共起語に基づくトピックの経年変化を利用した情報推薦について

1R-3

渡邉倫　大囿忠親　伊藤孝行　新谷虎松
名古屋工業大学　大学院工学研究科　情報工学専攻
e-mail: {watanabe, ozono, itota, tora}@ics.nitech.ac.jp

1　はじめに

本論文では，共起語に基づいたトピックの経年変化を利用した情報推薦について述べる．これまでの情報推薦において，推薦システムが推薦する情報は，新規なもの，またユーザにとって未知である情報とは限らない．そこで本論文では，データベース中の文書に含まれる語の共起に着目し，語の共起数からトピックの経年変化を観測する．本論文では，推薦される情報が有用なトピックであるかを判定する．また，推薦される情報がユーザにとり未知の情報であるかを判定する．

2　共起語に基づくトピックの経年変化

本論文では，研究トピックの経年変化を測定するため，トピックモデルを構築する．本トピックモデルは，語の共起頻度，および語の新近性に基づいている．本トピックモデルは，研究支援システムにおけるデータベース中の文献より作成する．データベースに含まれる文献には，情報工学に関する文献情報，および，発表された年度を検索することができる．年度ごとの文献において，語の共起の頻度，および語の最新性を観察することで，年度ごとにおいて，出現する語の傾向を測定することができる．しかし，共起文献数も年毎に増加する傾向にあり，その傾向をつかむのにはふさわしくない．一方，Jaccard 係数 [1] は情報検索において，キーワード間の関係の強さを表すために用いられる係数である．あるキーワード X で情報検索したときのヒット数を $|X|$ とすると，キーワード A と B の間の Jaccard 係数は $Jaccard(A,B) = \frac{|w_A \cap w_B|}{|w_A \cup w_B|}$ と定義される．この Jaccard 係数の経年変化を観測することで，そのキーワードに関する研究がどの程度まで進んでいるかを判断する．トピックモデルの構築において，語の共起の出現，および語の最新性を観察するが，ここでは，以下の４つに場合分けすることができる（図 1）．

- 語の共起頻度が高いならば，その分野は既知の知識である．
- 語の共起頻度が低いならば，その分野は知られていない．もしくは，その分野を研究する研究者が少ない．
- 語の最新性が，年ごとに上昇しているならば，その分野はよく研究され，今後伸びるトピックである．
- 語の最新性が，年ごとに減少しているならば，その分野は，今後すたれていくトピックである．

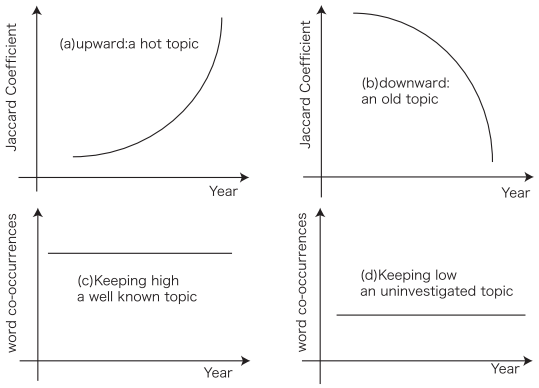

図 1: 語の共起頻度と最新性の関係

本論文では，研究支援システム Papits のデータベースより，文献を選択する上で，語 w_n および語 w_m のトピックモデル $T_{w_n w_m}$ を以下のように計算する．

$$T_{w_n w_m} = T_{freq}(w_n, w_m) \cdot T_{recency}(w_n, w_m) \quad (1)$$

ここで，式 (1) における，語 w_n および語 w_m は，同一文に共起する語である．$T_{freq}(w_n, w_m)$ は，研究支援システム Papits のデータベースに含まれる文献に出現する語の共起数である．そして，$T_{recency}(w_n, w_m)$ は，語 w_n および語 w_m からなるトピックの新しさを測定する値である．$T_{recency}(w_n, w_m)$ は，以下の式 (2) を用いて計算する．

$$T_{recency}(w_n, w_m) = \begin{cases} \frac{R_{w_n w_m}(t)}{R_{w_n w_m}(t-1)} \\ (R_{w_n w_m}(t-1) \neq 0) \\ R_{w_n w_m}(t) \\ (R_{w_n w_m}(t-1) = 0) \\ R_{w_n w_m}(0) = 1 \end{cases} \quad (2)$$

$$R_{w_n w_m}(t) = Jaccard(w_n, w_m) = \frac{|w_n \cap w_m|}{|w_n \cup w_m|} \quad (3)$$

An Information Recommendation Using the Change of the Research Topics Based on Word Co-occurrence

Satoshi WATANABE, Tadachika OZONO, Takayuki ITO, Toramatsu SHINTANI
Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology Gokiso, Showa-ku, Nagoya 466-8555 JAPAN

ここで，$R_{w_n w_m}(t)$ は，時刻 t における，語 w_n，および語 w_m の最新性である．語 w_n，および語 w_m の最新性は，式 (3) を用いて計算する．式 (3) では，Jaccard 係数を用い，$R_{w_n w_m}(t)$ の値を求める．$R_{w_n w_m}(t)$ と $R_{w_n w_m}(t-1)$ を比較することで，語 w_n，および語 w_m の最新性を求める．すなわち，トピックの最新性を値の上昇，もしくは下降を利用し観察する．

3　トピックの経年変化を利用した情報推薦

本論文では，論文 [2] において述べた，語の共起グラフがスケールフリー性を持つことを利用する．以下に本提案手法を用いたユーザモデルと文献の類似度判定関数を示す．ユーザの所持する文献，および執筆した文献から得られたユーザモデルを用い，文献ファイルから得られた語の共起関係を比較し，類似度の高い文献を示すことにより関連性の高い文献を推薦することが可能になる．式 (4) はユーザモデル U_X，文献 P_Y の類似度を表す関数である．

$$sim(U_X, P_Y) = \frac{1}{n^2}\sum_{i=1}^{n}(\sum_{j=1}^{n}\Pi_i\Pi_j) \qquad (4)$$

ここで，n は前処理を適用した後の文献 P_Y に含まれる語数である．前処理とは，不要後の除去，および接辞処理のことを指す．また，Π_i は語 i における最新性および頻度の積である [2]．式 (4) を用いて計算を行うとユーザモデル U_X と文献 P_Y の類似度を求めることができる．式 (4) は，語の共起頻度，および適応度を利用した計算式である．本類似度計算式の値を推薦する値として使用することが可能である．上記の式 (4) を用いた場合，推薦する文献が，有用なトピックであるのか，また，現在対象となるユーザにとって興味があるのかを考慮していない．そこで，式 (4) を以下のように拡張する．

$$sim(U_X, P_Y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{T_{w_i w_j}}{k_{w_i w_j}}\Pi_{w_i}\Pi_{w_j} \qquad (5)$$

ここで，$T_{w_i w_j}$ は，2 で述べた，語 w_i，および語 w_j に対する，トピックモデルの値である．また，$k_{w_i w_j}$ は，語 w_i，および語 w_j の語の共起である．

4　評価実験

実験の設定は以下の通りである．まず，8 か月間に被験者が閲覧した文献を収集する．これらの文献を閲覧した順に処理し，前処理を行う．文献に含まれる語をノード，文中における共起関係をリンクとし，グラフに付加する．本実験では，研究支援システム Papits のデータベースに存在する文献を用いる．データベースに含まれる文献は，英語文献であり，情報科学に関する文献が格納されている．ここでは，評価実験を通じて，ベクトル空間モデル，語の共起シソーラス [3]，および IRM[4] との比較を行う．図 2 は，それぞれの手法に関しての適合率である．本実験での適合率は，有用なトピックの情報，およびユーザが未知の情報を示す．図 2 における，横軸は，時間経過を表し，縦軸は，適合率を示す．ベクトル空間モデル，シソーラスを用いた類似度計算，IRM などの既存の手法および，式 (4) の適合率は，変化しないもしくは，下降しているが，式 (5) による適合率は，時間経過ごとに上昇している．

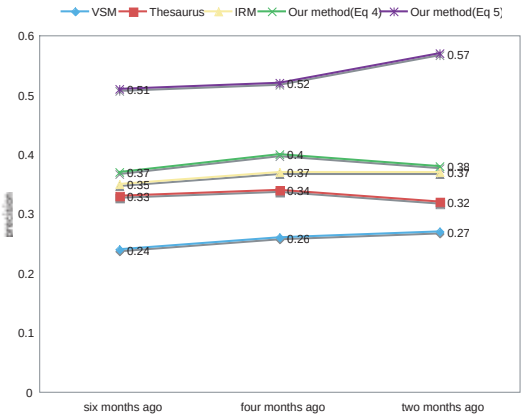

図 2: 実験結果

5　おわりに

本論文では，共起語に基づいたトピックの経年変化を用いた情報推薦について述べた．共起語に基づいたトピックの経年変化を利用することで，有用なトピックの情報，およびユーザにとり未知の情報を推薦することが可能である．

参考文献

[1] H. Kautz, B. Selman, M. Shah. The Hidden Web. AI Magazine. Vol. 18, No. 2, pp. 27-36, 1997

[2] 渡邉倫，伊藤孝行，大囿忠親，新谷虎松，“研究活動におけるスケールフリーネットワークを用いたユーザモデルの試作”，日本ソフトウェア科学会第 20 回大会論文集, 日本ソフトウェア科学会, 2003.

[3] 後藤将志，大囿忠親，新谷虎松，“シソーラスを用いた情報間類似性評価手法について，” 第 64 回情報処理学会全国大会，2002.

[4] 松尾豊，福田隼人，石塚満，“ユーザ個人の閲覧履歴からのキーワード抽出によるブラウジング支援，” 人工知能学会論文誌 Vol.18，　No.4，pp.203-211，2003.