E-012

ニュース検索のための格構造を用いたユーザの興味表現と分類手法

A Representation and Clustering of User's Interests by Using Case Structures for Retrieval of News Documents

東原　智幸†　　渥美　雅保†
Tomoyuki Higashihara　　Masayasu Atsumi

1. まえがき

インターネット上の文書数の増大によりユーザ自身の求める文書を検索することが難しくなっている．近年，その問題を解決するため，ユーザが興味を示す文書をシステムが自動的に検索し，提示する研究が多く行われている[1][2]．また，ブログなどの文書から話題(トピック)を抽出するサービスも提供されている．それに伴って，ブログからのトピック抽出について研究されており，[3]では文書ベクトルに基づいて，クラスタリングを行い，得られたクラスタをトピックとしている．その際，興味や文書の表現としてtf・idf法が多く用いられているが，文の係り受け構造や格情報などの情報が利用できない点に問題がある．本研究では，格構造の集合からユーザの興味概念構造を表現し，それらをニュース検索に適用する方法について提案する．ユーザがニュースに興味を示した場合，そのニュースのすべてを興味として保持することは興味とは関係のないデータもふくまれるため，検索率の減少の原因となる．本システムでは，ユーザが興味を示したニュースに対して，展望台システム[4]を適用しニュースの主題を選択する．その結果からトピック構造集合を抽出する．次に，その集合をクラスタリングし，ユーザの興味概念構造を作成する．

最後に，興味概念構造を使用した検索実験を行い，ユーザの興味に適応した検索が行えるか評価を行う．

2. 興味構造

興味構造はトピック構造集合をクラスタリングすることにより構築する．以下でこれらについて説明を行う．

2.1. トピック構造

トピック構造は，ニュース中の主題と推測される文章内に含まれる格構造集合である．

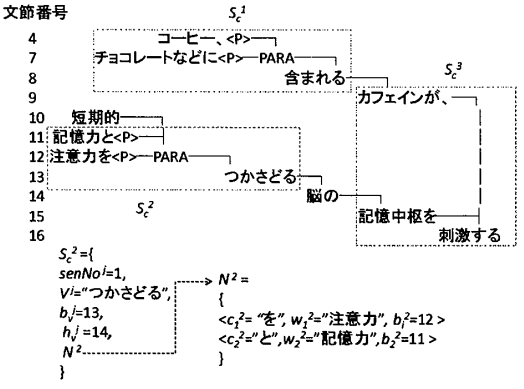

図 1　構文解析結果と格構造(一部略)

ニュース中のj番目のトピック構造S_c^j(図 1)は

$$S_c^j=\{senNo^j, V^j, b_v^j, h_v^j, N^j\} \quad (1)$$

と表現される．$senNo^j$は，ニュース中の文番号，V^jは動詞，b_v^jは動詞の文節番号，h_v^jは動詞の係り先文節番号である．

N^jは動詞V^jに係るM個の名詞集合で，表層格c_m^j，その格の要素(単語または複合語)w_m^j，その要素の文節番号b_m^jの組の集合からなり，

$$N^j=\{<c_1^j,w_1^j,b_1^j>,\ldots,<c_m^j,w_m^j,b_m^j>,\ldots,<c_M^j,w_M^j,b_M^j>\} \quad (2)$$

と表される．

例えば，ニュース中の 1 番目の文章に"カフェインが記憶中枢を刺激する"が存在する場合は，格構造は
{1,"刺激する",16,-1,{<"が", "カフェイン",9>,<"を","記憶中枢",15>}}と表現される．動詞の係る文節がない場合には，係り先文節番号は-1 となる．

2.2 トピック構造のクラスタリング

クラスタリングでは，トピック構造類似度と単語間の概念類似度を用いる．

2.2.1 トピック構造間類似度

トピック構造間類似度$s(x, y)$はトピック構造$x=S_c^i$と$y=S_c^j$間の概念的，格構造的な類似度を求める式で，以下の式により定義される．

$$s(x,y)=sim(V^i,V^j)+avr(N^i,N^j)$$

$sim(V_i, V_j)$は動詞の概念類似度と$avr(N_i,N_j)$は名詞集合の概念類似度平均である．名詞集合間の概念類似度平均は，

$$avr(N_i,N_j)=\frac{1}{K}\sum_{s.t.\ c_k^i=c_k^j} sim(w_k^i,w_k^j)$$

で表され，表層格の同じ単語同士の概念類似度の和の平均を求めたものである．K は表層格の同じ単語の組数である．

2.2.2 単語間の概念類似度

単語間の概念類似度は，概念体系辞書[5]を用いて概念の深さを検索し，その値を用いて類似度を計算する．
単語w_0, w_1の概念体系の深さをd_0, d_1, 共通の深さをd_cとするとき，単語間の類似度は

$$sim(w_0,w_1)=\begin{cases}\dfrac{2\times d_c}{d_0+d_1} & \cdots w_0,\ w_1\text{が登録されている}\\ 0 & \cdots\text{それ以外}\end{cases} \quad (3)$$

で表される．w_0, w_1が辞書には登録されていないが，形態素要素に分割できる場合には，各形態素要素毎に類似度を計算し，それらの最大値を$sim(w_0,w_1)$とする．

2.2.3　クラスタリング

クラスタリングは抽出されたトピック構造集合 *TpSet* に対して実行される．トピック構造集合 *TpSet* は，主題文 1 文ごとに形態素解析[6]・構文解析[7]を行い抽出される．

$$TpSet=\{S_c^1,\ldots,S_c^j,\ldots,S_c^J\} \quad (4)$$

ここで，Jはトピック構造の総数である．S_c^jの名詞集合N^jには，V^jに直接係っている名詞w_i^j，その名詞と並列(PARA)関係または同格関係である名詞が含まれる．同格関係とは“長男太郎”などのように複数の分節が対等で同一の対象を示している関係である．
抽出した *TpSet* を用いて階層併合的クラスタリング[10]によりクラスタリングを行う．基本的なアルゴリズムは以下の通りである．ここで，G はクラスタ集合，C はクラスタ数，index はクラスタ番号である．

1．$S_c^i \in TpSet$をクラスタ$G_i=\{S_c^i\}$として割り付ける．C=J, index=C+1 と設定する
2．各データの類似度を以下の式にて計算する

$$s(G_i,G_j)=\min_{x\in G_i,\,y\in G_j} s(x,y)$$

3．類似度最大のクラスタ対G_q, G_rを結合する．Gに$G_{index}=G_q$

†創価大学大学院工学研究科情報システム工学専攻

$\cup G_r$を追加し，G_q，G_rをGから除く．C=C-1 と更新する．

G_{index}を作成する際に，G_q，G_r間共通の概念を求め，G_{index}の概念とする．G_q，G_r間の動詞の組，表層格が同じ名詞集合の組でそれぞれ共通概念を求める．共通概念は，概念体系の共通の深さd_cにある概念である．

4．もし，C=1 ならば終了．C>1 ならばi<indexなるすべてのG_iについて$s(G_{index}, G_i)$を計算後，index=index+1 と変更し，ステップ 3 に戻る．

図 2はクラスタリングされた結果を樹状図で表現した図である．クラスタに概念 ID(動詞のみ表示)，クラスタ番号，クラスタ間の類似度が付与されていることが分かる．

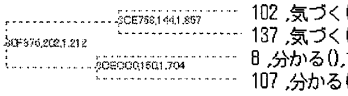

図 2　樹状図(一部)

2.3 興味概念構造による検索

上位のクラスタに与えられた概念を検索に用いることにより概念的に広い情報を検索することができる．例えば，図 2の{"気づく", <"に","異臭",>,<"が","妻">}と{"気づく", <"が","女性経営者",>,<"で","上">}では上位クラスタの概念は{"気づく",<"が",女性という概念>}となり，より広い“女性”という概念に近い主語をもつ“気づく”にも興味があると考えることが可能となる．

3.ニュース検索実験

3.1 テスト環境

本実験では，ニュース提供システム(図 3)[8]において，実験を行った．ニュース提供システムは，ユーザの要求を処理する音声対話部，ユーザの興味概念構造を管理する興味構造管理部，ユーザに提供するニュースを検索するニュース選択部から構成される．興味構造管理部とニュース選択部の一部を実装している．

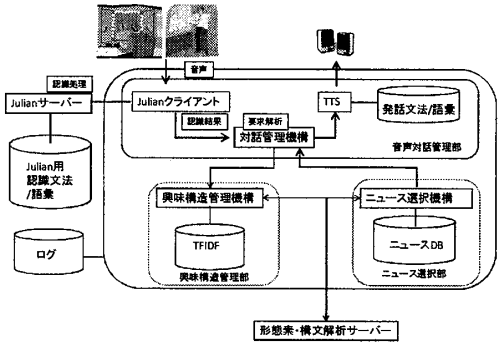

図 3 システム構造

3.2 評価データの作成

以下の手順でユーザのニュースに対する評価を収集し，ニュース検索性能に対する評価データを作成する．使用するニュースは，毎日新聞 2007 年版[9]の社会面，1 月 1 日～17 日のニュース 500 件(2 日は 0 件)である．

(1)ニュースをユーザに提示

(2)提示されたニュースに対して A（興味がある），B(どちらかといえば興味がある)，C(どちらかといえば興味はない)，D（興味はない）の 4 段階で評価する．

3.2 評価結果

興味概念構造の深さの違いが検索性能に及ぼす影響について評価を行った．興味概念構造の深さは，葉の部分が 0，根に向かうにつれて大きくなっていく．深さを D とし，D=(0，1，2，5)に変更し，それぞれの深さに対する検索結果(図 4)を調査した．図 4は，上位 N 位(N=1～20)まで検索した場合の精度を示している．評価データを 5 つに分割し，1 つ選択し，興味概念構造を作成，残りの 4 つを評価用データとして使用した．20 位における精度値は，D=0 で 0.663，D=1 で 0.630，D=2 で 0.640，D=5 で 0.615 となっている．5 位以降では，D=0 で 18 位 0.681，D=1 で 9 位 0.672，D=2 で 17 位 0.653，D=5 では 7 位 0.657 が最大であった．また，20 位までの平均では，D=0 で 0.666，D=1 で 0.652，D=2 で 0.633，D=5 で 0.634 であった．深さが大きくなっても精度に大きな違いは見られなかった．深くなれば検索に用いるクラスタ数も少なくなるが，クラスタは下位のクラスタの概念も含んでいるため検索基準数の影響が少なかったと考えられる．

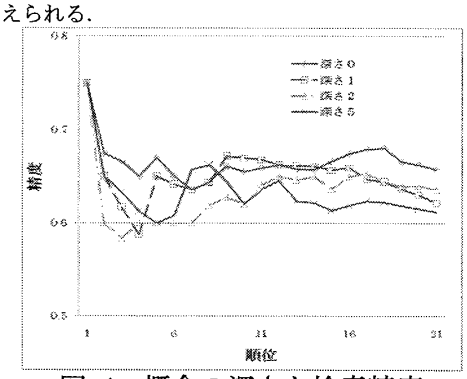

図 4　概念の深さと検索精度

表 1は，興味数評価データごとの興味総数の順位において検索された興味ニュースの割合の表である．

表 1　興味の割合

データ番号	興味総数(順位)	最小値(D)	最大値(D)
データ 1	55	0.605(0)	0.618(2)
データ 2	43	0.564(5)	0.605(0)
データ 3	24	0.480(5)	0.542(0)
データ 4	42	0.464(2,5)	0.482(0)
データ 5	57	0.711(1,5)	0.754(2)

4.まとめ

ユーザが興味を示した格構造の集合からユーザの興味概念構造を表現し，ニュース検索において，興味概念構造の深さを基準とした検索について評価を行った．今後は，古い興味の削除や，興味構造の更新(追加)，トピック構造の文章中における重要度を考慮したクラスタリング方法について検討を行う．

参考文献

[1]河合,熊本,田中:印象と興味に基づくユーザ選好のモデル化手法の提案とニュースサイトへの応用,日本知能情報ファジィ学会誌，vol.18 No.2，pp.173-183，2006.

[2]野美山,紺谷,渡辺他:個人適応型情報検索システムー個人の興味を学習する階層記憶モデルとその協調的フィルタリングへの適用ー，情報学基礎研究会報告，Vol.96，No70，pp.49-56,1996.

[3]戸田,黒田,福田,石川:ブログにおける多視点からのトピック抽出手法の提案,DEWS2008(第19回データ工学ワークショップ)，B4-2,2008.

[4]相良,砂山,谷内田:サブトピックを考慮した重要文抽出による報知的要約生成，電子情報通信学会論文誌 D，Vol.J90-D，No.2，pp.427-440,2007

[5]日本電子化辞書研究所,EDR 電子化辞書(第 2 版)仕様説明書，TR2-006(改)，2001.

[6]黒橋:日本語形態素解析システム JUMAN ver. 5.1，東京大学大学院情報理工学系研究科,2005.

[7]黒橋:日本語構文解析システム KNP ver. 2.0，東京大学大学院情報理工学系研究科,2005.

[8]東原,三吉,渥美:スマートホームにおける音声ニュース提供システムアーキテクチャの構築，FIT2008(第 7 回情報科学技術フォーラム),E-025,p.193-p.195,2008

[9]CD-毎日新聞(データ集)2007 年版

[10] データマイニングの基礎,元田ほか,オーム社,2006