D-10 反復型情報検索のためのウェブサイトの多段階要約手法

A Multi-Stage Summarizing Method of Web Sites for Repetitive Web Searches

李 龍　　上林弥彦
Ryong Lee　　Yahiko Kambayashi

京都大学情報学研究科社会情報学専攻

1. はじめに

従来のウェブ情報検索エンジンの主な目的はユーザが与えた質問に対して適した情報を含むページを上位のランキングとして出すことであったが、実際にこのように個々のユーザ質問を扱う個別的な情報検索方法とともに、検索結果から関係ある他の情報を繰り返して探すことを支援する**反復型検索方法**が重要である。現在の主なウェブサイトでは全文検索まで可能な機能を備えて、ユーザが質問を文字列として毎回繰り返して出して情報を探すようにしている。あるいは、関連あるページへのリンクをページごとにつけることによってユーザが関連内容を見るようにしていることもある。しかし、いずれにしてもユーザの質問作成のためのインタラクションとウェブサイトの膨大な情報を観覧して重要な内容を要約する作業は減らないという問題点がある。このような反復型情報検索を支援するために、各ウェブサイトの管理者としては自分のサイトの全内容からもっとも重要な内容の要約とそれに関わるページへのリンクを周期的に管理する必要がある。しかし、その管理作業はサイト中のユーザの数や全般的なウェブの規模が大きくなるにつれ、そのような要約を手で作ることはコストがとても高くて困難な作業となる。さらに一人の個人が作成すると客観性やその記述のレベルを維持することは難しくなる。従って、**ウェブサイトの要約を訪問したユーザの目的に応じて動的に生成**する必要がある。本稿では、ウェブサイトが含んでいるページ集合からユーザの目的に応じて要約を概念ネットワークとして提供する手法を提案する。

2. ユーザ要求に応じたウェブサイトの動的な要約機能

あるウェブサイトを反復型情報検索の目的で訪問したユーザがそのウェブサイトに含まれる情報を検索する時、最初からウェブサイトに対する詳しい内容よりも、最初は簡単な要約から次のステップでは興味のある部分に対して詳しい内容を知ろうとするだろう。その行動を次の三つの段階として分けて行うことが考えられる。

① **ウェブサイト全体のイメージの把握**：ターゲットページサイトからどういう知識が得られるかを知りたい。
② **イメージの拡張**：ユーザにとっては最初に与えられたウェブサイト全体のイメージをもっと詳しくみたい。すなわち、要約結果をより大きくする必要がある。
③ **特定の部分に関する詳細化**：上記の段階で与えられた要約からユーザの興味がある部分にフォーカスして要約を見たい。

以上のように、ウェブサイト自体が持っている知識をユーザにどう与えるかが課題となる。ウェブサイトの要約作成に対する我々のアプローチは、ウェブページ集合から概念と概念間の関連によるネットワークを構築することである。たとえば、京都に関する多くのウェブサイトからは{"京都"}, {"観光"}, {"銀閣寺"}のような概念がよく表れる。 その間には{"京都"}→{"観光"}、{"観光"}→{"銀閣寺"}のような概念関連があると考えられる。さらに、それらを結合すると{"京都"}→{"観光"}→{"銀閣寺"}のような総合的な概念ネットワークを構成することになる。このような要約作成方法はウェブページ中に隠れている知識を利用するという特徴がある。そして、それぞれの関連は上記の三つのステップの要求から動的に生成・結合されて概念ネットワークを構築している。本稿では、上記のようなユーザの目的に応じたウェブサイトの要約をそのイメージのレベルと詳細度による概念ネットワークとして提供する方法を提案する。

3. 概念ネットワークの構築

ウェブページからこのような知識構造を作るために従来のデータマイニングの手法を適用している。本稿で目的としている概念ネットワークは、図1のように現在のウェブページの内容からよく現れる連想ルールを導いてそれによる結合構造をしている。特に、その連想ルールは、A→B の型で A が現れるページから B が現れる比率を計算する。ここでは、A,B が一つの語である場合について考える。そして、ターゲットするデータセットから A と B が一緒に現れる比率をこのルールの Support、そして A が現れた時に B が現れる比率をこのルールの Confidence と定義する。その概念構築は次のステップによって行う。なお左辺が複数である場合もほとんど同様である。

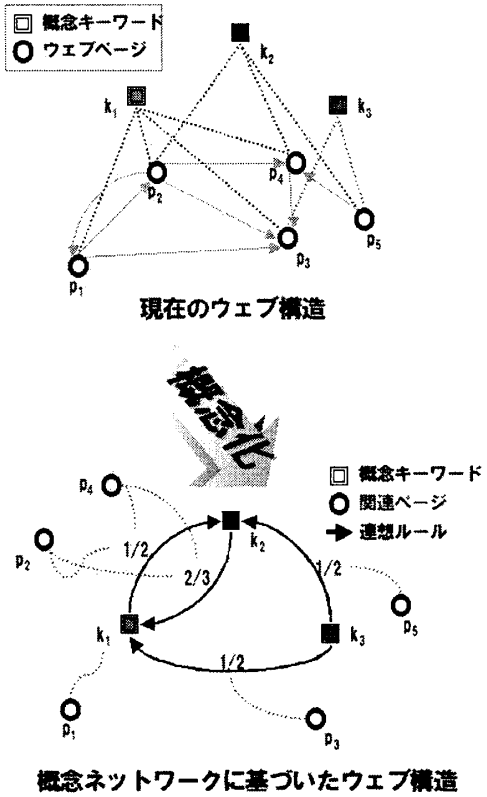


図１．ウェブ構造の概念化

① **概念抽出対象ページの収集**:ウェブは一般的に非常に異質な内容を含めているので、連想ルール抽出での Support をあげるためには、類似性が高いページ集合を絞る必要がある。

② **概念キーワードの抽出**:上記のデータセットのようにウェブページからの一つページ中で現れる単語の数が予測できないほど大きくなる場合があるなど、ページごとの中心的で代表的なキーワードを絞って計算しないと計算時間と効率の問題が起こる。そのため、一つのウェブページに対応する概念キーワードを、TF/IDF 手法によって計算して各ページで tf•idf の値が高い n 個までの特徴単語を選択する。

③ **連想ルールの計算**:ターゲットするウェブページ全体のセットからよく現れるルールを導くために、Apriori[1]アルゴリズムを適用する。なお、Apriori 計算の結果は、{'京都','観光'}→{'銀閣寺'}のような、ルールが三つ以上の要素から構成されることまで計算される。本稿では簡単に二つの要素からもの構成されるルールについて説明する。

④ **要素ルールから概念ネットワークの構成**:上記のステップまでの結果の連想ルールはそれぞれ異なる Support と Confidence を持っているので、概念ネットワークにおけるこれらの結合の制約に対する考慮が必要となる。要素ルールに対する Support が高いほど全般のデータセットに対するそのルールの出現度が高いということで、この値が高いルールから構成される概念ネットワークはデータセット全体に対するおおまかなイメージを提供する。Confidence はルール間の関連の強さを表すので概念ネットワークでの各概念における関連概念の数をどの程度にするかを決める尺度となる。従って、概念ネットワークの構成には、要素ルールからどの程度のイメージと関連を求めるかを決める必要がある。

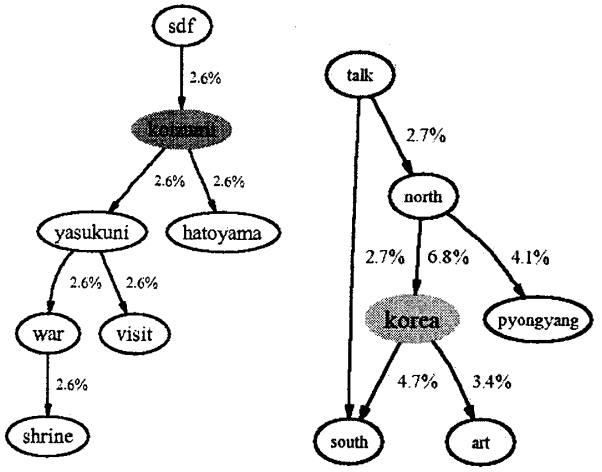
上記の各ステップにおいて、対象ページ集合、概念として使える単語、有効な要素ルール、ルールの結合条件を考慮している。これらの制約によって異質的なウェブ情報空間から知識を有効に取り出すことが可能である。

4. ウェブの実データを利用した実験

概念ネットワーク構築の実験では、データセットとして日本の朝日新聞の英語ページサイト[2]と韓国の Korea Herald 英語新聞サイト[4]からページを収集した。そして、ステップ②による各ページの概念キーワードのインデックスを生成して、ステップ③における連想ルール抽出を可能にした(ストップワード処理はこの段階で行った)。実際の連想ルール抽出では、全データセットから任意の100ページを選び最小 Support は、1.0%から 5.0%の範囲で、最小 Confidence は 80%に固定した条件で行った。これにより、80%のルールに対する確信度で、ルールの数を調整できるようになった。その結果としては次の三つにグループとして分けられた。

Group A: 約 1%以下の Support で数千個のルールが出現
Group B: 約 1.5-2.5%までは２５～80個程度のルールが出現
Group C: その以上に対しては、数個以下のルールが出現

個々の連想ルールは知識を与えるには十分でないが、このようにルールの結合によって図2のように概念の全般的な関連がわかるようになる。しかし、そのコンポーネント中には、ノードが二つしかないようなものがあり、例えば、{'yasukuni'}→{'visit'}のような関連は新聞記事と関連した何かの知識を連想させるが情報としての価値が低くなるという問題があった。そのため、少



(a) 朝日新聞英語サイト　(b) Korea Herald

図２．新聞社のウェブサイトからの概念ネットワークの例

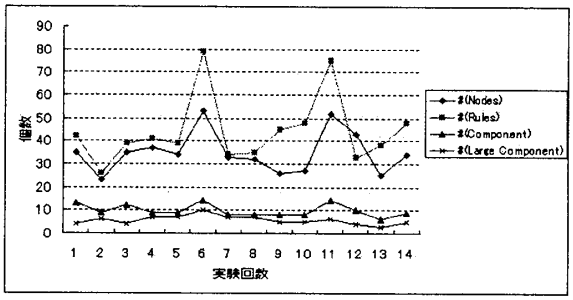


図3. Group B (最小 Support: 1.5～2.5%)の実験結果

なくとも三つ以上のノードからなるコンポーネント(Large Component)が概念ネットワークとして意味を持つようになると仮定して、図3の#(Large Component)/#(Component)の比率を比較したら、約 60%程度のコンポーネントが概念ネットワークを構成できた。すなわち、3章で提案した方法とコンポーネント制約によって概念ネットワークを十分に構築できるようになった。

5. おわりに

本稿では、反復型ウェブ情報検索を支援するために概念ネットワークを動的に構築してウェブサイトの要約としてユーザに提供する手法を提案した。その要約は Support と Confidence という尺度によって変えることができる。実際にウェブは多くの知識を含んでいてそれをどう抽出して利用するかはウェブ利用の重要な側面の一つである。我々の以前の研究でも現実空間に対する人間の知識抽出を連想ルールとして表現した[3]。それに加えて概念ネットワーク構築はウェブから知識を掴むのに非常に有効な方法として使える。なお、本研究は科学技術振興事業団(JST)・戦略的基礎研究推進事業(CREST)における「デジタルシティのユニバーサル」プロジェクトの支援によって行われた。

参考文献

[1] R. Agrawal and R. Srikant. "Fast algorithm for mining association rules," In Proc. of the 20th VLDB Conference, pages 487-499, Santiago, Chile, 1994.

[2] 朝日新聞英語サイト(2001/6/26～2002/6/10: 9,766pages) http://www.asahi.com/english

[3] Y. Kambayashi, R. Lee and T. Tezuka, *"Generation of Location-Related Knowledge from Web Contents,"* NSF-OntoWeb Invitational Workshop on DB-IS Research for Semantic Web and Enterprises, Atlanta, April 2002.

[4] Korea Herald 新聞サイト(2000/10/7~2002/6/10: 5,006pages) http://www.koreaherald.co.kr