1D4-06

Webページ集合からの階層的知識の構築

Constrcution of the hierarchical knowledge from Web pages

間瀬 心博[*1] 山田 誠二[*2]
Motohiro Mase Seiji Yamada

[*1]東京工業大学 Tokyo Institute of Technology
[*2]国立情報学研究所 The National Institute of Informatics

This paper showed the system, which provide construction of hierarchical knowledge from structured keywords. User inputs structured keywords that are related with the information obtained by Web browsing of user. The system extracts new keywords with the extraction rules consists of lexico-sysntactic pattern and HTML tag structure, and show hierarchical knowledge added new extracted keywords. The system supports that a user obtains information efficiently.

1. はじめに

現在 WWW を情報源として利用した情報収集が広く行われている．一般にユーザは検索エンジンを用いた絞込みを行い，目的の情報の含まれる Web ページを探索する．ユーザが獲得できる情報は選択したリンクに依存しているため，閲覧した Web ページの周囲に有用な関連情報が存在していたとしても，情報の存在に気づかずに見落すこともある．しかし，ユーザが網羅的に Web ページを閲覧することは現実的ではなく，閲覧する Web ページ数が膨大であれば，獲得できる情報間の関係を把握することは困難になる．そのため，ユーザが閲覧して得られた情報と関連する未見な情報について，情報間の階層的な関係と各情報を得るのに必要な Web ページの集合を提示し，ユーザの情報獲得を支援する必要があると考える．

そこで本研究では，ユーザにブラウジングによって獲得した情報を特徴づけるキーワードを提示してもらい，関連する情報のキーワードとの階層的な関係と各キーワードを理解するのを助ける Web ページ集合を合わせて提示するシステムを提案する．ユーザはシステムを利用することで，興味のある未見の情報を発見し情報間の関係を把握することで，効率的な情報獲得が可能となると期待できる．

2. システムの概要

2.1 システムの対象

本システムでは，ブラウジングを行うことで目的の情報を獲得し，さらに周辺に存在する未見の関連情報も獲得したいと考えるユーザを対象とする．システムはユーザのブラウジング中にオンラインで情報を提示するのではなく，ブラウジング終了後にユーザが獲得した情報をもとに関連情報を収集しその関係を抽出し提示する．ユーザが効率的に情報獲得するにあたり，システムが提示する情報の間の関係は階層的である方が望ましいと考える．しかし，単純に関連する Web ページを収集し情報間の関係を抽出しても，実際の階層的な関係を抽出することは難しい．そこで本研究ではユーザにブラウジングで獲得した情報を特徴付けるキーワードを用いた階層的な構造を指定してもらい，その階層構造をもとにして周辺に存在する未見の情報のキーワードとの階層的な関係を抽出する．ユーザによって入力された階層構造を利用することで，ユーザの情報に対する視点から大きく外れることなく階層構造を展開することが可能となる．さらに，複数のキーワードを指定されることで，膨大な Web ページから無関係な分野のページを集める無駄を省くことができる．現在は，ユーザによるキーワードや階層構造の明示的な入力を行っているが，今後はユーザのブラウジング履歴から自動的に抽出することを検討している．また，ブックマークのフォルダ構造の一部分を入力としフォルダ名をキーワード，フォルダ構造を階層構造として考えれば，フォルダ構造の展開にも利用できると考えられる．

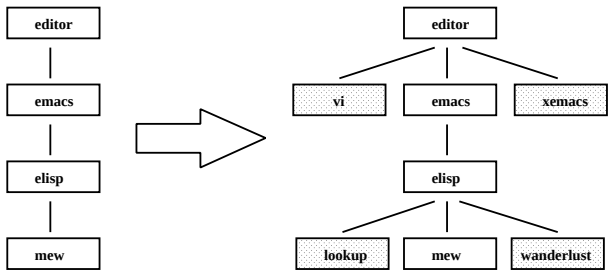

図 1: 階層構造の展開

2.2 システムの構成

ユーザがブラウジングによって獲得した情報を特徴付けるキーワードを階層的に構造化してものをシステムに入力すると，システムは最上位の階層を除いた各階層について，上位階層 d のキーワードと下位 階層 $d+1$ のキーワード間の関係を持つ新たな 階層 $d+1$ のキーワードを抽出し，階層構造に付け加えてユーザに提示する．図 1 に階層構造の展開について示す．

システムの構成を図 2 に示す．ユーザはブラウジングによって獲得した情報を特徴付けるキーワードを階層的に構造化してシステムに入力する．システムはキーワードに関連する Web ページを検索エンジンを用いて収集する．収集した Web ページから抽出規則を用いて，周辺に存在する情報を特徴付けるキーワードを抽出し，そのキーワードを入力された階層構造に追加して，新たな階層構造を生成する．階層的な知識の構築の詳細については次章で述べる．ユーザは入力したキーワードと追加されたキーワードとの関係を見ることで，ブラウジングで獲得した Web ページとその周辺ページから得られる情報の関係を効率的に理解することができる．

連絡先: 間瀬 心博，東京工業大学 総合理工学研究科 知能システム科学専攻，mase@ntt.dis.titech.ac.jp

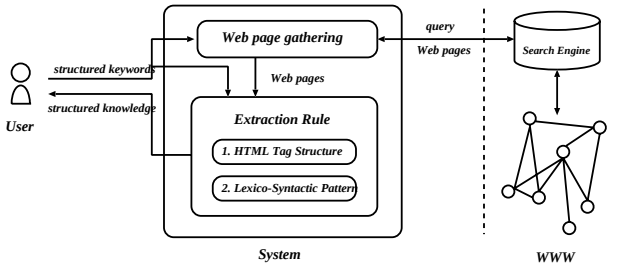

図 2: システムの構成

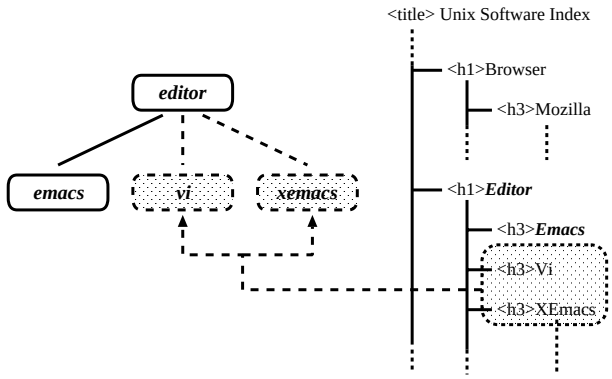

図 3: HTML タグ構造によるキーワード抽出

3. 階層的知識の構築

本研究では Web ページ集合から階層的知識を構築するために，ユーザから入力されたキーワードとその階層関係を利用する．階層関係を形成するキーワードのペアを抽出するために，キーワード間の上位・下位関係を示す構文構造パターンと，HTML ファイル中のタグ構造を利用した 2 つの手法を用いる．以下に，入力された上位キーワードに対応する下位キーワードの抽出手順を示す.

3.1 情報収集

キーワードの上位・下位関係を抽出するためにキーワードに関係する Web ページ集合を収集する．検索エンジンに 階層 d と 階層 $d+1$ のキーワードをクエリとして与える．得られたヒットリストから上位 100 ページを収集し，さらにそれらのページからリンクされているページも収集する．上位キーワードが複数の分野で出現するキーワードだとしても，下位キーワードを同時に与えることで特定分野に関連するページ集合を収集することになり，ノイズとなる無関係の分野のキーワードをあらかじめ除去できる．

3.2 構文構造パターンを利用した情報抽出

収集した Web ページから上位・下位関係を示す構文構造パターン [1] を探し出し，上位キーワードに対応する下位キーワードの候補を抽出する．構文構造パターンは以下のものを用いる．

1. NP such as $\{NP,\}$* {(or|and)} NP
2. such NP as$\{NP,\}$* {(or|and)} NP
3. NP $\{,NP\}$* {(or|and)} other NP
4. NP, including $\{NP,\}$* {(or|and)} NP
5. NP, especially $\{NP,\}$* {(or|and)} NP
6. NP is NP
7. NP has NP

1〜4 のパターンでは，上位，下位の両キーワードが同時に出現している場合にのみ，下位キーワードの候補を抽出する．例えば，上位キーワード「editor」，下位キーワード「emacs」の場合に,「... such editors as emacs, vi ....」からは「vi」が下位キーワードの候補として抽出される．

3.3 HTML タグ構造を利用した情報抽出

収集した Web ページの HTML ファイルを解析し，特定のタグに関する木構造を生成する．title，h1〜h5，li のタグに注目し，これらのタグでタグ付けされ文字列から下位キーワード候補を抽出する．生成された木構造の中に下位キーワードが出現するノードがある場合に，その親ノードの文字列に上位キーワードが含まれていれば，その親ノードから派生する他の子ノードの文字列中には下位キーワードの候補が含まれていると考え [2]，その文字列を抽出する．(図 3) これらの文字列から stopword を取り除いて，名詞のみを下位キーワード候補とする．

3.4 下位キーワード候補の選定

上述の構文構造パターンと HTML タグ構造の抽出規則を用いて，下位キーワード候補を含む文列を抽出する．これらの抽出した文字列を一文書として，文書に出現するキーワードに TFIDF による重み付けを行う．全文書から最も重みの高いキーワードを取り出し，頻度の高い上位数個のキーワードを選択する．これらのキーワードを 階層 d の上位キーワードに対応する新たな下位キーワードとする．

以上の手順を各階層ごとに繰り返し，新たな階層構造を生成する．

4. まとめ

ユーザにブラウジングで獲得した情報の特徴的なキーワードを提示してもらい，周辺に存在する関連情報のキーワードとの階層的な関係を提示するシステムを提案した．ユーザは獲得した情報の特徴的なキーワードをあらかじめ階層的に構造化してシステムに入力する．システムは構造パターンと HTML ファイルのタグ構造を利用した抽出規則を用いて関連情報の特徴キーワードを抽出し，入力された階層構造に抽出したキーワードを加えユーザに提示する．本システムを利用することでユーザは，未見の関連情報の存在を発見し，情報間の関係を理解することで効率的に情報収集が可能になると期待できる．

参考文献

[1] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. Technical Report S2K-92-09, 1992.

[2] Ellen Spertus, editor. *ParaSite: Mining Structural Information on the Web*. The sixth International World Wide Web Conference, 1997.