

ブラウジング履歴からの階層的知識の構成

間瀬心博[†]

山田誠二[‡]

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[†]
国立情報学研究所[‡]

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．ユーザが目的とする情報を持っている場合には，検索エンジンを用いた絞込みが可能である．ユーザは検索エンジンに目的の情報に関連するキーワードを入力することで，キーワードにマッチした Web ページの URL リストを得ることができる．これらの Web ページを起点にリンクをたどり目的の情報の含まれる Web ページを探索することになる．ユーザが収集できる情報は選択するリンクに依存しており，当然選択しなかったリンク先にある Web ページに含まれる情報を得ることはできない．ユーザにとって有用な関連情報が周囲に存在していたとしても，その情報の存在自体を発見することができない．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各 Web ページから得られ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になると期待できる．

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，さらに周囲に存在する未見の関連情

Hierarchical structured information from browsing history

[†]Motohiro MASE, CISS, IGSSE, Tokyo Institute of Technology

[‡]Seiji YAMADA, National Institute of Technology

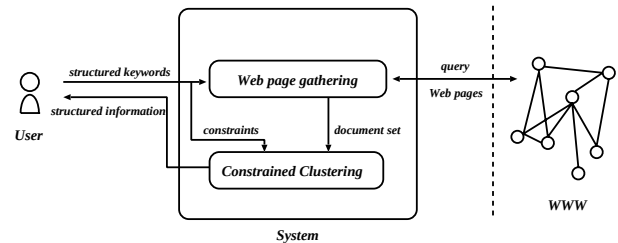


図1 システムの構成

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元にして，関連情報との関係を抽出し提示する．また，ユーザが効率的に情報獲得するにあたって，システムが提示する情報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，ユーザには獲得した情報を特徴付けるキーワードを用いた階層構造を指定してもらい，その階層構造を元にして周辺に存在する未見である情報のキーワードとの階層的な関係を抽出する．ユーザによる入力を制約として用いることで，階層構造の構築の精度をあげることが可能となる．詳細は 3.1 章で述べる．現時点では，ユーザによるキーワードや階層構造の明示的な指定を行っているが，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．

2.2 システムの構成

システムの構成を図1に示す．ユーザはブラウジングによって獲得した情報の特徴語を階層的に構造化してシステムに入力する．システムは特徴語に関連する Web ページを検索エンジンを用いて収集する．さらに収集した Web ページ集合について各ページに出現する単語の共起頻度を用いたクラスタリングを行い，階層構造を抽出する．階層的に構成された各クラスタには，クラスタを構成する Web ページ集合を特徴づけるキーワードと，そのキーワードについて情報を得るのに適した Web ページ集合が関連づけられている．各クラスタに関連づけられたキーワードと

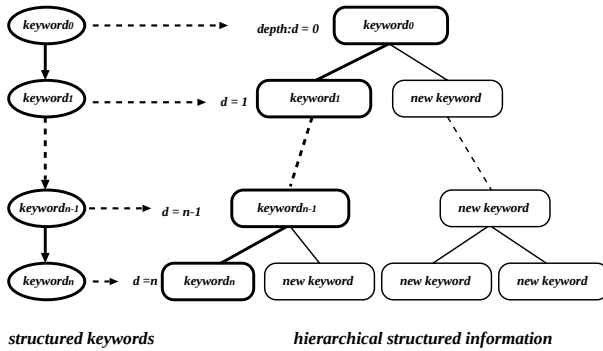


図2 クラスタリングの制約条件

そのクラスタ間の関係から，ユーザの閲覧した Web ページと周辺ページから得られる情報の階層的な関係を得ることができる．

3 階層的知識の構築

本研究では階層的知識を抽出するために，ユーザから入力されたキーワードとその階層構造を制約として用いる．これは，抽出される階層知識の上位下位関係の矛盾をできるだけ減らすためである．

3.1 制約条件

システムに入力されるユーザ指定のキーワードとその階層構造は，ユーザがブラウジングした後に明示的に指定されるものであり，妥当なものであると考えられる．この階層構造を反映しかつ周辺に存在する Web ページから獲得できる情報との関係を階層構造として構築することが望ましい．図2にユーザの指定による制約条件と構築されるべき階層構造の関係を示す．

ここでは，入力されたキーワードの階層構造を以下の制約とみる．

- 深さ d の階層のクラスタには，キーワード d によって特徴付けされるクラスタが必ず 1 つ存在する．
- そのクラスタの子クラスタには キーワード $d+1$ で特徴付けられるクラスタが存在する．
- 構築される階層構造の深さは入力されたキーワードの階層構造と同じ深さとなる．

3.2 階層構造の構築

本研究では一般的な階層的クラスタリング [1] を行わない．各階層ごとにクラスタリングの対象となる Web ページ集合を変更し，かつ 3.1 で示した制約を

満たすように再起的にクラスタリングを行う．以下に手順を示す．

1. 分割するクラスタを選択する．

- a) 分割するクラスタがユーザ指定のキーワード n で特徴付けされたクラスタの場合には，キーワード n ，キーワード $n+1$ をクエリとして検索エンジンで検索をかけ，得られたヒットリストの上位 N 個の Web ページを収集し，この Web ページ集合 D をクラスタリングの対象とする．
- b) 特徴キーワードがクラスタリングによって決定されたクラスタの場合には，分割するクラスタの特徴キーワードをクエリとして検索エンジンを用いて Web ページ集合 D を収集する．

2. 収集した Web ページ集合 D の各ページについて出現する語の共起頻度を計算する．分割するクラスタを特徴付けるキーワードと共起する語の出現頻度に注目した各 Web ページのドキュメントベクトルを生成する．

3. k-means 法によるクラスタリングを行い， k 個のクラスタに分割する．ここで，a) の場合は，キーワード $d+1$ が特徴キーワードとなるクラスタが生成されるように制約をかけてクラスタリングを行う．

4. 深さが d になるまで再起的にクラスタリングを行う．

4 まとめ

ユーザがブラウジングによって獲得した情報をもとに，周辺に存在する関連情報をとの階層関係を提示するシステムを提案した．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] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques, 2000.