データベースシステム 113−47
（1997. 7. 16）

グラフ構造を持つデータのカット分割に基づく検索

水内 祥晃† 田島 敬史‡ 田中 克己*
†神戸大学大学院自然科学研究科情報知能工学専攻
‡神戸大学工学部情報知能工学科
*神戸大学大学院自然科学研究科情報メディア科学専攻
{mizuuchi,tajima,tanaka}@in.kobe-u.ac.jp

本稿では、グラフ構造を持つデータをデータベース化する際に、検索の単位として、個々のノードではなく、カットという意味的につながっているノード群を用いる手法について提案する。例えばあるネットニュースの記事群をデータベース化する場合、特定の記事ではなく、ある話題について検索したい場合が多い。よって、個々の記事を検索単位とするよりも、同一の話題を論じている一連の記事群を一つのカットとし、これを検索の単位とする方がより適切である。同様に、WWW のページを検索する場合も、個々のページではなくある意味的なまとまりを検索の単位とした方が効果的である。そこで、本稿ではこれらのデータの検索に対して、カットの概念を適用する方法について述べる。

Retrieval for Graph Structured Data based on Cut Partitioning

Yoshiaki Mizuuchi† Keishi Tajima‡ Katsumi Tanaka*
†Division of Computer and Systems Engineering,
Graduate School of Science and Technology, Kobe University
‡Dept. of Computer and Systems Engineering,
Faculty of Engineering,Kobe University
*Division of Media and Computer Sciences,
Graduate School of Science and Technology, Kobe University

This paper proposes a method to construct a database of graph structured data. We divide a graph into several *cuts*, which are connected subgraphs consisting of correlated nodes, and we use a cut as a unit of query instead of each node. For example, in a database of Netnews articles, we usually want to query not a specific article but some topic. In such a case, it is appropriate to use cuts consisting of consecutive articles discussing the same topic as a data unit. Similarly, when querying WWW pages, it is more effective to regard a set of correlated pages as a data unit. In this paper, we explain how to apply the concept of cuts to the query for those graph structured data.

1 はじめに

インターネット上に流れている、電子メイルやネットニュース、WWW ページなどの情報の特徴の1つとして、メイルやニュースでは議論の中での引用関係、WWW ページではハイパーリンクによって、グラフ構造を構成することが挙げられる。これらをデータベース化し検索を行う場合、何を検索の単位とするかが問題になる。一番単純な方法は個々のノードを検索の単位とする方法だが、一通のニュース記事や一つの WWW ページの大きさは、短い物から長い物までばらつきが大きく、検索の単位としては短すぎる場合も多い。そこで、単純に個々のノードを検索の単位とするよりも、意味的につながりのあるノード群からなるあるまとまりを考え、それを単位として検索を行う方が有効である。そこで本稿では、この意味的なまとまりのある単位を表す「カット」という概念を導入し、このカットを単位として検索を行う枠組について述べる。

そのような枠組の例として、まずネットニュースの記事をデータベース化する場合について述べる。大量のニュース記事群の中から、ある特定の話題に関して検索を行う場合、検索結果として個々のページが返されるよりも、その話題に関する議論の流れ全体をつかむことができるように、関連する一連の記事全体を返して欲しい。さらに、複数のキーワードで検索をかけるような場合、それら全てを同時に含む記事があれば良いが、一連のやり取りの中で複数の記事に分散してそれらのキーワードが現れているような場合は、個々の記事単位の検索ではうまく検索できない。そこでこの研究では、同一の話題に関して議論している一連の記事群を検索の単位とする。

次にカットという概念を、WWW のページ群にも適用する。現在、Alta Vista など様々な検索エンジンが WWW 上に存在している。これらは収集した WWW ページに関するインデックスを作成し、与えられたキーワードを含むページを検索してくれる。しかしこれらのインデックスはページ1枚毎に対して作成されているため、ニュース記事の場合と同様に複数のキーワードによる検索を考えた場合、それらのキーワード全てがたまたま同じページにあれば良いが、ある内容が複数のページにわけて記述され、与えられたキーワードがこれらの複数のページに散らばって現れる場合は、うまく検索できない。そこで、インデックスを作る際に、意味的にまとまりのあるカットを単位として作成すれば、そのような問題は解決される。本研究ではそれを実現する方法の1つとして、既存の検索エンジンを利用して、これに wrapper をかけるという形を提案する。

以降、第二節でカット分割されたネットニュース上の記事に対する検索、ついで第三節で WWW ページのカット分割と検索について述べ、第四節でまとめとする。

2 カット分割されたネットニュース上の記事に対する検索

ニュース記事群のカットへの分割の方法としてまず考えられる物に、同一のサブジェクトを持ち、かつ引用関係のあるメイル群を一つのグループとする方法がある。我々はこのようなメイル群をスレッドと呼ぶが、一つのスレッド内には話題の転換があったにもかかわらず、同一のサブジェクトのままになっている場合や、逆に話題の転換点ではないにもかかわらず、サブジェクトが変更される場合がある。

そこでカットという概念を導入し、その話題の転換点を検出する。そして、記事群を意味的にまとまりのある単位に分割し [1]、それらを検索の単位とする検索機構を提案する。

2.1 カット分割

カットの検出は、特徴ベクトルという記事の特徴を表すベクトルを生成し、それを利用して類似度を算出することによって行っている。ここではまず特徴ベクトルの生成方法について述べ、さらにそれらを用いた類似度判定について述べる。

2.1.1 特徴ベクトル

記事の本文を、奈良先端大松本研究室で開発された、日本語形態素解析システム『茶筌 (ChaSen)[2]』を用いて形態素解析し、その結果出力されたものから名詞を抽出、各名詞当たりの出現頻度をカウントする。特徴ベクトルはそのデータを元に生成した空間ベクトルのことである。特徴ベクトルの各要素の値 v_i は

$$v_j = tf_j \cdot idf$$

$$= \frac{w}{W} \cdot \log \frac{N}{n}$$

によって算出される。

tf_j(term frequency) は、その記事(群)中の単語 jの相対出現頻度、つまり、Wは特徴ベクトルを作成する単位内の名詞数、wはその特定のキーワードの単位内での出現頻度である。そして、Nは総記事数、n はその単語を含む記事数である [3]。また、この特徴ベクトルはカット及び記事単位でそれぞれ作成する。

2.1.2 記事の類似判定

類似判定を行なう記事を M_1、M_2とし、それぞれの特徴ベクトルを$\vec{v_1}$、$\vec{v_2}$とすると、その2つのベクトルの間の類似度 $A_{\vec{v_1},\vec{v_2}}$は以下の式で算出される [3]。

$$A_{\vec{v_1},\vec{v_2}} = \frac{\vec{v_1} \cdot \vec{v_2}}{|\vec{v_1}||\vec{v_2}|} \quad (1)$$

比較を行なう記事の両特徴ベクトルの内積を取るという操作により、頻度情報により重みづけされた各ベクトルの要素を掛け合わせた値の和が算出される。この値を類似度とする。

そして類似しているか否かという判定は、同じレベルでの比較の時には J_b、参照／被参照関係の記事の組の時は J_pという閾値を用いる。閾値より大きければ、その二つの記事間は類似性が高いということである。ここで J_b、J_pは任意の値で、$(0 \leq J_b, J_p \leq 1)$ を満たす。

2.1.3 カット検出

記事の参照関係は木 (tree) で表すことができる。

カットの検出は、葉と、葉が引用した記事の間の検査から始め、順に根に向かって同様の検査を繰り返すことで行なう。手法は以下に示す。

カット検出の手法

1. カット検出を開始する記事（出力枝を持たないもの）を選択し、それのみを構成要素とする暫定のカットを定義する。
2. その暫定のカットと同一レベルに位置する（同じ記事を参照して記述された記事を構成要素として含む）他の暫定のカットの特徴ベクトルを逆横型探索[1]によって探索し、それぞれの暫定カットの間で類似度を算出する。もし同レベルにどの暫定カットにも所属していない記事が存在すれば、その記事の所属暫定カットを先に決定する。
3. 類似度の高かった暫定カット同士を合併し、それを1つの暫定のカットとして再定義する。この際、暫定カットの存在するレベルに位置していた記事を含んでいるものが複数あれば、その暫定カット同士は合併は行わないこととする。
4. 検出を行っているレベルにあるそれぞれの暫定のカットとその上位の記事の間で類似判定を行う。
   (a) 類似していれば、その暫定のカットの特徴ベクトルに重み付けしてその上位の記事の特徴ベクトルに加え、暫定のカットにこの記事を追加する。
   (b) 類似していない時、暫定のカットの構成要素に暫定のカットの位置するレベルよりも下にある記事が含まれていれば、それを除外し、そのすぐ下のレベルのみの構成要素の特徴ベクトルを再生成し、それと上位の記事の特徴ベクトルとで類似度算出を行う。
      i. 類似していれば、先ほど再生成した同じレベルの記事のみで構成されるカットとその上位の記事で暫定のカットを生成し、また今まであった暫定のカットも存続する。
      ii. 類似していなければ、今ある暫定のカットは正式なカットとなり、別にその上位の記事のみを構成要素とする暫定のカットを生成する。
5. 2へ戻る。

[1]グラフの探索法の一種で横型探索があるが、本稿で行う探索は横型探索とは異なり、根からではなく葉から行うものであるから、我々はこの探索法を逆横型探索と呼ぶ。

2.2 検索機構

先程述べた手法により、記事群はカット単位に分割されている。ネットニュースやメイリングリストなどでは、そのとき行われている議論が既に行われていたということがある。その際、ユーザの「この議論が既に行われていたかどうか？」という質問に対して、単一の記事を返すような検索機構では十分とは言えない。そこで、単一の記事ではなくカットを返す検索機構を提案する。まず、記事群に対してどのような質問が与えられるのかについて述べる。

- 文字列 (or, and を含む) を与える
  これは普通の検索エンジンなどで検索をする方式と同じである。
- 記事を与える
  “こんな記事が新しく来てるけど、以前にこの内容について議論が行なわれていないか？”という問いかけに対処できる。呈示する記事の特徴ベクトルと各カットの特徴ベクトルとの類似度判定によって実現できる。特徴ベクトルは、与えられた記事と、すでにあった記事群とを統合して作成し直す。
- カットを与える
  質問を投げかける記事群とは別に、すでにカット分割が行われているスレッドで、その中のあるカットに対応する議論があるかどうかといった質問に対処できる。

図 1は実装画面である。システムの流れを以下に示す。

1. まず記事群が 2.1 節で述べた手法によりカット分割される。
2. ユーザは、自分が検索したい方式 (文字列、記事、カット) を選択する。
3. カットに対して検索がかけられ、マッチしたカットが呈示される。

このように、ただ単に１つ１つの記事に対して特徴ベクトルを比較し、類似度で判定するよりも、同一の話題性を持つ一連の記事群であるカットとし、これを検索の単位とする方が、より効果的にユーザの質問要求を満たすことができる。

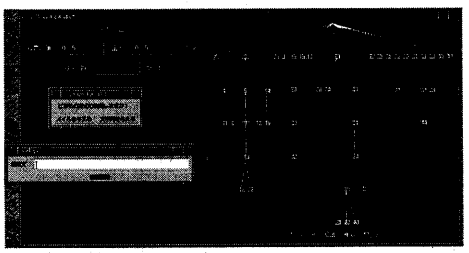

図 1: ネットニュース組織化システム (検索機能付き)

3 WWW ページのカット分割

一般の検索エンジンでは、検索結果として呈示される情報はページの名前や URL などのごくわずかな情報であるため、本当にそのページがユーザにとって必要かどうかを知るためにはユーザが実際にリンクをたどる必要がある。しかし、出力結果が大量である場合、１つ１つリンクをたどって行く作業は困難であり、もしも本当に必要とするページが後に呈示されていたとしても、見つけることはほとんど不可能である。

そこで、検索結果の大量のページを、お互いに類似しているもの同士でグループ化して呈示してやることにより、ユーザが効率良く必要なページを見つけられるようにしたい。そのようなグループ化を実現する手法として、ユーザのアクセスパターンのログを取り、それに基づいてグループ化を行うという手法が提案されているが [4]、ここでは、WWW ページの本文中の単語の出現頻度に基づいたグループ化を行う。しかしその際に、検索結果として呈示されたページの内容だけでは類似度を判定するには情報量が少なすぎる場合があるので、これらのページからリンクをたどり、そのページを含む意味的にまとまりのあるカットを作成し、このカット全体の情報を使って類似度を判定することにする。

3.1 検索エンジンの出力結果に対するカットの適用

以下にシステムの流れを示す。

1. まず、検索結果から得られる URL の情報を用いることによって、明らかに同一作者のものであるページについては、予めまとめておく。
2. 検索エンジンより得られたページ１つ１つについて、そのページを含む、意味的にまとまりのあるページ群 (カット) を生成する。カット生成方法については後述する。
3. その結果得られたカット同士の特徴ベクトルを比較し、類似性があるものを１つのカットとして呈示する。

3.1.1 カット分割

ネットニュースの記事のカット分割の場合、ある１つの記事をルートとして全体の記事数が決まっているため、特徴ベクトルを相対的に作成し、それらを基に類似度判定を行うことができた。

しかし、WWW ページの場合はリンクをたどることにより、無数のページがでてくる可能性がある

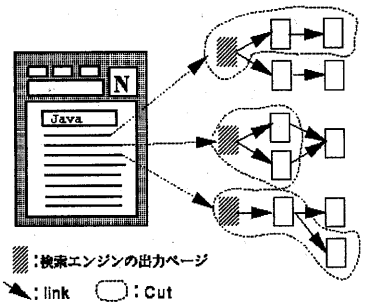

図 2: 検索エンジンの出力結果に対するカットの適用

し、また今回の場合、ルートが含まれるカットだけを検出すればよい。したがって、全体の数を確認し特徴ベクトルを算出して、類似度判定を行うということはせず、1回類似度判定を行う度に対象となるカットとページで特徴ベクトルを生成し直し、それを基に類似度判定を行うものとする。また、記事群のカット検出は、逆幅優先型探索により行っていたが、WWW ページのグラフ構造の場合、葉と葉の関連性よりも、葉とその根との関連性が強いことから、深さ優先探索で行うことにする。以下にカット検出の手法を示す。

カット検出方法

1. 検索エンジンの出力に得られたページのみを構成要素とする、暫定カットを定義する。カット検出はこの暫定カットをルートとして行われる。
2. 暫定カットと、その暫定カットからリンクされているページとで特徴ベクトルを生成し、類似度判定を行う。もし類似度が高ければ、対象となったページを暫定カットに加える。
3. 先程対象となったページからさらにリンクが張られているページがある場合は、暫定カットとそのページとで更に類似度判定を行う。ただしこの時、ルートからのカット検出で、既に対象となったページへのリンクは無視する。つまり、逆戻りのリンクなどは考慮にいれないことにする。
4. 以降同様にして、暫定カットとリンク先のページ間で類似度が低いと判定され、カットが検出されるまで、深さ優先探索を続ける。
5. 再びルートに戻り、まだ探索が行われていないリンク先に対して、1からカット検出を行う。
6. すべてのページについてのカット検出終了後、それらで再び特徴ベクトルを生成し直し、類似性のあるものをまとめて呈示する。

実際のカット検出例

構成要素が m, n である暫定カットを $\{m, n\}$、その特徴ベクトルを $\vec{v}\{m, n\}$ と表す。例として、図3のような場合でカット検出行う。

1. カット検出が開始されるページを a とし、その暫定カットを $\{a\}$ とする。
2. $\{a\}$ と b で、それぞれ特徴ベクトル $\vec{v}\{a\}$、$\vec{b}$ を生成する。そして $\{a\}$ と b で類似判定を行う。その結果類似性があったとすると、暫定カット $\{a, b\}$ が生成される。なお、この暫定カットの特徴ベクトルは $\vec{v}\{a, b\} = \vec{a} + w \cdot \vec{b}$ と表せる。ここで、重み w は任意の値で $\{w|0 < w < 1\}$ である。(図 3(A))
3. 次にページ b からリンクが張られているページ c と、暫定カット $\{a, b\}$ とで再び特徴ベクトルを生成し直し、類似度判定を行う。$\{a, b\}$ と、c の間には類似性がなかったとすると、この道におけるカットは $\{a, b\}$ であったということになる。またページ c でカットが分割されたわけであるから、ページ d は見にいく必要はなくなる。
4. 上と同様にして $\{a\}$ と e で類似判定を行う (図 3(B))。類似性がなかったとすると、e の先にはどこにもリンクがないため、この道についてはカットが検出されなかったことになる。
5. つぎに $\{a\}$ と f で類似判定を行う (図 3(C))。類似性があったとすると、$\{a, f\}$ が生成される。そしてさらに f のリンク先である e と類似度判定を行う。(図 3(D)) このときも $\{a, f\}$ と e で特徴ベクトルを生成し直す。e は先程、$\{a\}$ 類似判定を行ったときに、類似性がないと判定されているが、$\{a, f\}$ と判定を行った場合に類似性がとあるという可能性がある。したがって、1度類似判定で却下されたページでも、他のページからリンクされている場合は、類似度判定の対象とする。今、判定の結

果類似性があったとされると、$\{a, f, e\}$ が生成される。

6. fからはまだリンクが張られているため、さらに続ける (図 3(E))。$\{a, f\}$ と gとで類似度判定を行った結果、類似性があるとすると $\{a, f, g\}$ の暫定カットが生成される。gからは a に対してリンクが張られているが、a はすでにカットに含まれているため、類似度判定の対象とはならない。

7. さらに、fのリンク先である、h と $\{a, f\}$ で類似度を判定する (図 3(F))。類似度が高いと判定されたとすると、$\{a, f, h\}$ が生成される。

8. hからは gに対してリンクが張られているため、$\{a, f, h\}$ と、g で類似度判定を行う (図 3(G))。類似性があったと判定されると $\{a, f, h, g\}$ が生成される。また、h からは fに対してリンクが張られているが、これも先程の gの場合と同様の理由で、類似度判定の対象とはならない。

9. 以上の手順により、得られたカットは、$\{a, b\}$、$\{a, f, e\}$、$\{a, f, g\}$、$\{a, f, g, h\}$ の 3 つであるが、$\{a, f, h\}$ と $\{a, f, h, g\}$ は重複しているため、より大きなカットである $\{a, f, h, g\}$ を採用する (図 3(H))。

4 あとがき

本研究では、まずネットニュースにおける記事をカット分割し、それに対して検索を行うシステムの設計を検討し、その開発を行った。そして次に、カットという概念を WWW のページ群にも適用し、既存の検索エンジンにおける問題点を解決する手法を提案した。なお今後の課題としては以下の事項が考えられる。

- WWW ページのカット分割の際、現在では出力されたページに対してリンクが張られているという情報は用いていないので、それらも含めた効果的なカット分割方法の考案
- 出力結果をブラウジングできるような GUI の開発

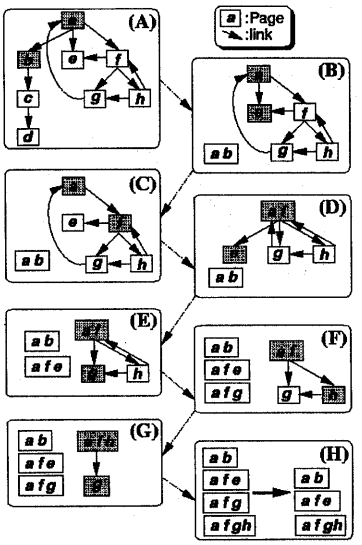

図 3: カット検出手順

謝辞

この研究を行うにあたり、コメントをいただいた河野浩之先生に感謝致します。本研究の一部は，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 (課題番号 08244103) による。ここに記して謝意を表す.

参考文献

[1] 北川 雅嗣、水内 祥晃、田島 敬史、田中 克己, "メイリングリスト情報のためのクラスタリングとカット検出,"第 8 回データ工学ワークショップ (DEWS'97),pp.221-226,(1997)

[2] 茶筌ホームページ、http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html

[3] Peter Ingwersen, 藤原鎮男監訳：『情報検索研究-認知的アプローチ-』、トッパン、1995

[4] Ming-Syan Chen, Jiawei Han, and Philip S.Yu. Data Mining: An Overview from a Database Perspective. *IEEE Trans. on Knowledge and Data Engineering*, Vol. 8, No. 6, pp. 866–883, 1996