'gvim'，'vi'，'sakura'，'秀丸エディタ' が分類されており，"プログラミング" クラスタには，プログラミング言語である 'C 言語'，'Perl'，'Python'，'Ruby' が分類されていることがわかる．

5. ま と め

本論文では，キーワードとクラスタの関連度，キーワード同士の類似度を用いたクラスタリングモデルを構築した．また，進化的計算手法を用いて Web 検索キーワードをクラスタリングする手法を提案した．

実験の結果，提案手法は，各キーワードを関連のあるクラスタに分類していることがわかった．また，関連のあるキーワード同士をできるだけ同じクラスタに分類していることがわかった．

本論文で提案した手法は，検索対象における知識が漠然としていたり，検索要求を満たすような Web 検索キーワードが分からない場合において，関連するキーワードを，クラスタリングして提示することで，Web 検索の検索補助が可能であると思われる．

参 考 文 献

1) 木谷強, 高木徹, 木原誠, 関根道隆. フルテキストと抽出キーワードを利用した情報検索. 情報処理学会報告, Vol. 96-NL-115, pp. 129–134, 1996.
2) エヌ・ティ・ティレゾナント株式会社. goo が目指す次世代検索サービスへのアプローチ, 2006. http://help.goo.ne.jp/public/goc04_01.pdf.
3) G. Salton and M. J. McGill, editors. *Introduction to modern information retrieval.* McGraw-Hill, 1983.
4) G. Salton, editor. *Automatic Text Processing.* Addison-Wesley, 1989.
5) J.H. Holland. *Adaptation in Natural and Artificial Systems.* The University of Michigan Press, 1975.