動的情報フィルタリングのための逐次的クラスタリング*

5Q－6

堀井 則彰† 上原 邦昭†‡

†神戸大学工学部情報知能工学科

‡神戸大学都市安全研究センター

1. はじめに

本稿では，領域知識を導入した動的なクラスタリングにより Web ページの概念木を構築し，「内容が類似しているページは概念木上で近接している」という仮定に基づき，ユーザの関心を満たすページを抽出する手法を提案する．領域知識は，ユーザとの対話を用いた「失敗からの学習」により漸増的に構成される．このため，情報の変化とユーザの意思に応じて概念木を更新させることができ，情報の動的性とユーザの好みに対応できるようになっている．

2. Web ページのデータ表現

一般に，テキスト文書のデータ表現には単語の出現頻度が用いられるが，データ表現の情報量が多くなるという問題がある．また，Web ページには多くのノイズ，つまりページの内容には関係のない単語が含まれているため，Web ページの内容を単語の出現頻度だけで正確に表現することは困難である．このため，本システムでは重要語と呼ぶ概念を導入し，Web ページのデータ表現には重要語の出現の有無を用いている．

しかしながら，ある重要語が Web ページに出現しているとしても，重要語が属する問題領域に含まれる他の重要語が出現していなければ，その問題領域に関する情報は Web ページ中にないと考えられる．たとえば，研究者の研究テーマの一覧や単語の箇条書などのページがこれに相当する．このため，出現している重要語と意味的に関連する他の重要語が出現していない場合には，仮に重要語が存在してもノイズとみなし，領域知識を用いて除去している．

領域知識とは，全ての重要語が葉となるように構成されているシソーラスであり，概念が類似する重要語は近接するように配置されている．本システムでは，領域知識を自動的に構築しており [1]，各重要語をその語義文に基づいて属性-値対の組で表現し，重要語間の意味的類似度を計算して領域知識を構築している．したがって，属性値が 1 で Web ページが持つ情報と関係がある重要語はシソーラス中で近接しているが，ノイズとなる重要語は孤立していると考えられる．このような場合には，シソーラス中で孤立している重要語の属性値を 0 にしている．

また，1 つの Web ページが相反する 2 つの問題領域について記述されている場合，いずれかの内容についてはあまり詳しく述べられていないと考えられるため，片方の問題領域に関する重要語をノイズとみなし，先験情報を用いて除去している．たとえば，先験情報とは「人工知能の分野で “reinforcement learning” の Web ページに “natural language” で使われる単語は出現しない」などという経験的な情報である．

先験情報を表現するために，検閲リンクという考え方を導入している．検閲リンクとは，ある単語の出現が他の単語の出現を無効にするリンクである．シソーラス中であるノード (Node1) から他のノード (Node2) に検閲リンクが張られている時，Node1 以下の部分木に含まれる重要語が Web ページに出現すれば，Node2 以下に含まれる重要語の属性値は全て 0 にされる．逆に，Node1 以下に含まれる重要語が Web ページに出現しなければ，Node2 以下に含まれる重要語の属性値はそのままとなる．

一方，仮に 2 つの Web ページが同じ問題領域について記述されていても，各ページに出現している重要語が意味的に類似していなければ，システムによって同じ問題領域について記述されているページと判断されることはない．この問題を解決するために，共通属性という考え方を導入する．共通属性とは複数の重要語の関係を示すもので，シソーラスでの中間ノードに相当している．共通属性の値は，共通属性が示す重要語の属性値のうち，少なくとも一つが 1 であれば共通属性の値が 1 となり，それ以外は 0 となる．

3. 動的クラスタリングの詳細

本システムでは，新しい Web ページが得られる毎に COBWEB [2] を用いてクラスタリングを行ない，内容が類似した Web ページが近接するように概念木を構築している．なお，ユーザが興味のある Web ページを検索する場合には，関心を持つ内容について記述されたテキスト，たとえば，ユーザが関心を持つ論文や Web ページをシステムに入力し，概念木への追加を試みる．最終的に，入力されたテキストの近くにクラスタリングされている Web ページが，ユーザの関

---

*Incremental Clustering for Dynamic Information Filtering
Noriaki Horii† and Kuniaki Uehara†‡
†Department of Computer and Systems Engineering, Faculty of Engineering, Kobe University
‡Research Center for Urban Safety and Security, Kobe University

心を満たす Web ページとして検索される．

COBWEB は逐次的なアルゴリズムであるため，1 つの Web ページをクラスタリングする毎に概念木を評価することができる．ある時点で Web ページの誤分類が発見された場合，クラスタリングを改善するために領域知識の更新が必要となる．このため，本システムではユーザとの対話による「失敗からの学習」を用いて，Web ページの誤分類をトリガとする領域知識の漸増的な構成を行なっている．

ユーザは，Web ページが誤分類されていると感じた場合，新たな重要語をシソーラスに付加することができる．新たな重要語をシソーラスに付加する方法は，付加する重要語の属するクラスが既にシステムに与えられている場合とそうでない場合とで異なる．既にクラスが与えられている場合には，システムは付加する重要語のクラスと同じクラスを持つ重要語が属するノードに付加し，クラスが与えられていない場合には，付加する重要語のクラスと最も関係のある重要語が属するノードに付加する．

Web ページの属性-値対の組で，属性値が 1 の重要語にもかかわらず，Web ページの内容と関係がないと感じた場合，検閲リンクを付加することができる．ユーザが属性値が 1 である重要語をページの内容と関係があるものとそうでないものに分けると，システムは各集合に対してシソーラス中での各重要語が属するノードのうち最上位にあるノードを求め，ページの内容と関係がある重要語の集合で求めた最上位ノードからページの内容と関係がない集合で求めた最上位ノードへ検閲リンクを張る．

以上の操作の繰り返しにより，情報の動的性とユーザの好みに対応した動的なクラスタリングが可能となっている．また，本システムは個人重視の情報フィルリングであり，個人の趣味範囲は限られているため，領域知識に含まれる重要語は膨大な量になることはない．

4. 実験と評価

提案したシステムを用いた実験を行なう．問題領域は “Artificial Intelligence” とし，検索エンジン Alta Vista[§] に領域名を入力して，1997 年 1 月に入手したページ (Jan-Data) と，同年 5 月に入手したページ (May-Data) をデータとして用いている．また，初期領域知識の構築には論文誌のタイトルを使用し，重要語は論文タイトルに含まれる単語，クラス名は論文タイトルが分類されているトピック名としている．なお，重要語の語義文には参考文献のタイトルを使用している．

[§]http://altavista.digital.com

実験では，入手した Web ページを 訓練データとテストデータに分け，訓練データを用いて領域知識の更新を行ない，更新する前と後の領域知識をそれぞれ用いてテストデータのクラスタリングを行ない，各クラスに属する Web ページのクラスタリングの精度を比較評価している．評価尺度には再現率と適合率を用いている．図 1 は再現率の結果のみを示している．

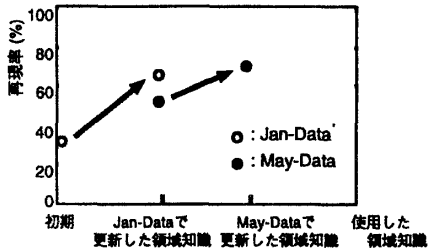

図 1: “Artificial Intelligence” での再現率

図 1 から再現率は向上していることが分かる．これは，漸増的に構成される領域知識が再現率の向上に有効であることを示している．また，Jan-Data で更新した領域知識を用いたクラスタリングでは，May-Data の再現率が Jan-Data よりも 9.8% 低い．これは，入手した Web ページの内容の時間的な変化が原因と考えられる．しかしながら，May-Data を用いた領域知識の更新により再現率は向上しており，漸増的に構成される領域知識は情報の変化に対して有効であることが確認された．

5. おわりに

本稿では，漸増的に構成される領域知識を導入した情報フィルタリングを提案し，情報の動的性に対して有効であることを確認した．実験では初期領域知識の構築に論文誌のタイトルを利用したが，このような情報がない場合は Web ページ自身を直接利用することもできる．すなわち，Web ページのタグおよびアンカー部分に含まれる単語を重要語，タグが示す段落部分およびリンク先のページを語義文として，初期領域知識を構築すればよい．もちろん，構築される領域知識は貧弱なものとなるが，漸増的な領域知識の構成によりユーザの好みを反映した信頼のおける領域知識が構成される．このため，適切な情報が利用できない場合でも有効な情報フィルタリングを行なうことができる．

参考文献

[1] 山崎 毅文, Pazzani, J.: 帰納学習アルゴリズムと階層型クラスタリング手法を用いた概念シソーラスの自動構築及び更新, 情報処理学会情報学基礎研究会, Vol. 39, No. 7, pp. 49–56 (1995).

[2] Fisher, D.: Knowledge Acquisition Via Incremental Conceptual Clustering, Machine Learning, Vol. 2, pp. 139–172 (1987).