ドキュメントの関連性を考慮したタグ付け支援手法の提案

A Proposal of Tagging System Based on Relevance of Documents

酒井修太郎　　奥野拓
Shutaro Sakai　Taku Okuno

はこだて未来大
Future University-Hakodate

1. はじめに

プロジェクト管理を行うにあたって，ドキュメント管理は重要である．プロジェクトの進行とともにドキュメントの数が増加することで，その量は膨大なものとなる．この中から迅速に目的のファイルを探し出すためにはドキュメントを分類する必要がある．

タグ付けは分類手法の一つである．本研究におけるタグはドキュメント自体の付加情報を表すメタデータとして定義する．タグ付けされたドキュメントにはその中身を推測できるだけでなく，付与されたタグから関連したドキュメントを探し出すことができるという利点がある．

本研究では，タグ付けを支援する手法を提案することで，タグ付けにおける問題点を解決し，よりドキュメント管理を便利にすることを目的とする．

2. タグ付けにおける問題点

タグを付与する場合の主な問題として，以下に示す 2 点が指摘されている[1].

2.1 同義語の問題

タグを付与する場合，ユーザはタグ名を自由に決定することができるため，タグ名は異なるが同じ意味であるタグが大量に作成されてしまう可能性がある．このようなタグは可能な限り一つの語句として統一することが理想である．同義語の問題の一例を図 1 に示す.

2.2 タグの分類抽象度の問題

ユーザは誰もが異なった感性を持っているため，様々な抽象度のタグが作成される可能性がある．このような抽象度が異なるタグについても可能な限り一つのタグとして統一することが望ましい．タグの分類抽象度の問題の一例を図 2 に示す.

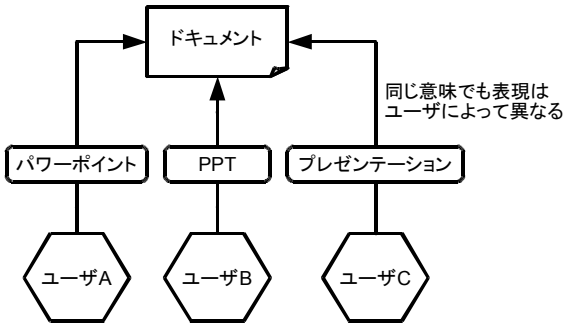

図 1　同義語の問題例

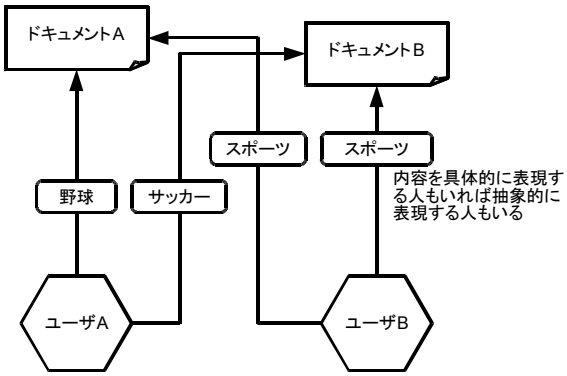

図 2　タグの分類抽象度の問題例

タグを用いて関連したドキュメントを探す場合にこの 2 つの問題が発生することで，関連したすべてのドキュメントを参照することができなくなる可能性がある．このような問題を解決し，かつタグとして十分な情報が付与できることが必要とされる.

3. 関連研究

前節で挙げた問題点に対して，タグを特徴ごとにまとめることによって解決する手法が提案されている[1]．ある 2 つのタグが共起する確率を算出することで，その確率分布に見られる特徴から同じ内容を意味するタグや抽象度の異なるタグを探して分類することで同種のタグとしてまとめている.

また，タグを階層的に配置することで問題を解決する手法も提案されている[2]．この手法ではタグの共起関係に基づいてタグの特徴ベクトルを算出し，その距離を求めてタグの上位下位関係を決定することで関連したタグの検索を容易にしている.

しかし，これらの手法はすでに登録されたタグに対して行っているものであるため，ドキュメントに対してタグを付与する場合に適用するには不都合である．したがって，タグを付与する時点で前節の問題点を解決するためには別の手法を用いる必要がある.

4. 提案手法

4.1 アプローチと概要

ドキュメントの管理において 2 節で挙げられた問題点を解決し，かつユーザが判断に迷うことなくタグの付与を行うことができるようにするための一つのアプローチと

して，ドキュメントに付与するタグの候補をユーザに提示する方法が考えられる．

このアプローチに関連する技術として，パターンマッチングと形態素解析を用いてテキスト中から組織名や人物名，製品名などのキーワードを抽出できる手法が提案されている[3]．しかし，この手法を用いる場合，抽出されたキーワードをそのままタグとして利用するだけでは非常に多くの語句が出現し，その中で意味の似た語句も大量に出現することが予想される．このため，何らかの方法で抽出するキーワードを限定することが必要である．

キーワードの限定を行う方法については，文書内の総単語数や特定の語句の出現数からその語句の重要度を求め，その結果から同じ語句が含まれる文書の類似度を求める手法が提案されている[4]．また，キーワードそのものをタグにするのではなく，すでに登録されているタグの中からそのキーワードに近いタグ名のタグを探し出してユーザへの提示を行う．この方法を用いることで，意味が似ているキーワードからは同じタグの提示を行うことができるため，意味の似た複数のタグが作成される問題を解決することができる．これらの点をふまえて本研究で提案する手法を以下で述べる．

4.2 タグ付け手法の詳細

本研究で提案する手法を図 3 に示す．

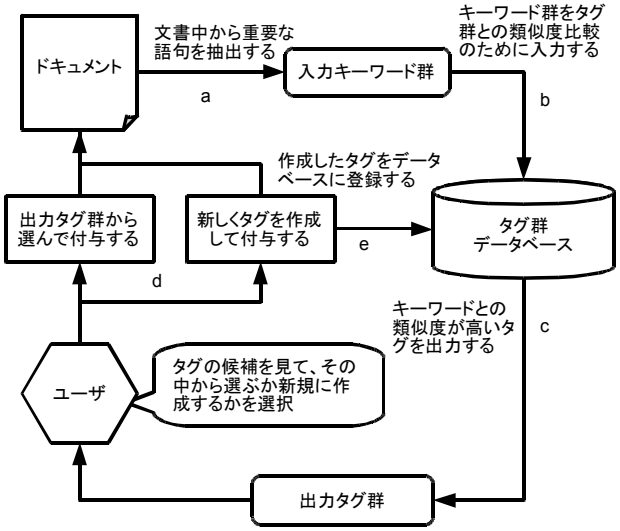

図 3　本研究で提案する手法の流れ

まず，本手法を実現するためにプロジェクトで頻出する語句がタグとして登録されているデータベースを用意する．以下にタグの付与を行う手順を示す．

a. タグ付けするドキュメントからキーワード群を抽出し，それらの重要度を求め，重要度の高いものを入力キーワード群とする．
b. 入力キーワード群をもとにタグ群が蓄積されたデータベースに問い合わせを行う．
c. 入力キーワード群と蓄積されたタグ群の類似度を算出し，類似度が高いタグを出力タグ群としてユーザに提示する．
d. ユーザは出力タグ群の中からドキュメントに付与したいタグを選択することでタグの付与が完了する．また，出力された候補以外に付与したいタグがある場合は新規にタグを作成して付与する．
e. 新規にタグを作成した場合，そのタグはタグ群データベースに追加登録する．

5. まとめと今後の方針

本研究ではタグ付けを支援するために，ドキュメントからキーワードを抽出し，そのキーワードとデータベース内のタグとの類似度を算出し，類似度が高いタグをユーザに提示することで，ユーザのタグ付けを支援する手法を提案した．今後の課題としては以下の項目が挙げられる．

5.1 キーワード群の抽出方法の検討

形態素解析を用いることでドキュメント内の文を形態素の列に分割することができる．しかし，その形態素の中のどの語句をキーワードとして扱うべきかを決定する必要がある．例として品詞に注目する場合でも，名詞をキーワードとして用いれば十分であるのか，他の品詞もキーワードの候補にする必要性があるかなどについて検討が必要である．

5.2 キーワードとタグの類似度を求める手法の検討

3 節で語句の重要度から文書の類似度を求める手法を挙げたが，提案手法に含まれている類似度の比較は単語同士のものとなるため，この手法をそのまま応用することは難しく，類似度の比較に関する更なる検討が必要である．

5.3 タグ群データベースの初期登録状態の決定方法

本研究で提案した手法を実現するためにはタグ群が登録されたデータベースが必要になるが，そのデータベースに初期状態として登録する語句を決定する必要がある．ソフトウェア開発プロジェクトのようにドキュメントの管理を頻繁に行う活動ではどのような語句が頻出するかを調査し，どの程度の数のタグが初期状態として蓄積されていれば十分であるかを検討しなければならない．

参考文献

[1] 丹羽智史，土肥拓生，本位田真一，“Folksonomy の 3 部グラフ構造を利用したタグクラスタリング”，セマンティックウェブとオントロジー研究会，2006.
[2] 江田毅晴，吉川正俊，山室雅司，“非巡回有向グラフによるフォークソノミータグの局所拡張可能な配置方法”，DEWS，2008.
[3] 松平正樹，大沼宏行，上田俊夫，渕上正睦，森田幸伯，“文書中のキーワードに関する多種多様な情報を収集･整理するシステム～システムの概要と固有表現抽出技術，オントロジー技術～”，沖テクニカルレビュー第 200 号，2004.
[4] 長沼潔，速水悟，“医療分野における Web 文書からの話題抽出方法”，人工知能学会第 19 回全国大会，2005.