論文検索支援のためのグラフィカルナビゲーションシステム

曽原　寿允† 堀　幸雄†‡ 今井　慈郎†‡
香川大学工学部†　香川大学総合情報センター‡

1．はじめに

近年インターネットの発達により，大量の論文が電子化され蓄積，提供されるようになり，国内外において数多くの Web を媒体とした論文検索サービスが提供されている．これらのサービスにおいて，ユーザは主に文字列情報を元に論文の検索を行うが，検索結果には目的の論文以外のノイズも多く含まれ，その中から必要な情報を峻別することはユーザにとって負担は少なくない．

そこで，本研究では論文検索の支援として論文関係をグラフィカルに提示するシステムを作成する．本研究では文字列マッチングだけでは検索できない情報を抽出するため論文をクラスタリングし，対象クラスタ間の類似度に基づき，クラスタのリンク関係を可視化する．クラスタのリンク情報をユーザに視覚的に辿らせ，論文検索をナビゲートすることで，利用効率の向上を図ることが可能となる．

2．関連研究

グラフィカルなインタフェースを用いた論文検索支援の研究として，類似・関連する論文をグラフィカルに検索・表示する研究[1]やノードとエッジ構造を用いたサーベイ支援の研究[2]があげられる．

前者のアプローチは，一つの論文を入力として，それに関連性の高い論文をその周りに配置しグラフィカルに表示しているが，論文の数が大量になった場合を考慮していない．後者の研究では，論文等の関係をノードとエッジを用いたリンク構造による視覚化を行っている．リンク構造により，データ構造をグラフィカルに可視化することは，データ同士の繋がりや，位置関係などを具体的に知覚することができ有用であると考えられる．

本研究では，論文のクラスタリングを行い，大量の論文に対応する．また，ユーザにインタラクティブな操作を提供するため，リンク構造を用いた GUI を選択し，クラスタのリンク間を探索してもらうアプローチをとる．

3．システムの概要

本システムは以下の 2 つのモジュールから構成されている．

・論文処理モジュール
・ナビゲーションモジュール

3．1 論文処理モジュール

論文処理モジュールでは，システムを構築するためのデータの前処理を行う．論文処理モジュールの処理の流れは以下のようになる．

(1)分析対象となる論文群を取得
(2)論文データからタイトルと抄録を抽出
(3)タイトルと抄録を形態素解析し単語に分解
(4)形態素解析によって得られた単語に対し，TF-IDF 値を求めて単語の特徴量を算出し，論文の特徴ベクトルを作成
(5)得られた特徴ベクトルを用い論文集合のクラスタリングを行い，クラスタへ分類
(6)分類されたクラスタ情報をナビゲーションビューによりユーザへ提供

論文処理モジュールの処理の流れを図 1 に示す．

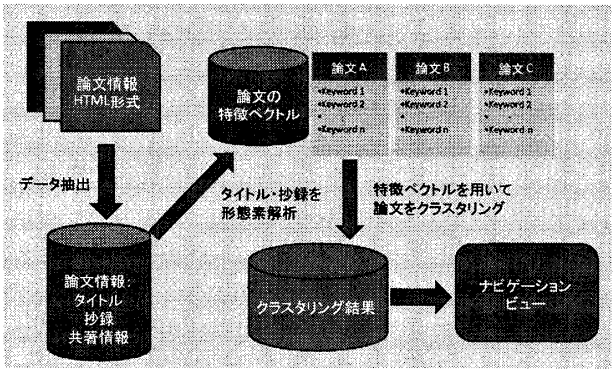

図 1. 論文処理モジュールの処理フロー

Graphical navigation system for searching paper
Toshimitsu Sohara†, Yukio Hori†‡, Yoshiro Imai†‡
†Kagawa University
‡ Information Technology Center, Kagawa University

Copyright © 2010 Information Processing Society of Japan.
All Rights Reserved.

3．1．1 特徴ベクトルの作成

まず，検索対象となる論文群を取得する．

取得した論文集合からそれぞれの論文に含まれるタイトルと抄録を抜き出し，形態素解析を行い単語に分解する．

得られた単語に対し，式1で表される TF-IDF 値 $w_{i,j}$ を求める．TF-IDF は、文書中の特徴的な単語を抽出するアルゴリズムであり、情報検索や文書要約などの分野で利用されている．

$$w_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i}) \quad (1)$$

$tf_{i,j}$ は文書 j における単語 i の出現回数，df_i は単語 i が出現する文書数，N は全体の文書数を表す．

論文データに含まれる単語のうち，TF-IDF 値の高い単語をその論文の特徴を表している単語として採用し，論文の特徴ベクトルを作成する．

3．1．2 クラスタリング

作成した特徴ベクトルには，論文の特徴を表した単語と，TF-IDF 値がセットで保存されている．この情報を元に論文集合に対してクラスタリングを行う．

クラスタリングの結果としてナビゲーションモジュールに提供する情報は以下があげられる．

・論文の所属クラスタ
・クラスタ同士の類似度

本システムでは，類似するクラスタ間のリンクを巡ることにより論文の探索を行うため，各クラスタ間の類似度が必要となる．

クラスタリングには軽量データクラスタリングツール bayon[3]を用いた．また，クラスタリング手法には Repeated Bisection 法を用いた．Repeated Bisection 法とは，クラスタリングツール CLUTO でも使用されているクラスタリング手法で、データ集合を繰り返し 2 分割することでクラスタリングを行う手法である。具体的には以下の 1-4 の処理を繰り返し実行しクラスタリングを行う．

(1) 分割するクラスタを1つ選択
(2) クラスタ中からランダムに 2 つ要素を選択し、それぞれが格納したクラスタを2つ作成
(3) 元のクラスタ中の全ての要素に対し、2 で選んだ要素との類似度を求め、類似度が高い方のクラスタに要素を追加
(4) 2 クラスタ間で要素の移動を行い、分割結果の洗練

3．2 ナビゲーションモジュール

ナビゲーションモジュールでは，クラスタリングの結果得られたクラスタ情報や論文の情報の探索を行う．ナビゲーションビューの作成には，SpringGraph[4]を用いた．図2は実際のナビゲーションビューである．中央にクラスタを表すノード表示され，その周りにそのクラスタに属する論文が配置されている．また，別のクラスタに移動するためのリンクが張ってあり，別のクラスタを選択することにより，クラスタ間を移動することが可能である．別のクラスタに移動した場合でも，同様にそのクラスタに含まれる論文の情報を確認することができる．

論文のノードには論文のタイトルを表示しており，マウスオーバーで抄録が表示される．探索支援の機能として，探索経路の保存や，気になる論文をストックしておけるエリアを用意している．

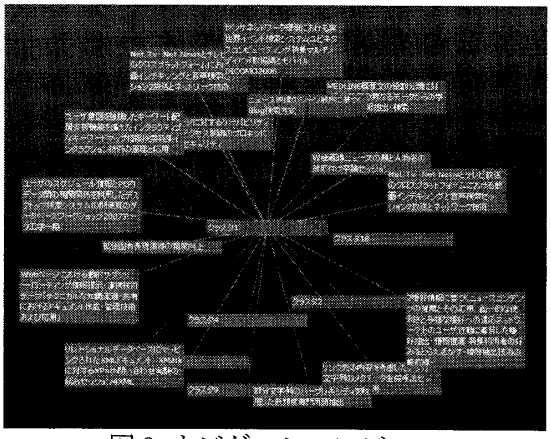

図2. ナビゲーションビュー

4．おわりに

本稿では，研究者が効率的かつ俯瞰的に論文検索を行うためのナビゲーションシステムの実装を行った．今後の課題として，実際にユーザに利用してもらい評価実験を実施したい．

参考文献

[1] 鈴木 雅人．リッチインターフェースを備えたグラフィカル論文検索支援システム，情報処理学会研究報告ヒューマンコンピュータインタラクション研究会報告，pp. 87-91，2008
[2] 小池諭，三末和男，田中，二郎．書誌情報ネットワークのビュー操作に基づく文献サーベイ支援，第六回知識創造支援システム・シンポジウム報告書，日本創造学会，pp. 220-227，2008
[3] bayon http://code.google.com/p/bayon/ 2010 年 1 月 15 日
[4] SpringGraph Flex Component http://mark-shepherd.com/blog/springgraph-flex-component/ 2010 年 1 月 15 日

Copyright © 2010 Information Processing Society of Japan.
All Rights Reserved.