部分グラフを基本単位とするWeb文書群の検索モデルと自動分類について

4L－2

段 一為　佐野 綾一　波多野 賢治　田中 克己
神戸大学大学院自然科学研究科

1 はじめに

WWW 上で Web 文書を検索するために、現在、様々な検索エンジンが利用されている。この際、質問として入力されたキーワード全てがページ内に存在すれば検索は可能だが、該当する内容が複数のページにまたがっている場合も考えられ、そのような場合にはそのページ群を検索することができないという問題がある。

このような背景から、リンクでつながった関連のある一連のページの中にキーワードが分散して現れている場合も検索可能とするために、Web 文書のリンク構造をグラフ構造とみなして[1]、ページ単位の検索ではなく複数のページ、すなわち、問い合わせに対する極小部分グラフを検索の基本単位として検索・分類するためのモデルを提案する。

2 極小部分グラフを基本単位とする検索モデル

本研究では Web 文書を扱うため、問い合わせを行うデータベースを $DB = \{U_1, U_2, \cdots, U_l\}, (l \geq 1)$ と定義する。ただし、各 $U_i (i = 1, \cdots, l)$ は各 URL をノード、URL 間をつなぐリンクを有向枝とする連結グラフであり、Web 文書の特徴を考えれば、$U_j \cap U_k = \phi, (j, k \in \{1, \cdots l\})$ となる。次に、このデータベースに対する問い合わせ $Q = k_1 \wedge \cdots \wedge k_m (m \geq 1)$ の解として $G = (V, E)$ を以下のように定義する。ある $i \in 1, \cdots, l$ に対して $U_i \supseteq G$ が成り立ち、G にはキーワード $k_1, \cdots, k_m$ が全て出現する。さらに、$k_1, \cdots, k_m$ を全て含む、$G' \subset G$ なる G' が存在するとき、G は $k_1, \cdots, k_m$ を全て含む極小部分グラフである(図1参照)。

3 極小部分グラフの評価尺度

極小部分グラフを問い合わせの解と考えた場合も、その解として複数の極小部分グラフが返される。よって、それら極小部分グラフのスコアリングを行うために評価式を定義し、それにより極小部分グラフを評価する必要がある。しかし、極小部分グラフを評価する際には、以下の点を考慮する必要がある。

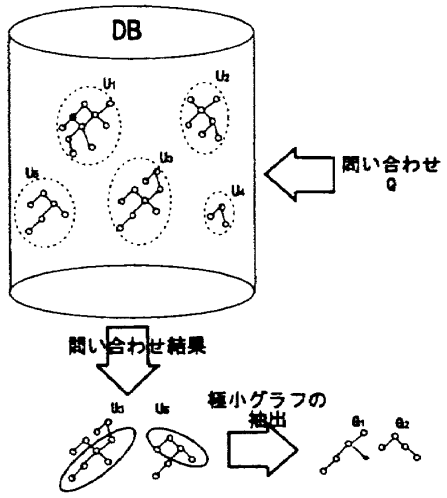

図 1: 検索モデルの事例

- グラフの直径。ここでいう直径とはグラフの最長パスのノード数を表す。
- ノード自身が持っている問い合わせ Q にマッチするキーワード数。

以上のことを考慮し、本研究では次の評価式を提案する。

$$\text{評価尺度}1 = \frac{1}{n'} \tag{1}$$

$$\text{評価尺度}2 = \frac{1}{\sum_{v \in V_i} \frac{m}{n \cdot K(v)}} \tag{2}$$

ただし、n は極小部分グラフ中のノード数、n' はグラフの直径、m は問い合わせ Q のキーワード数、v は直径上のノード、V_i は極小部分グラフのノード集合、$K(v)$ は v における Q にマッチするキーワード数を表す。ただし、式2において、$K(v) = 0$ の場合は $\frac{m}{n \cdot K(v)} = C (C \geq 1$ である定数$)$ とする。

式1は、グラフの直径のみを考慮した評価式であるから、図2のようなスコアとなるが、図中の (3)、(4)、(5) のようにグラフの構造が同じであるが、ノードにマッチするキーワードが異なるような場合でも同じスコアが与えられてしまう。そこで、各ノードに問い合わせキーワードがどのような割合で割り当てられているかを評価し、式2を定めてある。ただし、図中の (4)、(5) のようにノイズのノードが含まれている

"A Query Model and Automatic Classification for Web Documents Based on Minimal Subgraph structures As Retrieval Units"
Yiwei Duan, Ryouichi Sano, Kenji Hatano, and Katsumi Tanaka.
Graduate School of Science and Technology, Kobe University.

極小グラフ	評価尺度1のスコア	評価尺度2のスコア
(1) k1 k2 k3	1	1
(2) k1 k2 k3	1/2	1/2
(3) k1 k2 k3	1/3	1/3
(4) k1 k2 k3	1/3	2/5
(5) k1 k2 k3	1/3	1/2

図 2: 評価の一例

場合も考えられ定数 $C(C \geq 1)$ を定めている。C はこれらノイズのノードをどう扱うかにより変化させることができるが、図2では $C = 1$ として多少のノイズに対しては無視している。

4 検索モデルの実現

本システムを実現するためには、具体的に以下の手順で極小部分グラフを生成する必要がある。

1. $Q = k_1 \wedge \cdots \wedge k_m$ を検索エンジンの問い合わせに変換する。具体的には $Q' = k_1 \vee \cdots \vee k_m$ として、検索エンジンに対して OR 検索を行う。
2. 検索された Web 文書それぞれのリンク構造を調べ、部分グラフ U_i を生成する。
3. 生成された部分グラフ U_i の中から極小部分グラフを生成する。例えば、問い合わせ $Q = k_1 \wedge k_2$ により、図3のような部分グラフ U が与えられた場合、部分グラフ U からは(1)、(2)、(3)、(4)、(5)の極小部分グラフが得られる。

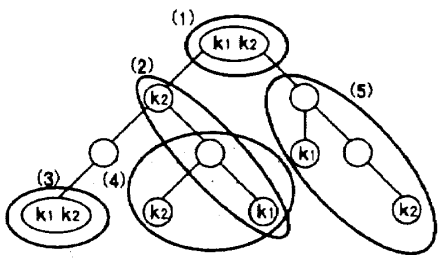

図 3: 極小部分グラフの切り出し

5 極小部分グラフの自動分類

以上のように導出した極小部分グラフを自動分類するために、本研究では自己組織化マップ (SOM)[2] の利用を考えている。

まず、各極小部分グラフの各ノード h_l ごとに特徴ベクトルを生成する。その手順は以下のようになる[3]。

1. 全ての極小部分グラフを構成する各ノードから取り出された単語の出現頻度を調べる。
2. 出現頻度の高い順に上位 500 の単語 $(t_1, t_2, \cdots, t_{500})$ を特徴ベクトルの要素とする。
3. 要素 $(t_1, t_2, \cdots, t_{500})$ の各ノード h_l での出現頻度 $(f_1, f_2, \cdots, f_{500})$ を求め、各マッチノード内の全出現単語数 N_l で正規化することによって、各マッチページのベクトル

$$\boldsymbol{F}(h_l) = (\frac{f_1}{N_l}, \frac{f_2}{N_l}, \cdots, \frac{f_{500}}{N_l})$$

が生成される。

こうして生成された各ノードの特徴ベクトルを利用して極小部分グラフの特徴ベクトルを生成する。ここでは単純に極小部分グラフに含まれている各ノードの特徴ベクトルの和を極小部分グラフの特徴ベクトルと考えることにする。もちろん、この特徴ベクトルの生成方法には様々なバリエーションが考えられ、前述したようにグラフの直径や、各ノードにマッチしたキーワード数を考慮する必要がある。

6 おわりに

本研究では極小グラフを無向グラフとして扱ったが、Web 文書は本来有向グラフであり、非連結グラフの場合も考えられる。さらに、同一キーワードが多数極小部分グラフに現れる場合のグラフの評価など様々な問題が挙げられる。今後はこれらの問題点を考慮した評価尺度を考え、実際に極小部分グラフを分類するシステムを構築していく。

謝辞 この研究は，一部，日本学術振興会未来開拓学術研究推進事業における研究プロジェクト「マルチメディア・コンテンツの高次処理の研究」および文部省科学研究費重点領域研究「高度データベース(No.275)」(課題番号 08244103) による．ここに記して謝意を表す．

参考文献

[1] K. Tajima, Y. Mizuuchi, M. Kitagawa, and K. Tanaka. "Cuts a Querying Unit for WWW, Netnews, and E-mail" *In Proc. of 9th ACM Conference on Hypertext and Hypermedia*, pp.235-244, Jun. 1998.

[2] T. Kohonen. The self-organizing map. *Proceedings Of The IEEE*, Vol.78, No.9, pp.1464-1480, 1990.

[3] 佐野雅一, 波多野賢治, 田中克己. 「自己組織化マップを用いた Web 文書の対話的分類とその視覚化」. 情報処理学会研究報告, Vol.98, No.57, 98-DBS-116-5, pp.33-40, 1998年7月.