ネットワーク分析にもとづく複数形式文書分類手法の提案

矢部 大輔† 木村 昌臣‡

芝浦工業大学大学院工学研究科† 芝浦工業大学工学部情報工学科‡

1. 研究背景

近年，情報ライフサイクル管理(以下 ILM)が注目されている.ILM は様々な形式の電子文書(以下文書)やデジタルデータのデータ作成から廃棄までを効率的に管理する手法であり，膨大な数の様々な形式の文書を管理する場合に有用である.文書を管理するためには類似の特徴を持った文書群に対し，管理ポリシーを適用する.そのためには，それらの文書を分類する方法が必要になる.しかし，これまでテキストマイニング手法等で提案された既存の分類手法の多くは共通の形式を持つ文書を対象としているため，形式の異なる文書群へ適用することが容易ではない.

そこで本研究では，文書の形式が異なっていても関連する内容をもつものは同じグループに分類するべきであると考え，文書の内容をもとにした文書分類手法を提案し，その有効性を検討する.

2. 提案手法

2.1 提案手法の概要

文書をノード(以下文書ノード)として考え，文書間の関連性を共通のキーワードを含むか否かで表し，関連性を有する場合にエッジを張ることでグラフ構造の構築を行う.次にネットワーク分析手法のクラスタリングを適用することにより文書を分類を行う手法を提案する.利用者の意思に反する分類であれば利用価値がなくなり意味がないので，利用者の意図を反映させることが必要となってくる.そこで分類先のディレクトリをノードの 1 つ(以下ディレクトリノード)としてグラフ構造に追加してクラスタリングを行う(図 1).

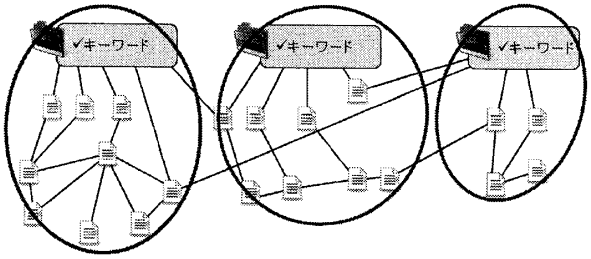

図 1.文書とディレクトリをノードとしたグラフ

このときディレクトリノードにキーワードを設定することで文書ノードとの関連性を見ること

A proposal for classification of the various format of documents using network analysis

† Daisuke Yabe, ‡ Masaomi Kimura,

† Graduate School of Shibaura Institute of Technology

‡ Shibaura Institute of Technology

ができるようになり、かつ利用者の意図を汲んだ分類が可能になる.

2.2 グラフ構造の構築

文書内容をもとにしたグラフ構造の構築は，文書内に設定されたキーワードに共通のものがある場合に関連性があると判断して文書間にエッジを張ることにより行う.共通のキーワードを持つか否かで関連性の有無を決めるが，その際どの単語をキーワードとするかが問題となる.重要性が高い言葉は文章では特徴的な使われ方をしていると仮定し，TFIDF 法を用いて名詞を抜き出しキーワードとして分類する全ての文書に付与する.文書の文字列を見て関連性の有無を決めるため,形式を問わず様々な文書に対して適用可能である.文書に含まれる文字列を文書からテキスト要素を取得する xdoc2txt[1]を用いて抽出し，形態素解析器である ChaSen [2]を用いて形態素解析を行い名詞のみを取り出す.取り出した名詞に対して TFIDF 値をそれぞれ計算し，値の高い名詞をその文書にキーワードとして付与する．共通のキーワードを一定数(共起回数)以上含む文書間には関連があると判断し，文書間にエッジを張る.しかし，文書内容に関係ない名詞がキーワードになって共起され；る可能性があるので共起回数はグラフ構造に接続されないノードがでない数のうち最大のものをとる.ただし，ディレクトリノードに関してはより多くのエッジを接続してハブの役割をさせることにより利用者の設定したキーワードが反映されたクラスタになりやすくなるので文書ノード間の共起回数はなるべく少なくする.この作業を全ての文書間で行うことによりネットワークを作成することができる.

2.3 クラスタリング

統計解析ソフト R のパッケージ igraph に含まれている spinglass.community 関数を利用してクラスタリングを行う.エッジでつながっているノードは同じクラスタへ，つながっていないノード同士は違うクラスタへ分類されると評価が良くなるよう設計した評価関数を物理学におけるスピングラス系の Potts モデルハミルトンと定義し，SA による最適化を行う手法である.

3. 実験

本研究室で毎週行う進捗報告の資料や発表資

料等の文書を用いて提案手法を利用した分類を実際に行う.対象の文書の形式は Microsoft の Office の Word292 個, PowerPoint37 個である.

用いる文書データ数は 329 個, ディレクトリノードを 3 個とし,合計のノード数は 332 個である. 文書データは主に 3 つの研究分野に分かれているのでディレクトリノードを 3 個とし, 関連する名詞をキーワードとして設定した.キーワードが少ないと共起が起きにくいため一文書あたりのキーワード数を 50 とし, キーワードが 5 回共起したときにエッジを張る.ただし, ディレクトリノードに設定されたキーワードの共起回数は 1 回とする. 事前に類似しているドキュメント群を作成しておき, これを正答として実験結果のクラスタに内容が類似している文書がどれだけの割合で入っているかを調べる.

4. 結果

構築したグラフ構造から隣接行列を作成してクラスタリングを行った結果, 6 個のクラスタが作成された.クラスタのノード数と正しい分類となっているかを調べ, 表 1 にまとめた.

表 1. 実験結果

	24	72	55	24	90	67
	17	68	54	21	50	44
	70.8	94.4	98.2	87.5	55.5	65.6

実験結果よりクラスタ 2,3,4 は高い確率で内容が類似した分類であったが他のクラスタには誤った文書が混じる結果となった.このため誤った分類になってしまった文書を調査したところ, 次のいくつかの特徴があることが分かった.

1. 文書に文字がほとんど含まれていない
2. フォーマットが似ている(目次, 日時など)
3. 文書内容が他のクラスタの文書内容と類似

1 の場合は文書内容のほとんどが画像であるものや作りかけの文書であった.2 の場合は目次や日時, 所属などの部分がキーワードとして共起してしまったことが原因であると考えられる.3 の場合は文書内容が別のクラスタに属する文書と類似している部分があったためだと考えられる.

ディレクトリノードはクラスタ 2,5,6 に含まれ, キーワードと近い内容のクラスタに属していた.ディレクトリノードの含まれていないクラスタ 1, 3, 4 はそれぞれ 6,2,6 に内容が類似していた.類似した文書であるが, 複数のクラスタに分かれる場合があるためだと考えられる.そこでクラスタ間のノードの何本接続本数を図 2 に表した.

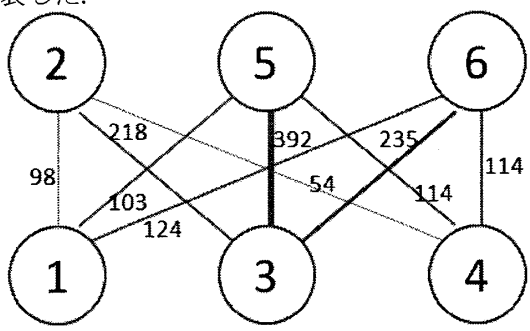

図 2 クラスタ間のノード数

クラスタの含包はノードの本数に関連が見られる.しかし, クラスタ 5 と 3 のノード数が非常に多いため内容を調べたところ異なる研究分野ではあるが同じ分析手法を利用していた.

5. 課題

本研究で提案している手法では文書内容が類似しているものをエッジで結ぶことでグラフ構造を構築しているため, 内容が類似しているかどうかという切り口のみでの分類になっている.このため他の基準による分類を実現しようと考えた時上記で提案した手法にどのようにとり入れるかは, その基準がいかにネットワーク構造に反映できるかによる.グラフ構造は分類結果に非常に大きな影響をもたらすので, その構築方法次第で様々な種類の分類に対応できるようになる可能性がある.今後はどういった分類に対し, どのようなグラフ構築方法をとるのが最適であるか調査する必要がある.

6. まとめ

本研究では、文書内容より作成したグラフ構造をネットワーク分析することで文書を分類する手法を提案し, 複数の形式の文書を混ぜて実験を行い, 内容にもとづいた分類が可能であることを示した.

今後はグラフ構造の構築方法を改良することで様々な分類への対応を行いたいと考えている.

参考文献

1) xdoc2txt, http://www31.ocn.ne.jp/~h_ishida/xdoc2txt.html
2) 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸: 日本語形態素解析システム『茶筌』version 2.3.3 使用説明書, 奈良先端科学技術大学院大学(2003)
3) 湯浅夏樹, 上田徹, 外川文雄:大量文書データ中の単語間共起を利用した文書分類, 情報処理学会論文誌, Vol.36, No.8, pp.1819-1827 (1995)