

共起グラフを用いたキーワード抽出

大澤 幸生 Nels Eric Benson 谷内田 正彦
email: {osawa,nels,yachida}@yachi-lab.sys.es.osaka-u.ac.jp

大阪大学基礎工学部 システム工学科

あらまし

学術論文のキーワード、特に、既存のどの分野に属するかではなく、著者が新しく主張したい内容を捕えたキーワードの抽出を行うための手法を提案する。本手法のポイントは、論文を意味の構造物と見なし、もしなければ論文が支離滅裂になってしまう、いわば柱に当たる単語を見つけることである。ある単語が、文書中の単語間の共起関係を描いたグラフを繋ぎ止める役割を果たし、文書全体を一つの流れとするような力の強さをその単語の文書全体にとっての重要度とみなし、重要度の高い一定個数の単語を求める。この手法を英語論文に対して適用し、その性能を評価する。

キーワード： キーワード抽出、共起グラフ

Keyword Extraction based on Co-occurrence Graph

Yukio OHSAWA, Nels E. BENSON and Masahiko YACHIDA
email: {osawa,nels,yachida}@yachi-lab.sys.es.osaka-u.ac.jp

Department of Systems Engineering, Faculty of Engineering, Osaka University

Abstract:

The goal of this paper is to catch the newly invented concept in a technical paper. Such content can not be always featured by index terms which appear frequently in the text, because terms on which the author of a document assign strong importance do not always appear many times. Also, the main idea in a technical paper does not appear at a fixed position in the text which is structured as chapters or sections.

In order to extract keywords which might appear less frequently but are more important for the author, we use KeyGraph, a graph representing the co-occurrence among terms in the text. Keywords, or index terms are detected as such that connect the overall KeyGraph of the overall document. We show some experimental results which show the performance of KeyGraph.

Keywords: Keyword Extraction, Co-occurrence Graph

1. はじめに

インターネット、電子図書館、という新しいメディアについて述べるまでもなく、従来からキーワードを用いて文書を探すことの有用さは十分認められてきた。多くの場合、ある研究者が自分と関連の深い分野の文献を探したり、又は特定の論文が頭にあるがキーワードしか思い出せない場合の検索に有効な鍵としてキーワードが用いられてきた。

しかし、この論文で扱うのはこれらとは異なり、指向性の高い検索のための、文献の主張を要約したキーワードである。例を用いて言えば、ある研究者Aがサッカーボールの形をしたロケットを思いついたとしよう。サッカーとロケット、この二つはありふれた単語だが組み合わせると彼のオリジナルな研究となる。かれの論文にはサッカー型ロケットの表面効果や宇宙の現象が細かく述べられるだろう。ある日、別の研究者Bが同じ着想を抱くとする。彼は同様のアイデアが従来からあるかどうかを知りたい。キーワードは「サッカー」と「ロケット」である。この様な、文書の独創性を表したキーワードをテキストから取り出すのが本論文の目的である。

ところで、日常的にわれわれが用いているキーワードには、実際には数種類のタイプがあることに気付く。それらは例えば、次の様なものである。

- 1) 該当する分野名を与えられた項目から著者が選択したもの
- 2) CD-ROMの編集などで、同一の団体が統一的に付けた分野名
- 3) ただキーワードを示せと言われて著者が付けたもの
- 4) テキスト（又は見出しなど）から自動的にキーワードを抜き出したもの等々。

しかし、1や2の様に該当分野を示そうとしても、高い新規性を有する学術論文は既存の分野に収まりにくい。実際、学会などで自分の論文が意外なセッションに振り分けられたという経験はよく耳にする。3の様に著者がキーワードを付す場合も、該当する分野を示せば良いのか、

それとも自分の主張をこめた単語を示すのか迷う場合が多いので、検索者から見れば著者の意図がまちまちなキーワードしか使えないことになる。そこでこの論文では、著者やCD-ROM編集者をこういう迷いから解放するためキーワードはテキストから自動抽出することにする。これにも、

- 1) 見出し語などの文章の構造情報や
- 2) It is important to ...などという明らかな前置き

に頼る方法など様々あるが、本論文では、見出し語に頼る1) は諦める。何故なら、「仮説推論の高速化と制約充足問題」の様に従来の用語で内容を記したタイトルもあれば、「心配ないからメッシュでいこうぜ」(Mesh GAアルゴリズムの提案に関する論文)という様な乱暴な、しかし著者の言い分を明確に伝えるタイトルも少なくないからである。又、新聞記事では記事の冒頭部に重要な単語が多いとの報告がある[1]が、文書の種類によって主要部分の位置は様々で[2]、学術論文に限っても著者によって章建ての仕方に流儀があったりする。又、[3]でも、重要な意味分類が少数の大段落に集中するとされているが、これは被験者を文書の読者としたことにも因るのではないだろうか。実際、本稿でも、正解のキーワードを読者が決めると読者の興味の対象に応じて重要な段落が集中してしまうことが分かったので、著者自身のコメントを採用したところ、「正解」は大きく異なるものとなった。又、本稿では、統語解析可能な文型で記述されているとは限らないので2)も用いない。

さらに、

- 3) テキスト中からある単語 w の、平均的な文献と比較した対象文献 D での頻度 $f(D, w)$ の多いものを選ぶ方法

は理論的な基盤性が高く[4-6]、簡単な計算で処理ができる。しかし、これらは

- 仮定A) 平均的な文献での頻度 $F(w)$ が意味を持つ
- 仮定B) 頻出語は重要である

という二つの仮定をおいていることになる。これでいいのだろうか？

仮定Aは、頻度の測定から言えば、得られた頻度データが統計的に有意であることを必要としている。しかし、筆者が調査したところ、「仮説」という言葉の人工知能分野での出現頻度はロボット工学における出現頻度の3倍を上回るが、人工知能での平均頻度を上回るロボット工学の文献は1割を超える（assumptionはロボットでは環境についての仮定として用いられ、人工知能では仮説推論における仮説としても用いられることが多い）。これは、assumptionという単語の用いられ方にばらつきが多いことを意味しており、特にこの場合にはばらつきは多義性にも原因を發している。多義性の解消のためには著者の独創的な視点を考慮しなければならないが、ソースや新聞記事での共起性から分かる語の関連性を用いて関連語群を取り出す手法[3,7]では平均的な視点に文書の解釈が大きく左右されてしまい、本研究の目的とはそれてしまう。

又、仮定Bは、さらに信頼できない仮説であると筆者は考えている。何故なら、学生に重要なことを伝える時は筆者はよく、「一度しか言わんからよう聞いとけ」というからである。学術論文の場合にも、一般に次の様な傾向があるようである。

傾向 A) 論文の主張の内容よりも、読者が理解するのに必要な単語を何度も繰り返す。

例えば、ロボットビジョンの研究で、「ロボット」という語は非常に多く出現する。ロボット工学の分野で最も当り前のこの単語が、読者をこの分野に誘い込み、理解の土台を確保する役割を果たしている。「画像」という単語もやはり頻度が高い。低頻度語を除去しても、特許文書の分類精度が殆ど低下しないのもこのためである[8]。しかし、ロボットビジョンに関するある論文の著者は、これらより頻度の遥かに少ない「垂直エッジ」などが主張したい単語だと言う。又、プランベジアンネットワークにつ

いての論文の著者は、「ベジアンネットワーク」や「プラン」よりも「即応性」などの稀出語が重要とした。ここでの「即応性」や、「垂直エッジ」はその論文で従来にならぬ用いられ方をしているので統計的に有意な出現頻度を測定しにくいし、測定しても特にこれらの論文で頻度が高い単語ではないことが分かる。

更に、大阪大学基礎工学部システム工学科谷内田研究室の大学院生とスタッフ5人の論文について同様のインタビューを行ったところ、この傾向は全ての論文で成り立つことが分かった。彼等はロボット、画像、人工生命の研究者であり、筆者のキーワード抽出のメカニズムについては全く知らせていない。

そこで、本論文では、新しい手法でキーワードを抽出することを提案する。すなわち、本論文では、構造情報を用いず、統語解析を行わず、頻度だけで重要度を比較しない様なキーワードの自動抽出を考える。

2. 共起グラフKeyGraphを用いたキーワード抽出
われわれの基本的なコンセプトは、

文書を建物と見なそう。この建物は、壁(必要な説明部分)もあるし、ドアもあるし窓もある(詳細な記述)。他にも様々な装飾がある(比喩や例など、付加的な記述)。しかし、ちゃんと建つにはそれらを繋ぐ柱が一番大事である。

ということである。この「柱」がキーワードだというのがわれわれのアイデアの原点である。

学術論文は、冒頭に用件が密集している新聞記事[1]とは異なり、文章が論理的な鎖状に構成されている。その中には、数式やその説明、例証などのまとまりもあるが、これらはある単語集合が繰り返される(この論文の1節における「ロボット」や、数式での変数名などがその例である)。しかし、これらの頻出語は細部の記述であって要点とは別のいわば「壁」に相当する。これらのまとまりは文中にいくつもあ

それだけでは文章は、枝葉だけ取り出した、ばらばらで支離滅裂な内容になってしまう。これらのばらばらのまとまりの群れを一つに束ねるものこそ文章の要点であり、そうして束ねられて一つの鎖となったのが一つの文書だと考える。

2.1 KeyGraphによる

キーワード抽出の基本アルゴリズム

具体的には次のアルゴリズムを用いて、文書Dのキーワード集合KEYWORDSを求める。

[キーワード抽出の基本アルゴリズム]

- Step 1 テキスト中に2回以上現われ、辞書USUALに載っていない単語の集合をWORDSとする。
- Step 2 共起度 $c(X, Y)$ (X, Y が同じ文中に出現する頻度) の上位 |WORDS|-1 までの単語ペア 2 単語 $\{X, Y\} \in 2$ WORDS を選択し、エッジ $X-Y$ で結んだグラフ G_e を作る。
- Step 3 $Z \in D$ なる単語 Z のうち、複数の連結グラフを連結する強さが上位 10 位までの単語 Z を集合 KEYWORDS1 とする。ここで、連結グラフ G_a と G_b を連結する効果の強さ $link(G_a, G_b, Z)$ の定義は次の通りである。ただし、 $X \cup Y = X + Y - XY$ とし、複数の連結グラフに対して $link(G_a, G_b, G_c, \dots, Z)$ も同様に定義する。

$$link(G_a, G_b, Z) =$$

$$\frac{\sum_{j=1}^s |W_j| \cdot |Z_j|}{\sum_{j=1}^s |W_j| \cdot |S_j|} \left(\begin{array}{l} W_j \equiv \forall \text{words} \subset (j\text{-th sentence}) \cap (G_a - Z_j) \\ Z_j \equiv Z \text{ s } \subset j\text{-th sentence.} \\ S_j \equiv \forall \text{words} \subset j\text{-th sentence.} \end{array} \right)$$

∪

$$\frac{\sum_{j=1}^s |W_j| \cdot |Z_j|}{\sum_{j=1}^s |W_j| \cdot |S_j|} \left(\begin{array}{l} W_j \equiv \forall \text{words} \subset (j\text{-th sentence}) \cap (G_b - Z_j) \\ Z_j \equiv Z \text{ s } \subset j\text{-th sentence.} \\ S_j \equiv \forall \text{words} \subset j\text{-th sentence.} \end{array} \right)$$

(1)

$link(G_a, G_b, Z)$ の $G_a - Z - G_b$ を結んで G_e と併せて出来るグラフを、KeyGraph と呼ぶ。

- Step 4 Step 3 で結ばれたリンクの数の上位の 10

単語を KEYWORDS2 とし、

KEYWORDS = KEYWORDS1 ∪ KEYWORDS2 とする。

以下、このアルゴリズムの意味を説明する。

・Step1の意味 Step 1 で、対象とする論文の内容説明に必要となりそうな単語にスポットを当てる。基本的には、これは頻度で選ぶ、というのは、1節における「理解の土台を確保する役割」を果たす前提知識の単語や、例証や変数名のまとまりを拾い集めるのが Step 1 の目的だからである。これを上位数単語などとせず2回以上出現する単語としたのは、網羅的にこれらを拾うためである。尚、辞書USUALは文法上の単語ではないit, is, where などの構成要素(即ち、functional words) からの辞書である。

・Step2の意味 次に、Step2では、1で拾った単語数よりも1だけ少ないリンク数でWORDSを結んだグラフ G_e を得ている。

何故このリンク数を用いるかと言うと、 G_e が単結合な連結グラフとなる場合のリンク数に等しい(オイラーの定理)からである(連結グラフ: G 中のノードの全てのペアを結ぶリンクのパスが G に含まれるようなグラフ G)。図1-aの様に、このような単結合グラフは、文章が単語から単語へ論理的に導くプロセスが冗長性を含まないことに当たる(注: 上記のアルゴリズムでは導出の順序、即ち図1の矢印までは得ない)。

逆に言えば、このリンク数で単結合グラフでない図1-bの G_e では、 G_e の全体が連結グラフとならず、しかも G_e の連結な部分グラフ中に冗長な導出関係(単語3が単語2と単語1の両方から導かれる)が含まれることになる。

文書が意味的には鎖状に連なると考えると、単語3から単語4への導出が実際には存在する可能性がある。しかし、ローカルな共起性から分かるのはこの様なグローバルな導出関係ではなかろう。むしろ、ローカルな情報で「理解の土台を確保する」レベルの記述としては、単語3は G_e の中で単語4を導く役割よりも G_a の中で導かれる立場(単語1から単語3への矢印)の方が顕著

に現われる。言い替えると、Step2のリンク数は、文書中のローカルな情報から取り出し得るローカルなつながりを取り出し、取り出せても信頼しにくいグローバルなつながりを一旦断ち切るのに丁度良い値なのである。

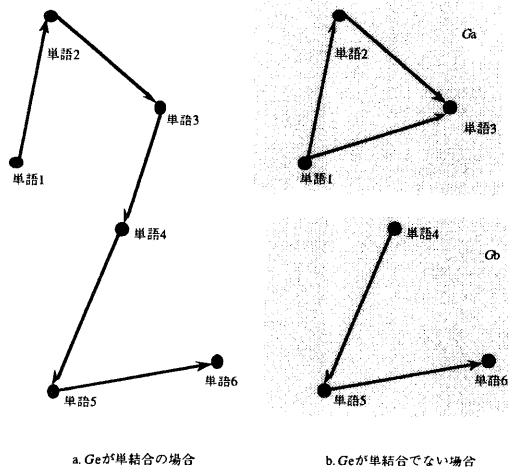


図1 Step 2のリンク数を用いたGe

・Step3の意味 Step2では、意味的なまとまりを単語の共起グラフとして拾い集めた。今、Step2でGeがGaとGbの二つのグラフに分かれて取り出せたとして。このままでは、文書Dは両方の意味的なまとまりに分割されて支離滅裂となってしまう。Dが一貫性を保っているのは、Dの著者が一貫した主張を主張する目的でGa、Gbに当たる内容を著述したからである。即ち、Z(図2の単語7)がGaとGbを結び付ける効果の強さを測定することができれば、その効果の強い単語Zに著者の主張が込められていると見なすことができる。

ここでは、この結び付ける効果の評価関数として、GaとGbの内容の両方について著者が考えている時に、単語Zについてどの程度述べているかの確率を用いる。論理的にはこれは

$$Z \leftarrow (Ga \text{ AND } Gb),$$

即ち

$$Z \leftarrow Ga. \text{ または } Z \leftarrow Gb.$$

なるZの確率であるから、 $p(Z \leftarrow Gb) \cup p(Z \leftarrow Ga)$ をとったのが式(1)である。

尚、式(1)では $Z \leftarrow Ga$ と $Z \leftarrow Gb$ が独立という仮定が用いられている。これは、先述のようにGa、Gbが読者がDの内容を理解するための前提であって、互いに導かれるものではないという仮説から来た仮定である。以上の基本アルゴリズムをKeyGraph 1と呼んでおく。

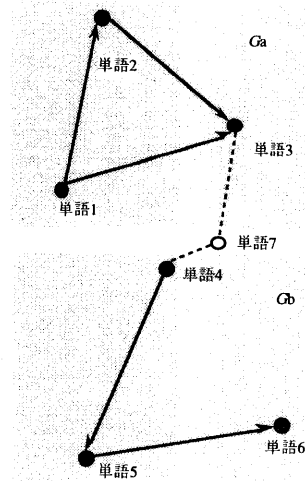


図2. Step3によるキーワード(単語7)抽出

・Step 4 の意味 KEYWORDS1に対し、文章で読者の理解の土台としての役割の強さをStep3で付与されるリンク数で評価する。ただし、後で検索する人にとってKEYWORDS1とKEYWORDS2はどちらが土台であるかは不明なので、両者をマージしたものをキーワード集合とするのである。

2.2 KeyGraph 1の解の評価

KeyGraph 1で以下の文章を解析した結果を示す。
[例文] (全文を示すために短い例とした)

This paper describes a novel method to interleave planning with execution in a dynamic environment. Although, in such planning, it is very important to control deliberation: to determine the timing for interleaving them, few research has been done. To cope with this problem, we propose a method to determine the interleave timing with the success probability, SP, that a plan will be successfully executed in an environment. We also developed a

method to compute it efficiently with Bayesian networks and implemented SIP system. The system stops planning when the locally optimal plan's SP falls below an execution threshold, and executes the plan. Since SP depends on dynamics of an environment, a system does reactive behavior in a very dynamic environment, and becomes deliberative in a static one. We made experiments in Tileworld by changing dynamics and observation costs. As a result, we found the optimal threshold between reactivity and deliberation in some problem classes. Furthermore we found out the optimal threshold is robust against the change of dynamics and observation cost, and one of classes in which SIP works well is that the dynamics itself changes.

[出力結果] 図3で、黒い●がStep1で得られた単語、実線がStep2で得られたGeのエッジ、点線はStep3のZとそれまでのGeの連結グラフの連結エッジである。二重丸の単語の集合が、結果として得られる集合KEYWORDSである。"reactivity"は一度しか出ていない希少語であるが、著者はreactivityとdeliberationを主要語とした。

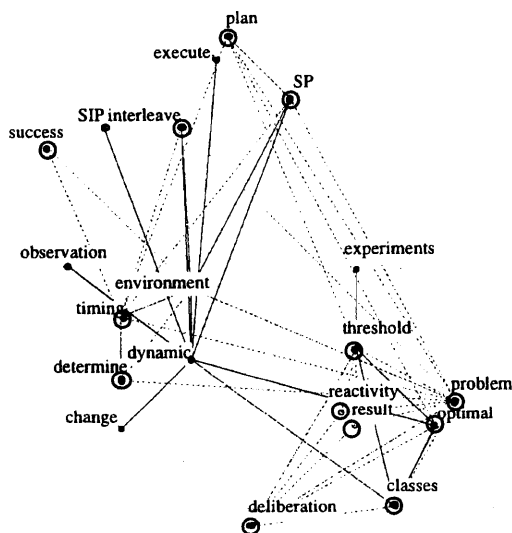


図3 KeyGraph 1の出力例

結果のデータを集めてKeyGraph 1の評価をする。現時点では十数文献の試行しか行ってい

ないが、それらの上位の延べ180単語を各文献の著者に見せた。すると、そのうち1-2割程度の単語は著者の新しい主張と無関係だが、他の単語は新しい主張の要点に属すると返答した。これに対し、相対頻度 ([5,6]参照) の上位の単語20個を各文献についてとる手法の結果に対しては、7割以上が分野の紹介あるいは枝葉の説明でしかないとのコメントであった。

以上から、KeyGraphが論文の主張点を捕えていることが分かる。尚、重要なのに拾えていないと指摘された単語が7語あった。この対策は3節で述べる。

2.3 KeyGraph 1の計算時間

上のアルゴリズムの計算時間を評価してみる。Step 1 は全ての単語の頻度を数えるので、文中の単語数Wに対して線形時間である。Step 2 は、WORDSの間の各単語ペアの共起頻度を調べるからその計算時間は $|\text{WORDS}|^2$ のオーダーとなる。このWORDSはWにはほぼ比例するので、Step2には W^2 の計算時間がかかる。最後に、Step 3, 4 では、D中の各単語が各連結グラフの単語と同じ文中にあるかどうか数えるのに時間を費やすので、最悪 W^3 かかる。高速化が今後の課題である。

3 キーワード抽出のアルゴリズムの発展型

3.1 共起グラフの分割手法の改良

KeyGraph 1 のStep2では、テキストの意味的まとまりを共起グラフとして取り出している。しかし、ここでは見かけの共起性から共起グラフを生成しているため、Step3で初めて(図3の点線の様に)取り出されるのが望ましいリンクがStep2で取り出されてしまうことがある。その例を図4に示す。これは、NBP法という仮説推論の高速化手法に関する筆者らの論文[9](4996words)をKeyGraph 1 で解析した結果である。この研究は、命題論理の仮説推論を高速化していた推論手法であるNBP法を、述語版に拡張したものであり、predicate logic (述語論理)という熟語はこの文書において非常に重要であるが、KeyGraph 1ではこの熟語が抽出されていない。

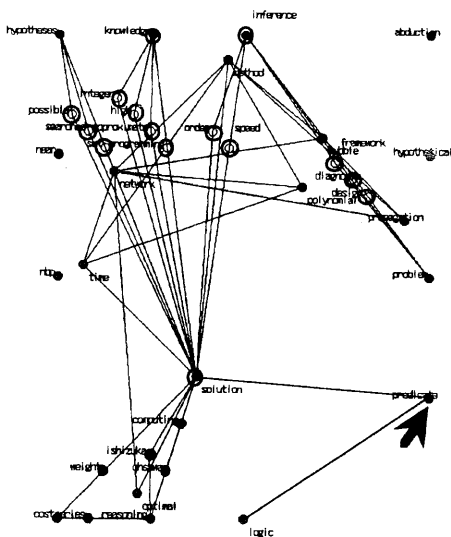


図4 KeyGraph 1 の失敗例 (注：縮小前のグラフが大きかったので点線が実線と区別できないが、白い◎の乗った線が点線である)

その理由は、predicate (図4の太い矢印の先) と solution の間のリンクが太いために Step2 で Ge のエッジとして取り出され、predicate-logic というグラフが Ge から分離できなかったからである。こうなると、分離されたグラフを結ぶ Step3 の手続きでは predicate をキーワードに含められなくなる。そこでわれわれは、グラフ Ge で複数の連結グラフが細いリンク (具体的にはエッジ一本) で連結されている場合には、その細いリンクを取り去るという方法を取り入れた。即ち、先の Step2 の直後に次の Step2.1 を挿入する。

Step 2.1

begin pruning

```
for all e ∈ E(edges in Ge),
  if no_path(end1(e), end2(e)), then
    prune Ge=Ge-e.
  endif
```

endfor

endpruning

ただし、no_path(X, Y) は X からエッジを辿ってノード Y にたどり着くことができないことを表わす述語で、その真値を計算する時間は最大で Ge 中のエッジ数に比例する。従って、Step2.1 の計算時間は E^2 と、2.2 節に述べた Key Graph の計算時間に比べると十分小さい。

実際、この Step2.1 を挿入することによって上記の例だけでなく、Key Graph1 で取り出せない単語がキーワードとして正しく取り出せるケースが多かった。先と同様に文献の著者に解の評価をしてもらったところ、拾えていないと指摘された7語のうち、6語までは Step2.1 によって抽出できるようになった。

しかし、例えば図4では time の代わりに speed が抽出されているが、time - solution, network - reasoning のリンクは Step2.1 を改良することによって切断すべきなのかも知れない。現在、Step2.1 の改良を練っているところである。

3.2 熟語、語尾の処理

現在、単語が複数個、一定の順序で繰り返し出現するパターンを熟語としている。これらは単語と同様に扱い、その要素単語が単独で現れるものを区別して計測している。図4の "predcate logic" は、この方法と Step2.1 の併用により取り出すことができる。

語型を考慮したもう一つの点は、語尾変化である。例えば、planning と plan などの様に、同一の動詞の名詞型と原形は近い意味を持つと考えられるので、同一のノードにまとめてグラフを作成した (ただし、図3の例はこの改良以前のものであり、reactivity と reactive を区別している。即ち、一回しか出現しない reactivity を得た例である)。具体的には [] の語尾変化の手法に準じた。これでは動詞の不規則変化が捕え切れないが、多くの不規則変化する動詞は USUAL で削除済みであることを考えると大きな誤差は生じない対応と考えている。

4 今後の改良予定

Key Graph が狙うキーワードは、対象文書の関連分野ではなく、著者が新しく主張しようとする

る内容を表現する単語である。この目的には、頻度が少なくても、著者にとって重要な語を取り出せる本手法は有効であると考えている。

キーワード候補語との共起度を考慮して単語の重要度を考えている点では、KeyGraphの考えは[10]の手法に近い。しかし、Step2でローカルな単語同士のリンクを数えるのではなく、Step3で単語のグローバルな重要さに当たるリンク強さに重きをおいて計算している点では方針から異なっている。

今回、文書の構造情報を用いてない理由は1節に述べた。しかし、論文の章立てが読者に与えるわかりやすさを考えると、やはり全く用いないのは誤った戦略であろう。

例えば、図3や図4では中央の連結グラフだけが極端に大きく、他の連結グラフは1ノードだけからなっている。各連結グラフを意味のまとまりと見るには、これは不自然に見える。現在進めている改良として、対象論文の各節でKeyGraphを別々に作成することによってグラフのアンバランスを解消しようとしている。ただし、この改良では難問がある。本来、基本アルゴリズムのStep2の連結グラフは、著者の章立てからは見えない意味的まとまりを取り出したものである。実際、論文の執筆において著者は、グラフで書くと3次元になる内容を、文章が一次元情報であるという制約に押し込めるために戦い続けるものであり、その一つの解決が上手な章立てによってなされる。しかし、KeyGraphでは多次元のままグラフ化するのであるから、章立ての役割のうちこの部分は不要となる。即ち、両者は共通の目的を持つ別の手法と言えるから、KeyGraphに加えて文章の構造をどの程度まで細かく反映させるかが難しい。

5 おわりに

今後の改良を行う上で何より必要なのは、著者から直接提供頂ける文献とコメントである。読者の方々からメールなどにて（アドレスは本稿冒頭）論文をご提供頂ければ幸いである。

研究へのコメントを頂きました東京大学工学

部の石塚満教授、コメントと文献の提供を頂きました東京工業大学の山田誠二助教授、大阪大学助教授の八木康史助教授、日本学術情報センターの杉本雅則先生に記して感謝します。

◇ 参考文献 ◇

- [1] 野本, 松本: テキスト構造を利用した主題の推定について, 情処研究会NL114-8 (1996)
- [2] 木本: 日本語新聞記事からのキーワード自動抽出と重要度評価, 信学論, J74-D-1, No.8, pp.556-566 (1991)
- [3] 鈴木, 増山, 内藤: 語の意味分類の出現傾向を考慮したキーワード抽出の試み, 情処研究会, NL98-10 (1993)
- [4] H.P. Luhn, A Statistical Approach to the Mechanized Encoding and Searching of Literary Information, IBM J. Research and Development, 1:4, 309-317 (1957)
- [5] G.Salton, C.S. Yang, On the Specification of Term Values in Automatic Indexing, J.Documentation, 29:4, 351-372 (1973)
- [6] C.T. Yu, et.al., Term Weighting in Information Retrieval Using the Term Precision Model, J.ACM, 19:1, 152-170 (1982)
- [7] 白倉, 梅村, 小川: 新聞記事における事件特定のための単語群の抽出, 情処研究会, NL113-17 (1996)
- [8] 西野: 日本語テキスト分類における特徴素抽出, 情処研究会, NL112-14 (1996)
- [9] Yukio Ohsawa and Mitsuru Ishizuka: A Polynomial-time Predicate-logic Hypothetical Reasoning by Networked Bubble Propagation Method, Advances in Artificial Intelligence, G. McCalla, eds., Springer Verlag, LNAI-1081, 375-387 (1996)
- [10] M.F. Porter, An Algorithm for Suffix stripping, Automated Library and Information Systems, 14:3, 130-137 (1980)
- [10] 原, 中島, 木谷: 単語共起と語の部分一致を利用したキーワード抽出法の検討, 情処研究会, NL106-1, 1-6 (1995)