D-020

内容の同一性を考慮した類似ユーザ群の Web ページタギングに基づく意味情報抽出

Ontology Extraction Considering Content Concordance from Web Pages Tagged by Similar Users

伊藤 真也 ‡ Masaya Ito　原田 史子 † Fumiko Harada　島川 博光 † Hiromitsu Shimakawa

1. はじめに

一般的な検索エンジンは，ユーザが入力するキーワードと Web ページに含まれる文字列の一致に基づいて検索する．しかしながら，入力されたキーワードのユーザにとっての使い方を一般的な検索エンジンは考慮していない．ユーザは，同一のキーワードであっても辞書に含まれない暗黙的な意味を付与している場合がある．また，同一のキーワードであっても，ユーザによってその指し示す範囲は異なることがある．例えば，“情報推薦”というキーワードを考える．“情報推薦”という言葉は一般的には，検索エンジンなどの pull 型システムやポップアップ広告などの push 型システムを含む．ここで，ユーザ A が “情報推薦”というキーワードに “所属研究チームの研究内容”という意味を暗黙的に付与していたとする．検索エンジンは，ユーザ A の所属チームについての情報を優先的に検索結果として提示できない．またユーザ B が “情報推薦”というキーワードを pull 型システムのみを指し示す言葉として使用していたとする．検索エンジンは，ユーザ B の望まない push 型システムについてのページも提示してしまう．検索エンジンが個人の言葉の使い方を考慮できれば，よりユーザに有益な情報を提示できる．

個人の言葉の使い方を考慮した検索エンジンを作るためには，まず個人の言葉の使い方を把握し，計算機で処理できる形式で表現する必要がある．表現方法として知識や用語を体系化するのに用いられるオントロジが適用できる．本論文では，階層関係と類義関係を定義したオントロジを言葉の使い方を表現する形式として用いる．階層関係とは，言葉同士の上位と下位を定義したものである．類義関係は，意味が似通っている言葉同士がもつ関係である．個人の言葉の使い方を表現するためには，言葉の階層関係と類義関係を表現するオントロジが必要である．本論文では，類似する嗜好のユーザ群内における，内容が類似した Web ページ群へのタグ付与から個人の言葉の階層関係と類義関係を抽出する手法を提案する．

2. 研究背景

個人の言葉の使い方を抽出する方法として，ソーシャルブックマーク (SBM) の利用が考えられる．SBM では，ユーザは，自身のブックマークの整理のために，ブックマークに対して自由記述のタグを付与する．

言葉の使い方は，階層関係と類義関係の 2 つの関係を定義することで，表現できる．個人のタグ付与にのみに，着目することで，個人の言葉の階層関係を抽出することができる．一方，類義関係は，個人が付与したタグ群には存在し難い．これは，タグがそのユーザのブックマークの整理のために付与されるためである．そのため，嗜好が似たユーザ群から，言葉の類義関係を抽出することを考える．嗜好が似たユーザ群は，例えば，文献 [1] で提案されている方法を用いて SBM から抽出できる．

著者らは，文献 [2] において，あるタグが付与された Web ページの集合にもとづいて，個人の言葉の使い方を抽出する手法を提案した．この手法では，あるタグの名前として使われている言葉の意味は，そのタグがもつ Web ページの集合にあると考えている．この手法の問題点として，ユーザのタグの付け忘れが精度に影響する点が挙げられる．ユーザがタグを付け忘れた場合，そのタグがもつ Web ページ集合は，本来そのタグが持つ集合より小さいものになる．これは，そのタグの名前として使われている言葉の意味が狭義になることを意味する．そのため，この手法は，ユーザのタグの付け忘れがあった場合，正確に言葉の使い方を抽出できない．SBM から正確に個人の言葉の使い方を抽出するためには，ユーザのタグの付け忘れを考慮した手法が必要である．

3. タグに基づく特定ユーザの意味情報の抽出

3.1 同一性を考慮した特定ユーザの意味情報の抽出

正確な個人の言葉の使い方を抽出するために，類似ユーザ群内のブックマークの内容とタグの関係から個人の言葉の使い方を抽出する手法を提案する．これによりタグ付け忘れの精度への影響を低減することができる．例えば，ユーザがブックマークに対して，あるタグを付け忘れたとする．内容とタグの関係から抽出する場合，他の同一内容のブックマークに対して，そのタグが付与されていれば，そのタグの名前として使われている言葉の意味を正確に抽出できる．内容とタグの関係から言葉の使い方を抽出するためには，ブックマークを内容にもとづいて分類する必要がある．分類のさいに，同一のグループに分類された Web ページ群を，内容の同一性を

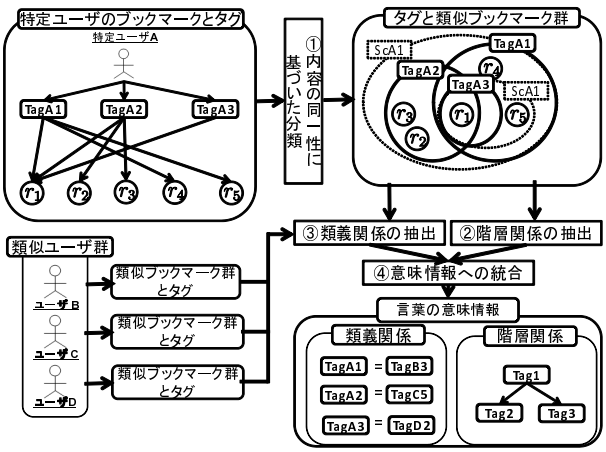

図 1: 提案手法全体像

†立命館大学情報理工学部
‡立命館大学大学院理工学研究科

持った Web ページ群と定義する．本論文では，ブックマークを内容にもとづいて分類し，内容とタグの関係から言葉の使い方を抽出する手法を提案する．

本手法は，SBM においてユーザがブックマークに付与したタグから，特定ユーザの言葉の意味情報を抽出する．図 1 に，提案手法の全体像を示す．手法の手順は，(1) 類似ブックマーク群の作成，(2) タグ間の階層関係の抽出，(3) タグ間の類義関係の抽出，(4) 言葉の意味情報への統合，となる．

3.2 類似ブックマーク群の作成

類似ユーザ群内のブックマーク群を内容の同一性にもとづいて分類する．ここで，同一の分類とされたブックマーク群を類似ブックマーク群と定義する．本節では，類似ブックマーク群の作成手法について提案する．本手法では，階層的クラスタリング手法とユーザが Web ページへ付与するタグを利用し，類似ブックマーク群を作成する．階層的クラスタリング手法として，例えば，文献 [3] の手法が適用できる．類似ユーザ群が，計 l 件の Web ページ $\{r_1, r_2, ..., r_l\}$ をブックマークしている．ユーザ u_i が付与したタグ群を $\{T_{i1}, T_{i2}, ..., T_{im_i}\}$ とする．ここで、ある T_{ix} の名前として，使われている文字列をタグラベル $L_i(T_{ix})$ とおく．T_{ix} が付与された Web ページ群を，タグクラスタ $C_i(T_{ix})$ とおく．階層的クラスタリング手法をブックマーク群 $\{r_1, r_2, ..., r_l\}$ に適用したさいに，生成された a 個の集合を，$\{G_1, G_2, ..., G_a\}$ とする．

特定ユーザが付与したタグ T_{ix} のタグクラスタ $C_i(T_{ix})$ から類似ブックマーク群を作成する場合を考える．まず，$C_i(T_{ix})$ と G_y を比較する．2 つの集合の類似度が高い場合，$C_i(T_{ix}) \cup G_y$ を類似ブックマーク群 $SC_i(T_{ix})$ として作成する．類似度が高い否かの判定では，式 (1) を用いる．BT は閾値である．

$$\frac{|\ C_i(T_{ix}) \cap G_y)\ |}{|\ C_i(T_{ix}) \cup G_y)\ |} \geq BT \tag{1}$$

$C_i(T_{ix})$ が，a 個すべての集合において，式 (1) を満たさなかった場合，$C_i(T_{ix})$ を $SC_i(T_{ix})$ とする．

3.3 タグ間の階層関係の抽出

特定ユーザの付与した任意の 2 タグのラベルについて，階層関係を同定する．本手法では，$SC_i(T_{ix})$ が $L_i(T_{ix})$ の意味を表すと考える．そのため，あるタグの $SC_i(T_{ix})$ が比較的大きければ，そのタグラベルは比較的広義な意味を持つと考える．あるタグの $SC_i(T_{ix})$ が比較的小さければ，そのタグラベルは比較的狭義な意味を持つと考える．ある 2 タグの $SC_i(T_{ix})$ と $SC_i(T_{ix})$ が，共通の要素を持てば，その 2 タグのタグラベルは，互いに関連している．階層関係にある 2 語は，一方が比較的広義な意味を持ち，他方が比較的狭義な意味を持つ．さらに，これらの 2 語は，互いに関連のある言葉である．そのため，本手法では，$SC_i(T_{ix}) \subset SC_i(T_{iy})$ が成り立つとき，タグ T_{ix} をタグ T_{iy} の上位の言葉として関連づける．

3.4 タグ間の類義関係の抽出

類義関係抽出の手法においても，$SC_i(T_{ix})$ が $L_i(T_{ix_i})$ の意味を表すと考える．類義関係にある 2 語は，互いに似通った言葉の意味を持つ．そのため，本手法では，2 タグの $SC_i(T_{ix})$ と $SC_j(T_{jy})$ が似通っていれば，その 2 タグのタグラベルは，類義関係にあると考える．

類似ユーザ群を S とおく．特定ユーザ u_i が付与したタグと $S - \{u_i\}$ の各ユーザの付与したタグを比較し，タグのラベル間の類義関係を抽出する．特定ユーザ u_i と $u_j \in S - \{u_i\}$ が付与したタグ群から，類義関係を抽出する場合を考える．まず，ユーザ u_i とユーザ u_j の持つすべての 2 タグの組み合わせを生成する．ユーザ u_i が m_i 個，u_j が m_j 個のタグをそれぞれ使用したとき，u_i が使用したタグを $\{T_{i1}, T_{i2}, ..., T_{im_i}\}$，$u_j$ が使用したタグを $\{T_{j1}, T_{j2}, ..., T_{jm_j}\}$ とおく．次に，特定ユーザ u_i とユーザ u_j がそれぞれもつタグの組み合わせ $(T_{i1}, T_{j1}), (T_{i2}, T_{j1}), ..., (T_{i1}, T_{j2}), (T_{i1}, T_{j3}), ..., (T_{im_i}, T_{jm_j})$ を生成する．各 (T_{ix}, T_{jy}) に対して $SC_i(T_{ix})$ と $SC_j(T_{jy})$ が似通っている場合，それらのタグのラベルを類義語とみなす．タグ T_{ix} とタグ T_{jy} が類義関係か否かの判断は式 (2) および (3) の指標を用いる．ここで，ユーザ u_i のブックマークしている Web ページ数を N_{u_i}，類義語かどうかの判定のさいに使用する閾値をそれぞれ θ_1, θ_2 と定義する．

$$\frac{|\ SC_i(T_{ix}) \cap SC_j(T_{jy})\ |}{|\ SC_i(T_{ix}) \cup SC_j(T_{jy})\ |} \geq \theta_1 \tag{2}$$

$$\frac{|\ SC_i(T_{ix}) \cap SC_j(T_{jy})\ |}{N_{u_i}} \geq \theta_2 \tag{3}$$

4. おわりに

本論文では，個人の言葉の使い方を考慮した検索エンジンを実現するために，類似ユーザ群における Web ページへのタグ付与から，Web ページの内容の同一性を考慮して，個人の言葉の意味情報を抽出する手法を提案した．本手法では，内容の同一性を考慮するために，類似ブックマーク群とタグの関係から，言葉の意味情報を抽出する．類似ブックマーク群は，階層的クラスタリング手法とユーザが付与したタグを利用して，類似ブックマーク群を作成する．内容の同一性を考慮することで，正確な言葉の意味情報の抽出が期待できる．今後は，提案手法を実現するシステムを実装し，有用性を検証する．

参考文献

[1] 矢島健太郎，井上潮: ソーシャルブックマークにおける文書解析を利用した類似文章および類似ユーザの推薦方法の提案，第 18 回データ工学ワークショップ，第 5 回日本データベース学会 年次大会，C9-3，2007 年 3 月．

[2] 伊藤真也，小河真之，原田史子，島川博光，Web ページへのタグ付けによる類似ユーザ群を利用した意味情報の抽出，電子情報通信学会/情報処理学会 情報科学技術レターズ，Vol.9, No.2, pp.21-24, Sep., 2010.

[3] 折原大，内海彰：HTML タグを用いた Web ページのクラスタリング手法，情報処理学会論文誌，一般社団法人情報処理学会，Vol.49，No.8，pp.2910–2921，2008 年 8 月．