論　文　　Technical Papers

WWW 上で文献検索プラニングを行なうソフトウェアエージェント NaviPlan

NaviPlan: a Software Agent for Planning Concept Understanding on WWW

大澤　幸生*1　　山田　誠二*2
Yukio Ohsawa　　Seiji Yamada

*1　大阪大学大学院基礎工学系研究科
Graduate School of Engineering Science, Osaka University
*2　東京工業大学大学院総合理工学研究科知能システム科学専攻
CISS, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology

19YY 年 MM 月 DD 日　受理

Summary

We propose navigation planning to automatically generate a sequence of Web pages by which a user systematically understand a target concept. First, with a planning framework, we formalize the browsing task in the WWW. Action is defined as the understanding of a Web page, and an operator for a Web page consists of conditional/effect knowledge. Here, we developed and employed a method to generate an operator from a Web page by extracting condition/effect terms with keyword extraction techniques. The implemented the navigation planning system is tested by comparing with methods using a search engine and link-tracing (like a Web robot). As results, we found out navigation planning is a promising approach to assist the concept understanding in the WWW.

1. はじめに

WWW 上の Web ページを見ることにより，“ある概念（**目標概念**と呼ぶ）を理解する” という作業が，多くのユーザにより日常的に行われている．しかし，この概念理解に有用な Web ページを探す作業は，検索エンジンや Web ロボット [Cheong 96] を利用することにより部分的には効率化，自動化できるものの，結局検索されたページを見てチェックするところはユーザがやらねばならず，それによる時間のロスは大きい．

本研究では，概念理解のための有用な Web ページの探索を人工知能におけるプラニングとして捉え，その作業全体を自動化することを目的とする．ここで，プランとは，目標概念を理解するために，ユーザが見るべき Web ページの系列である．このように，目標概念を理解するための Web ページの系列を生成するプラニングを，**ナビゲーションプラニング**と呼ぶ．

ナビゲーションプラニングでは，一つの Web ページを見て理解することを，プラニングにおける一つの行為として捉え，その行為を **U-オペレータ**として定義する．本研究で構築したナビゲーションプラニング・システム *NaviPlan* はプラニング過程に必要に応じて，Web ページから U-オペレータを自動生成する．

従来の Softbot[Etzioni 94]，Occam[Kwok 96] などは，情報収集の手続きプランとして，UNIX コマンドやデータベースの検索コマンドの系列を自動生成する．一方，*NaviPlan* はユーザの概念理解を誘導するようなプランを，ユーザが見るべき Web ページの系列で表現する．Web ロボット [Cheong 96] も検索エンジンの URL データベース構築のために URL 情報を自動収集するが，リンクの張られている Web ページ間を探索するに過ぎない．*NaviPlan* は，後述するようにリンクが明示的に張られていない Web ページも探索することが可能であり，目標概念の理解のために適切に制御される．又，WebWatcher[Armstrong 95]，Letizia[Lieberman 95] のように次に見るべきページを提示するだけでなく，*NaviPlan* は Web ページ系列を

生成／提示する．

2. ナビゲーションプラニング

ユーザが目標概念を理解するために有用なWebページをブラウズする行為は，以下のような対応によりプラニング [Russell 95] として定式化できる．

- **行為：**Web ページに記述されている概念を理解する．
- **状態：**ユーザの知識状態．既知の概念を表す単語の集合により記述．
- **初期状態：**ユーザの初期の知識状態．
- **目標状態：**ユーザが理解したい目標概念．単語の集合により記述．
- **オペレータ：**Web ページを見て，知識を獲得する行為を表す **U-オペレータ** U-Op(URL) は，以下の要素からなる．
  - **ラベル：**Web ページの URL でラベル付け．
  - **条件知識：**その Web ページを理解するために必要な知識 $C = \{c_1, \cdots, c_i\}$．Cの要素（Web ページ中の単語である）を条件語という．
  - **効果知識：**その Web ページを理解することにより得られる知識 $E = \{e_1, \cdots, e_j\}$．Eの要素（Web ページ中の単語）を効果語という．

3. Web ページからの U-オペレータ生成

NaviPlan は U-オペレータを自動生成する．というのは，世界中の膨大な数の Web ページすべてについて，U-オペレータを生成しておくのは，現実的には不可能だからである．以下 U-オペレータをどのように抽出するか述べる．

[タグ構造による抽出] 情報検索の分野において提案されたキーワード抽出法のほとんどは単語の出現頻度による．しかし多くの Web ページは HTML により構造化されているので，ここでは，そのタグ構造（`<TITLE>`, `<Hn>`, `<A HREF=...>`）を利用する．まず `<A HREF=URL>`と`</A>`の間の単語を条件語の候補とする．というのは Web ページにリンクされている単語は，その Web ページを理解するために重要なことが多いからである．また Web ページのタイトルと見出しは，そのページの主張点を表していることが多いので，`<TITLE>`と`</TITLE>`および`<Hn>`と`</Hn>`の間の単語を効果語の候補とする．

しかし，HTML タグだけでは条件／効果語を高い精度で求めることができない．そこで，次のキーワード抽出法 *KeyGraph* を併用することにする．

[KeyGraph による抽出] *KeyGraph* [Osawa 98] では，まず文章の土台となる概念を表す単語を得るので，これらを *NaviPlan* では条件語の候補とする．次に *KeyGraph* では，その文章の主張と見なされる単語を得るので *NaviPlan* ではこれらを効果語の候補とする．*KeyGraph* ではこれらのキーワードを条件／効果らしさを表わす実数値（図 1）と共に得ることができる．*KeyGraph* の併用により，タグ構造のないプレーンテキストからの条件／効果知識の抽出も可能となる．

ナビゲーションプラニングでは以下のように，条件知識，効果知識を，この *KeyGraph* とタグ情報の組み合わせによって抽出する．

4. プラニング手続き

ナビゲーションプラニングの手続きを，図 2に示す．簡単にいうと，目標概念を起点として U-オペレータをつないで行くのである．付加的な入力として，**文脈語**を用いている．文脈語は，目標概念が属する領域を表し，同じ単語でも使われる領域により意味が異なる場合に付加することで，探索の精度を上げる効果がある．

入力
- 入力ファイル（HTML ファイル，またはプレーンテキスト）とその URL

制御パラメータ
- `<TITLE>`の重み: α, `<Hn>`の重み: β, `<A HREF=URL>`の重み: γ
- 条件語数: N_C
- 効果語数: N_E

出力
- U-オペレータ U-Op(URL) = (URL, 条件知識，効果知識)

手続き

(1) *KeyGraph* により，入力ファイルから条件語の候補の集合 C_C と効果知識の候補 C_E を抽出．それらの候補は，[0, 1] の重みを持つ．
(2) `<TITLE>`と`/TITLE>`の間の単語を重みをαとして C_E に追加．
(3) `<Hn>`と`</Hn>`の間の単語を重みをβとして，C_E に追加．
(4) C_C 中の単語で，`<A HREF=URL>`と`/A>`の間に含まれる単語の重みをγに変更．
(5) C_C から重みが上位 N_C 個の単語を条件知識として，C_E から重みが上位 N_E 個の単語を効果知識として取り出して，U-オペレータを生成．`<A HREF=URL>`から抽出された条件語には，リンク先の URL を割り当てる．

図 1 Ext-OP: U-オペレータ生成手続き

5. *NaviPlan* の実験による性能評価

NaviPlan は，Perl で記述されており，ユーザとのインタフェースとして Web ブラウザを用いる．図 3に，目標概念が “concept formation”（概念形成）で，文脈御”AI”の場合の *NaviPlan* の出力プラン（深さ制限 $l = 4$）を示す．Web ページ 1 は，目標概念に直接結びつくものであり，概念形成について ILP（Inductive Logic Programming）を基にして説明されている．しかし，ILP 自体についての説明はページ 1 にはない．ILP については，Web ページ 2〜4 のうち，3 で詳しく述べられ，4 には関連する機械学習の論文アブストラクトが紹介されている．実際，ユーザが 1，3，4 の Web ページを読むことによって概念形成はもとより，それと ILP，機械学習との関係まで理解ができる．

入力
- 目標状態（= 目標概念）G_0：ユーザが理解したい概念を表す単語の集合.
- 文脈語 GC：目標状態の背景領域を表す単語集合.

制御パラメータ
- 初期状態 IS：ユーザの初期の知識状態.
- 深さ制限 l：プランの最大長.
- Web ページの制限数 w：検索エンジンにより得られる関連 Web ページの最大数.
- ビーム探索の幅 b
- 状態ノードの評価関数 H

出力
- プラン P: U-オペレータ U-Op(URL) の系列.

手続き

(1) $N_0 = [n_0] = [(G_0, [])]$, $i = 0$ で初期化.（N_iは，深さ i での状態ノード系列．状態ノードは，$n_i = (G_i, P_i)$ で記述される．ここで，G_iは副目標を表す単語の集合，プラン P_iは，U-オペレータ系列.）

(2) もし $i = l$，または，N_iが空なら，プラン Pを出力して終了.

(3) 以下の手続きを，N_iのすべてのノード n に対して適用して，$N_{i+1} = [\,]$ で初期化.

(a) 差異 $D = G \cap \overline{IS}$ を抽出.

(b) 検索エンジンに $D \cup GC$を入力して，w 個の関連 Web ページの URL を得る.

(c) 得られた URL の http サーバとの TCP/IP 接続により，その Web ページ（HTML ファイル，またはテキストファイル）を獲得.

(d) 得られた Web ページから，手続き **Ext-OP** を使って，U-オペレータを生成．そして，効果知識 Eと n の副目標との交わりが空でない U-オペレータに対して，下の手続きを適用.

i. その U-オペレータの Eを n の G から取り除いたものに，U-オペレータの Cを追加することにより，G'を得る.

ii. n のプラン Pの先頭に，その U-オペレータを加えることにより，P'を得る.

iii. 状態ノード (G', P') を N_{n+1}に追加.

(4) N_{n+1}の各状態ノードにつき，評価関数 Hによる評価を行い，その上位 b 個のノードで，N_{n+1}を更新.

(5) $i = i + 1$ として，(2) へ.

図 2 ナビゲーションプラニング手続き

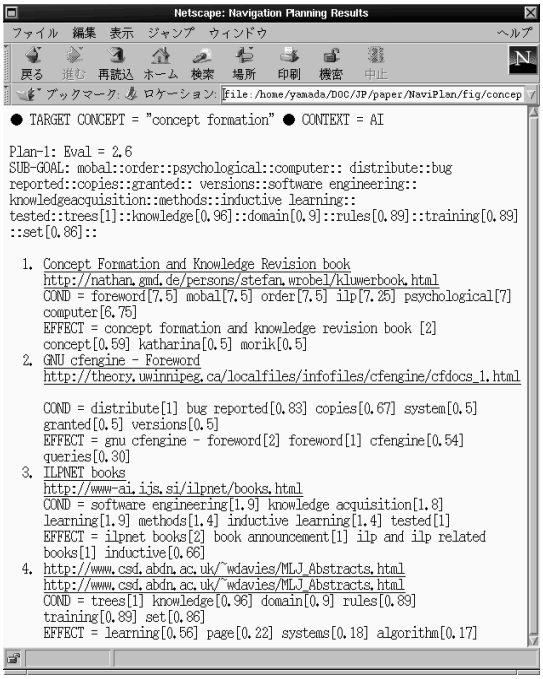

図 3 “Concept formation” のプラン

以下，*NaviPlan* の性能を，実際にシステムを用いた結果から評価した結果を示す．*NaviPlan* の中で用いる検索エンジンと以下の実験で比較に用いる検索エンジンとして，広範囲に検索可能な検索エンジンである MetaCrawler[*1] を用いた．

5・1 実 験 方 法

われわれは，以下の 5 つの手法を比較した．

手法 1: 検索エンジンのみ 目標概念と文脈語を検索キーとして入力する．ユーザは出力された Web ページを，ソートされた優先順に目標概念を理解するまで読み進める．

手法 2: 検索エンジン＋リンクの利用 手法 1 と同様の入力と出力ページをまず得る．次にユーザは，手法 1 と同様に優先順位に従って Web ページをゴールを理解するまで読み進めるが，途中でページ中のリンクに従って他のページに移っても良い．

手法 3: NaviPlan ユーザは，*NaviPlan* に目標概念と文脈語を入力する．そして，読み進めるべきと *NaviPlan* が判断した順番に目標概念を理解するま

*1 http://www.metacrawler.com/

で読み進める．

手法 4: NaviPlan＋リンクの利用　ユーザは，手法 3 と同様の入力を *NaviPlan* に与え，その結果として得られたページを，手法 3 に加えて手法 2 と同様のリンク利用を行いながら目標概念を理解するまで読み進める．

この実験において設定した *NaviPlan* のパラメータは，図 2では，$IS = []$，$l = 4$，$w = 8$，$b = 4$ で，図 1では，$\alpha = 2$，$\beta = 1$，$\gamma = 1$，$N_C = 6, N_E = 4$ とした．一度のプランニングに要した時間は 2 ～ 3 時間であったが，そのほとんどの時間は WWW サーバとの通信による HTML ファイルの獲得に費やされた．

5・2　実験結果と評価

43 通りの検索キーについて上記の 5 つの手法を比較した結果，手法 1 と手法 3 については，ともに 26 通りの目標概念が正しく理解された．リンク利用を手法 1 と手法 3 と併用した手法 2 と手法 4 ではそれぞれ，33 通りと 32 通りに増加したので，リンク利用が有効であることがわかる．

次に，ユーザが目標概念理解までにどれだけ労力を使ったか，すなわちどれだけ多くの Web ページを読んだかを評価した．リンクの利用がユーザの理解を促進することは先述のとおりであるから，リンク利用を合わせた手法 2 と手法 4 との比較を行なった．図 4は，そのデータを示すグラフである．横軸は，検索エンジン（MetaCrawler）で得られた出力ページの数（検索語がどれだけ Web で普及しているかを表わしていると考えられる）であり，縦軸は，実際に目標概念を理解できるまでに読む必要があった Web ページ数である．図 4から以下の 2 つの特徴が得られる．

- MetaCrawler（手法 2）による検索よりも *NaviPlan*（手法 4）によるプラニングがユーザの効率的な理解を助ける効果がある．
- Web 上でよく普及した目標概念では，手法 2 は多くのページをユーザに読ませてしまう．このことは，その概念が広く知れ渡っていて，改めて直接定義する（すなわち MetaCrawler で得られる）Web ページが少ないせいであろう．

6. まとめと課題

ユーザが理解したい目標概念の概念理解に有用な Web ページの系列を提示することができるナビゲーションプラニングを提案し，その実装システム *NaviPlan* を構築した．WWW 上でのユーザのブラウジングをプラニングの枠組で定式化し，必要に応じて Web ページから U-オペレータを自動生成して，プラニングを行い，有用な Web ページの系列を得る．検索エンジン等との実験的比較を行い，*NaviPlan* の有用性を示した．

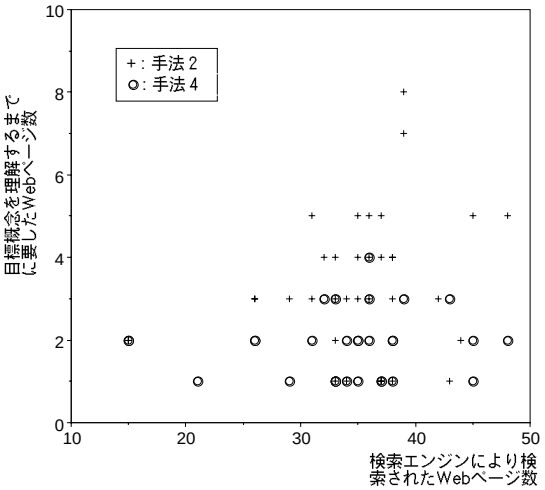

図 4　手法 2 と手法 4 の比較．

しかし，現状では U-オペレータの精度は十分に高くない．その原因は，U- オペレータで用いたヒューリスティックにある．例えば，見出しのタグ `<Hn>` を利用して効果知識の抽出を行っているが，見出し情報を用いて得られた効果知識が信頼できないことは [Osawa 98] でも指摘されている．このヒューリスティックの精度を高めることは，*NaviPlan* の第一の検討課題である．

◇ 参　考　文　献 ◇

[Armstrong 95] Armstrong, R., et al.: WabWatcher: A Learning Apprentice for the World Wide Web, *The AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environment*, (1995).

[Cheong 96] Cheong, F.C.: *Internet Agents: Spiders, Wanderers, Brokers, and Bots*, New Riders (1996).

[Etzioni 94] Etzioni, O. and Weld, D.: A SoftBot-based Interface to the Internet, *Communication of the ACM*, Vol.37 No.7, 72–76 (1994).

[Kwok 96] Kwok C. T. and Weld, D. S.: Planning to Gather Information, *Proc. of AAAI-96*, 32–39 (1996).

[Lieberman 95] Lieberman, H.: Letizia: An Agent that Sssists Web Browsing, *Proc. of IJCAI-95*, 924–929 (1995).

[Osawa 98] Ohsawa, Y. et al.: *KeyGraph*: Automatic Indexing by Co-occurrence Graph Based on Building Construction Metaphor, *Proc. of IEEE Advanced Digital Library Conference*, 12–18 (1998)

[Russell 95] Russell, S. and Norvig, P.: *Artificial Intelligence –A Modern Approach–*, Prentice-Hall (1995).

〔担当編集委員：××○○，査読者：○○××〕