共起シソーラスに基づく類似度比較を用いた情報推薦機構の試作

渡邉倫　大囿忠親　新谷虎松　（名古屋工業大学）

1　はじめに

文献サーベイを主に文献を読むことによる情報収集活動と捉えると，文献を探すこと，読むこと，文献に対する情報を管理することが必要となる．これらの作業に費やされる時間的コストは大変大きい．本研究室において開発中の研究支援システム Papits[1] は個人の所持しているファイル (論文) をユーザの興味と考え，そのファイルに基づいて，ユーザに新たな論文を推薦する．本論文では，Papits における，共起シソーラスをに基づく類似度比較について述べ，ユーザの興味があるファイルの推薦を行う機構を示す．

2　共起シソーラスに基づく類似度

本研究では，ユーザの興味や論文の内容を共起シソーラスによって表現し，これらの共起シソーラス間の類似度を計ることによって推薦を実現する．単語の共起関係によって，文書を特徴づけることが可能である．ユーザの持つファイル中と論文に出現する単語の共起関係はそれぞれユーザの興味と論文の特徴を表す．共起シソーラスとは語と語の共起関係を表現するグラフである．語を頂点とし関係を辺とする．本研究では，ユーザが所持するファイルの集合中に出現する単語の共起関係がユーザの興味を表現する．その例として，単語の共起パターンは（エージェント，オークション），（エージェント，情報検索）のようにユーザの興味を表す．

以下に共起シソーラス間の類似度判定関数を示す．ユーザの持つファイルから得られた共起シソーラスと論文ファイルから得られた共起シソーラスとの類似度を計算し，類似度の高い論文を示すことにより関連性の高い論文を推薦することが可能になる．式 (1) は共起シソーラス T_X，T_Y の類似度を表す関数である．

$$sim(T_X, T_Y) = \frac{1}{n} \times \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{1}$$

T_X と T_Y における異なり語数 n の正方行列 M で表すことができる．行列 M の要素は T_X,T_Y における，単語の出現回数と共起回数の指標から求められる値である．$x_i(y_i)$ は行列 M の i 列目の要素（単語）を元にした $T_X(T_Y)$ におけるベクトルである．式 (1) を用いて計算を行うと文書群 X と文書群 Y の類似度を求めることができる．この類似度を推薦する値として使用することが可能である．

3　Papits における情報推薦エージェント

Papits は図 1 のように大きくわけて 2 つのエージェントシステムから構成されている．1)Web から自律的にファイルを収集する情報収集エージェント，2) ユーザにファイルを推薦する推薦エージェントである．情報収集エージェントがデータベースに収集したファイルを蓄積し，推薦エージェントがデータベースに新たに追加されたファイルのシソーラスを作成する．

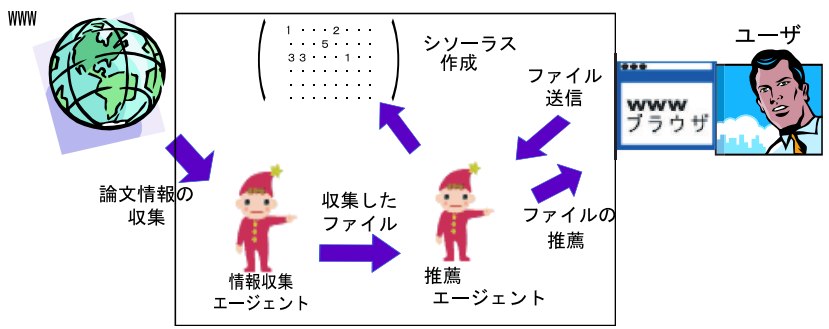

図 1: Papits の構成

推薦の準備として，ユーザは推薦エージェントにファイルを登録する．推薦エージェントはそのファイルの共起シソーラスを作成する．また，情報収集エージェントが Web からファイルを得たとき，推薦エージェントはそのファイルのシソーラスを作成する．その 2 つのシソーラスから類似度を計算し，推薦すべきファイルかどうかを判定する．その後，ユーザが推薦を希望したとき，エージェントは推薦するファイルをユーザに提供する．

4　まとめ

本稿では Papits における，シソーラス間の類似度比較方法の概略を説明し，情報推薦機構の実現について説明した．これにより，ユーザの興味があるファイルを推薦することが可能である．

参考文献

[1] 後藤将志，大囿忠親，新谷虎松：”文献情報共有支援システム Papits における Know-Who 検索機能の実現について ”，計測自動制御学会第 2 回システムインテグレーション部門学術講演会 (SI2001) 論文集，pp.381-382,2001.