NAIST-IS-MT9651004

修士論文

設計過程に関する情報の
構造化と可視化の研究

足立 秀和

1998 年 2 月 13 日

奈良先端科学技術大学院大学
情報科学研究科 情報処理学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
修士 (工学) 授与の要件として提出した修士論文である。

足立 秀和

指導教官： 西田 豊明
植村 俊亮 教授
武田 英明 助教授

設計過程に関する情報の
構造化と可視化の研究*

足立 秀和

内容梗概

設計を行うにあたり，過去の設計記録を効率的に利用するための技術に関心が持たれている．設計情報は，問題解決過程，設計対象の専門的知識など様々な情報で構成されるが，それらを将来再利用できる有益な情報として，記録し表現するための手法が模索されている．本研究では構造化されていない文書一般に対する構造化の方法について検討し，さらに実際の現場における設計情報の再利用を支援するシステムを試作した．

キーワード

設計情報, 設計過程, 弱構造化, 問題解決過程

*奈良先端科学技術大学院大学 情報科学研究科 情報処理学専攻 修士論文, NAIST-IS-MT9651004, 1998年2月13日.

Studies about Structuring and Visualizing of Design Process*

Hidekazu Adachi

Abstract

In this paper, we present the method of reusing poorly-structured documents(e.g. minutes of meetings) as design information. It is difficult for designers to retrieve design knowledge from minutes. So we propose a new approach that supports designers to retrieve design knowledge. In our approach, the first step is a construction of semi-organized knowledge from minutes, and the second step is an extraction of well-defined knowledge from semi-organized knowledge. In this way, designers can easily retrieve design knowledge from well-defined knowledge that is derived from minutes. With this approach, we implement a prototype of a system for extracting and visualizing resolving-process of design .

Keywords:

design information, design process, semistructured document, problem-solving process

*Master's Thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT9651004, February 13, 1998.

目 次

1 序論 **1**

1.1 はじめに 1

1.2 研究背景 2

1.3 研究内容 2

2 設計過程情報の取得 **4**

2.1 設計の概要 4

2.2 設計の明示化法 6

2.3 設計実験の手法 7

2.4 設計情報の具体例 8

2.5 設計情報から設計知識へ 8

3 文書の構造化 **11**

3.1 構造化とは 11

3.2 構造化の手法 12

3.3 文章を特徴付けるキーワードの抽出法 13

3.3.1 tf-idf 法 13

3.3.2 χ^2法 14

3.4 文書の文脈を把握する手法 15

3.4.1 ベクトル空間法 15

3.5 新しい空間ベクトル法の提案 18

3.5.1 従来の空間ベクトル法 18

3.5.2 単語出現頻度データの標準化 20

3.5.3 プロトコル解析文の文脈の解析例 21

3.5.4 解析実験の考察 23

3.5.5 文脈解析実験のまとめ 27

4 文書構造化の表現としての弱情報構造 **28**

4.1 弱情報構造 28

4.2 文書の弱構造化 29

5 設計情報可視化システムの構築 **33**

5.1 議事録の解析 33

5.2 設計情報構造化の過程 35

5.3 CM-2 上に弱構造化された設計知識 38

5.4 設計情報の弱構造化実験の評価 40

5.5 設計情報からの問題解決抽出 43

5.6 問題解決仮定抽出実験結果と考察 47

6 結論 **50**

謝辞 **51**

参考文献 **52**

図 目 次

2.1 設計の概要 5
2.2 設計局面の流れ 5
2.3 プロトコルデータ例 9
3.1 文脈ベクトル 17
3.2 自転車荷台の設計に関するプロトコルの一部 22
3.3 文章群の切り分け 24
3.4 データ標準化した後の文章類似度 25
3.5 単語出現頻度を用いた文章類似度 26
4.1 文書例 30
4.2 文書木構造 31
5.1 設計会議議事録例 34
5.2 ブロックの構成と文章内容の言明 37
5.3 言明と文章表現の関係 38
5.4 ”差動増幅回路”から検索した結果 39
5.5 ”差動増幅回路”から検索した議事録文ブロックの内容 40
5.6 ”差動増幅回路”から検索した議事録図ブロックの内容 41
5.7 議事録から文脈をたどる例 (1) 42
5.8 議事録から文脈をたどる例 (2) 42
5.9 議事録から文脈をたどる例 (3) 43
5.10 弱構造化された文章群 44
5.11 設計情報可視化システム概略 45
5.12 CM-2 データ 46
5.13 問題解決過程情報の表示例 47

表 目 次

4.1 文書要素 29

第 1 章

序論

1.1 はじめに

環境問題や自然破壊といった地球規模での視点による問題が近年盛んに提起されている．環境中に存在するダイオキシン，PCB，有機スズなどによる微量汚染問題，酸性雨や酸性霧，大量に山積みされる古タイヤや東南アジアで処理される鉛蓄電池．設計段階では考慮に入ることのなかった事象が実際には発生してしまったことがこれらの問題の原因であることが多い．

現在では人工物の流れには動脈系 (製造から販売) と静脈系 (廃棄から処理，処分) があると理解し，発生から消滅までを考慮に入れた人工物の設計が求められている．すなわち従来の設計では顧みられることのなかった事象や領域についても評価を下し，影響を正確に判断することが設計者には必要である．

たとえば自動車のタイヤの設計について考える．自動車の機能から導かれる条件は明確で，雨中における走行性の向上，走行中の振動の低減，耐摩耗性などが要求される．自動車の性能を向上させるのが目的の中心であり，高性能なタイヤが求められる．しかし廃タイヤが処理処分されその結果が与える環境への影響を考慮に入れた場合には，タイヤ処分コストの低減，擦り切れるタイヤの量や成分，リサイクル性の有無などの条件が加わり，自動車性能と環境影響とのトレードオフや，製造コストとタイヤ成分との兼ね合いなど，製造段階の根底から設計の変更が要求されてくるであろう．

このように設計対象はますます複雑になり，それにともない設計作業は専門度を深め，高度かつ難解なものになってしまった．その結果，設計者に要求される設計対象に対する専門知識はより深くかつより広範囲に及び，一人の人間がすべ

ての設計作業を行うことが困難となっている．

1.2 研究背景

複雑な設計対象に必要な専門知識を集団で分担し，一人の設計者の負担を軽減するのが現在では普通である．専門知識を集団で共有，利用するために，今まで設計者の頭にあった知識を情報の形で表さなければならない．かつ，その情報は集団での使用に便利な形態である必要がある．

また，設計作業自体の記録も貴重な情報である．同じような設計対象の設計を常に同じ人間が行うことは稀れで，後任の設計者が担当すると考えるのが普通である．そして協調設計においてその構成人物が常に同じであることは一般的な前提ではない．過去に設計作業を行った記録を利用すると似たような設計対象に対する設計作業の今後の経過に予測が付き，設計作業を円滑に進めることが可能である．

しかし協調設計や設計作業に関する情報の再利用を進めた結果，設計対象に関する専門知識の情報の増大だけでなく，設計作業に係る記録すべき情報の更なる増大を招き，設計者による設計情報の利用に負担を生じる結果となっている．同時に膨大な設計情報の作成は，設計者に大きな負担を課している．また過去の大量の設計情報が分類整理されないままで，再利用が促進されず貴重な情報が埋もれてしまうといったことも考えられる．

以上のことから設計情報の作成支援，設計情報の共有促進による協調設計の支援，設計情報の再利用促進による設計の効率化などを目的として，計算機による設計支援の研究が続けられてきた．

1.3 研究内容

本研究では設計情報の作成支援に関係して，設計情報として得られた構造化されていない文書を構造化する方法に関して検討した．具体的にプロトコル解析による設計実験で得られる文書を統計的な手法で構造化する方法について新しい手法を提案した．また設計情報の共有や再利用の促進に関係して，実際の設計現場

から得られた設計情報である設計会議議事録について，設計情報を可視化するシステムを試作した．

第 2 章

設計過程情報の取得

2.1 設計の概要

目的を設定し，それに対する実現の手段を考えた時点で設計が始まる．実現するには「何を達成したいか」という設計目標を明確にしなければならない．設計目標は必要な機能として書き下され，必要条件として書かれた必要機能は機能空間を形成する．

一方，設計対象は実世界のモノであるから，実世界における制約の支配下に存在する．それらの制約は属性として書き下され，属性空間を形成する．すなわち，形式的には設計は機能空間で指定された領域に対応する属性空間の領域を求めることである (図 2.1)．

設計過程はさらに以下のように具体的に解析される．

Suh によれば，設計作業は「問題設定」「創造過程」「分析過程」「最終チェック」という局面に区別され [7]，その局面が繰り返し反復されて設計プロセスが形成される．個々の局面は以下のようなプロセスである (図 2.1)．

問題設定　解決すべき事柄を明確な問題として設定する

創造過程　設定された問題を解決する具体的な手段を創造する．

分析過程　創造した手段が正しくかつ合理的であるか否かを分析する

最終チェック　分析なった手段が所期の問題への忠実度により評価する

また武田は設計実験により得られたプロトコルデータの解析を基に次の五段階の副過程からなる設計過程モデル [20] を提案した．これによりそれぞれの副過

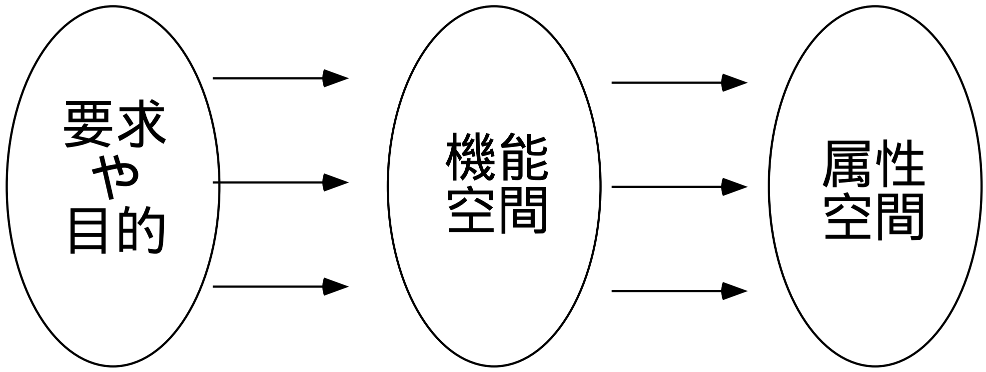


図 2.1　設計の概要

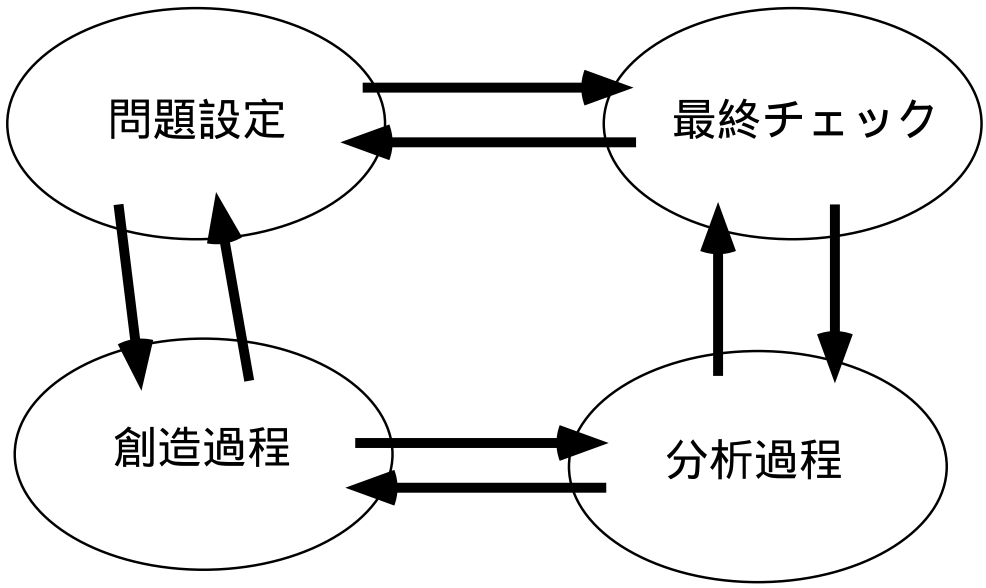


図 2.2　設計局面の流れ

程における推論の詳細が解明され，設計における問題解決過程のモデル化がなされた．

問題提起　解決すべき問題の提起

提案　問題提起を受けて，それに対する解決案を提案する

展開　提案した解決案を具体化する．

評価　解決案をある規準に従い評価する

決定　評価の結果を基に提案された解決案の採用を決定する

以上のことから設計にはその過程に明確な段階があり，かつ段階の遷移があることがわかる．また，設計過程は目的を達成するため手段の創造とそれに付随して発生する問題の解決の過程と言い換えることができる．したがって，設計を論理的な枠組で理解できるとともに，設計に関する情報はその枠組に沿った手法で解析することができると考えられる．

2.2 設計の明示化法

実際の設計ではどのように設計が行われているのだろうか．設計行為は個々の設計者が行うものであるが，第三者が設計行為を客観的に分析，解析するためには設計を明示化し，設計情報として取得する必要がある．

設計情報は設計過程情報，設計対象情報，設計行為情報などに分類される様々な情報から構成されるが，それらをすべて表現し，設計自体を正確に明示化する情報の記録，記述法は確立されていない．ここで，設計明示化には大きく分けて以下の手法がある [20]．

1. 個人的な経験や理解の体系化すr
   設計者が自らの設計経験や設計の理解を設計方法論として提示する．

2. 設計の実例の記録
   設計の実例を記録したもの

3. 設計の調査
   企業などで実際に行われている設計をなんらかの方法で調査するもの

4. 設計実験
   実験的方法により調べる．

今，設計過程や設計作業における思考の過程の変遷を記録するには実験的方法が適していると考えられている [20]．

2.3 設計実験の手法

具体的に設計実験の手法について説明する．設計実験は心理学的実験手法を利用して，設計作業における思考の過程を調べるものである．設計実験は極めて詳細なデータが得られ，設計者の行為，思考を直接解析することができる．その半面，個々の設計実験データを設計一般に適用することには注意を要する．

設計実験の方法として主に以下の 3 つのものが考えられる．

- プロトコル解析 (Protocol analysis)
  被験者が考えていることを常に声に出しながら作業を行う (thinking aloud) ことにより思考過程を記録する．

- Retrospective reporting
  作業終了した後，実験者が被験者に質疑を行い，質疑応答を記録する方法．

- Informal Reporting
  実験者は被験者が作業を行っているところを観察し，必要に応じて質問をする方法．

プロトコル解析は以下の手順に従って行われる．

1. 目的の決定
   設計プロセスの局面のどの部分を分析するのか目的を決める．

2. 実験環境の設定

3. 実験の実行とデータ収集
   被験者に課題を声を出しながら解いてもらい，被験者の発話等を VTR などで記録する
4. データの文字化
   音声記録として取得したデータを文字化し，画像記録の図面化などを行う．
5. データ処理
   プロトコルデータを分節し，記号化する．

プロトコル解析は発話のすべてを記録し，かつ VTR などを使って描かれた図面や身振りなども記録することから設計に関するデータの収集量が多い．しかし口に出しながら設計を行うため被験者の自由な設計作業を損なうことが考えられる．

2.4 設計情報の具体例

以下にプロトコル解析で取得された発話データの例を図 2.3示す．

このプロトコルデータは，3 人の設計者によるリュック (backpack) をマウンテンバイクに取り付けることが可能な荷台または金具の設計に関するものである．

2.3で述べたようにプロトコル解析ではすべての発話を記録するため，設計情報としてのプロトコルデータは情報量が多いことがわかる．また，発話データは時系列的に並んでいることから，設計の時間的経過を表現することに適していると思われる．しかしデータに冗長な部分が多く無意味な部分もすべて記録されていることから，プロトコルデータから設計過程を解明するにはデータの解析において意味のある情報を抽出するデータの解析に工夫が必要である．

2.5 設計情報から設計知識へ

設計過程情報を取得し，それを基に計算機によって設計支援を行うためには設計情報を計算機が扱うことの容易な情報形態に加工する必要がある．また Integrated Intelligent Interactive CAD[11] で議論されるように，計算機により知的に設計作業を支援するには設計情報を設計知識に高める必要がある．知識表現の具体化の

```
100 K do we have the preliminary design  for
      the em carrying slash fastening device?
101 X Yep the uh Batavus design
102 I yeah
103 K mm mm (inaudible)
104 J (reading) pack is firmly attached to the
      bike positioning of  the backpack was alright
      fact that the centre of gravity of the  backpack is
      placed rather far to the back of the bike (inaudible)
105 I do we have any em
106 J there's a problem with potholes .. the backpack tends
      to slide  up and down which adversely influences stability
      I guess when you hit  bumps
107 I isn't that in the negative
108 J mm yeah well the product was considered ugly well
      that's  solvable (laughter) we can fix that one if nothing
      else ... it takes a  while  to get used to cycling with
      this weight mistakes are made attaching the  fastening device
      to the bike so it has to be easy to attach
109 K with only one yeah gotta be fool proof so that's part of our
110 J yeah that should be in our spec
```

図 2.3 プロトコルデータ例

一つとして「知識メディア」[14][15] があるが，設計情報を集団が共有して利用し将来再利用することを可能にするには，設計情報を「知識メディア」を使い設計知識として表現することができる．

こういった観点に基づき，第 3 章ではプロトコル解析で得られた設計情報文書を構造化し，より有用な設計情報に近付ける方法を考察する．

第 4 章では議事録として得られた設計情報を構造化する方法を考察する．

第 5 章では議事録を知識メディア化し，設計知識として利用する設計環境の試作を行う．

第 3 章

文書の構造化

設計実験のプロトコル解析で得られる発話データは自然言語文で表現される．3 章では文書として得られたデータを構造化して文書に情報を加えることについて説明し，人手を使わず自動的に構造化を行うための手法に提案を行う．

3.1 構造化とは

自然言語文以外の情報を付加されていない文書は構造化されていない文書である．一方，自然言語文の意味を簡明，明確にするなどの目的で，自然言語文以外に様々な 2 次情報を付加することを構造化という．構造化された文書には例えば，HTML や SGML で記述された文書がある．

計算機とネットワークの発達によって電子文書が大量に生み出されているが，もし自然言語文のみの文書しかネットワーク上に存在しなければ大変不便なことであろう．HTML などの形式で他の文書の関係や文書内での文章間の関係が記述してあることにより他の利用者が有効に利用することが可能となっている．

文書の構造化にはたとえば次のような 2 次情報を付加することができる．

- 重要単語を抽出する
- 単語の重要度を記述する
- 重要な文章に意味のあるタグ付けをする
- 文章間の関連付けをする
- ある文章の文書内での重みを記述する

このような電子文書の構造化は文書の表現を豊かにし文書利用者の理解を助けるだけでなく，大量な文書を分類・整理する際の労を大幅に軽減し，また大量の電子文書に対して適合度の高い検索を行うことなどにも貢献する．

しかし，構造化作業は 2 次情報量を増やすにともない文書作成者の負担が増加するのが明らかである．また，文書の内容を完全に把握した者でなければ，意味のある情報を付加できないことは問題である．そこで，文書の構造化作業を計算機により自動的に行うシステムや，人間による構造化の負担を大きく減らしめるシステムについて盛んに研究されている．

3.2 構造化の手法

文書を構造化するとき構造を構成する最小単位は様々であるが，次のものが主な単位になるうる．

- 単語
- 文章の塊
  ある程度まとまりのある文章．
- 一つの完結した文書

構造を構成する単位にたとえば文書内での重要度を 2 次情報として付加したり，単位間の関係を 2 次情報として付加することにより構造化することができる．

計算機を利用した 2 次情報の取得方法は以下に分類される．

1. 表記の統計情報を用いた方法
2. 分類体系 (シソーラス) やオントロジーを用いた方法
3. 上記の両者を組み合わせた方法

統計情報を用いる方法 [18] は，単語の出現頻度などから文章の特徴付けを行ったり，文書の分類を行う方法である．この方法は計算機による処理が簡単で，実現が容易であるため汎用性が高い半面，表記による表層的な情報のみしか扱わないため得られる情報の精度が低くなってしまうことが指摘されている [10]．

シソーラスやオントロジーを用いた方法[21][9]では，シソーラスを用いて文書中の単語をグループ化する，またオントロジー(概念体系)を用いて単語の関係を表現することなどにより，文書の分類や単語の共起関係などを精度良く記述することができる．この手法では，分類体系への単語の割り当てやオントロジーの作成に人手が要する場合が多いが，電子辞書によるオントロジー作成の自動化による人手の負担軽減など現在盛んに研究されている．

統計情報とシソーラスなどを組み合わせた方法[10][12]では，統計情報をシソーラスで補うことにより分類精度を保ちながら人手の負担軽減をすることを目的としている．

3.3 文章を特徴付けるキーワードの抽出法

構造化のための2次情報取得の方法として，文章の特徴付けるためのキーワードの抽出方法について具体的に説明する．

ある一連の文書中の文章についてその文章をキーワードにより特徴付けることができれば，同じ特徴を持った文章，すなわち同じキーワードを持つ文章は同じような内容を持つと考えることができる．そこで，同じキーワードを持つ文章間の関係を文書に付加することにより文書の構造化をすることが可能となる．このようなことから文書中の重要キーワードを抽出する種々の手法が提案されている．

3.3.1 tf-idf 法

tf-idf 法は，ある文書中において一部の文章群の特徴を示すキーワードを統計的に抽出する手法で一般的なものである．

tf-idf 法では抽出するキーワードの評価基準として，ある一部の文章におけるキーワードの生起数 tf(term frequency, 相対出現頻度)と全体の文書におけるキーワードを含む文章の生起数 idf(inverse document frequency, 逆文献頻度)を考慮するところに特徴がある．Salton が提案した tf-idf 法[6]を以下に説明する．

- tf(term frequency) は，ある一部の文章におけるキーワードの生起数

    1. 文法的な機能を持つ語を削除する

2. ある文章 (ドキュメント) における tf_{ij}(あるドキュメント D_iにおける語 (t_j) の生起数を調べる

3. 生起数の閾値 TFを決定し，それぞれの文章 D_iについて

$$tf_{ij} > TF$$

を与える t_jをキーワードとする．

- idf(inverse document frequency) は，全体の文書におけるキーワードを含む文章の生起数を情報量で評価したもの

1. N個の文章の集まりの中で，ある語 t_jを含む文章の頻度を df_jとする．

2. その単語のその重みは

$$log_2 \frac{N}{df_j}$$

で与えられる．(以下，log_2を logと略記する．)

ここで，あるドキュメント D_iにおけるある語 t_jの重要度 w_{ij}は tfと idfを用いて，

$$w_{ij} = tf_{ij} * log \frac{N}{df_j}$$

と表すことができる．w_{ij}が大きいものがその文章の中で重要な単語と見做すことができ，これによりその文章のキーワードを抽出できる．また，キーワードにより文章の特徴を記述することができる．

3.3.2 χ^2法

何らかの手段により文書内容が属する分野の識別ができているとき，その分野に特有なキーワードを文書から抽出する手法にχ^2法 [18] がある．

χ^2法は個々の単語についてχ^2検定を行い，χ^2が十分大きい値となるときその単語の出現頻度が分野により偏りがあることがわかる．χ^2法を以下に説明する．

1. ある単語 i の分野 j における出現頻度を x_{ij}とする．

2. 総単語数を m，総分野数を n とすると，単語 i の理論度数 m_{ij}は

$$m_{ij} = \frac{\sum_{j=1}^{n} x_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}} \times \sum_{i=1}^{m} x_{ij}$$

となる．ここで理論度数とは全分野で等確率で単語 i が出現したと仮定したときの単語 i の分野 jでの出現頻度である．

3. 単語 i の分野 jでのχ^2項をχ_{ij}^2とすると

$$\chi_{ij}^2 = \frac{(x_{ij} - m_{ij}^2)}{m_{ij}}$$

となる．

4. 求める単語 i のχ_i^2値は

$$\chi_i^2 = \sum_{j=1}^{n} \chi_{ij}^2$$

となる．

今ある閾値χ^2を定めると，これ以上のχ_i^2値を持つ単語 i はある分野の文書で特徴的に出現するキーワードとして抽出することができる．

3.4 文書の文脈を把握する手法

個々の文章が特徴付けられているとき，文章の特徴が似ているものは同じような内容を持っていると考えることができる．このときどの程度類似性があるのか評価する必要がある．

3.4.1 ベクトル空間法

定めたキーワードの文章内での出現頻度をベクトルで表現したものにより文章の特徴表現し，ベクトル間の角度により文章の類似性を判断する手法 [6] である．

もっとも一般的な特徴ベクトルの生成法を以下に説明する．

一群の文章のそれぞれをキーワードの出現頻度によって特徴ベクトルを表現する．出現頻度を調査するキーワードを

$$word_1, word_2, word_3, ..., word_n$$

の n 個とし，文章は m 個あるとする．ある文章 i に含まれるキーワード jの出現頻度を v_{ij}とすると文章 i におけるキーワードの出現頻度は n 次元のベクトルで表現でき，

$$V_i = (v_{i1}, v_{i2}, v_{i3}, ..., v_{in})$$

とすることができる．

ここで，文章 a と文章 b の間の類似性 S_{a-b}はそれぞれの特徴ベクトル V_a，V_b の成す角度 (それぞれを絶対値 1 に正規化して内積をとったものの $\cos\theta$) で評価することができ，

$$S_{a-b} = \cos\theta = \frac{V_a \cdot V_b}{|V_a||V_b|}$$

とすることができる．$\cos\theta$の値が 1 に近いほど類似性が高いということができる．

単語の出現頻度により特徴ベクトルを生成する方法を発展させて，重み付けを施した後で文章の特徴ベクトルを生成する方法がある [5][4]．

単語の出現頻度による文章 i の特徴ベクトルを

$$V_i = (v_{i1}, v_{i2}, v_{i3}, ..., v_{in})$$

とする．tf-idf 法により文章 i における単語 kの重み w_{ik}をもとめて，文章 i の特徴ベクトル V_i'を，

$$V_i' = (v_{i1} * w_{i1}, v_{i2} * w_{i2}, v_{i3} * w_{i3}, \ldots, v_{in} * w_{in})$$

と定義する．文章間の類似度を特徴ベクトル V_i'の $\cos\theta$により評価することができる．

次に単語の共起を用いた特徴ベクトルの生成法 [19] を以下に説明する．

あるキーワード kは複数の文章で出現し，また共起するキーワードも存在する．そこで前述した文章 i での出現頻度ベクトルを利用してキーワード kの特徴ベクトルを考える．キーワード kの特徴ベクトルは，文章 i のキーワード出現頻度ベ

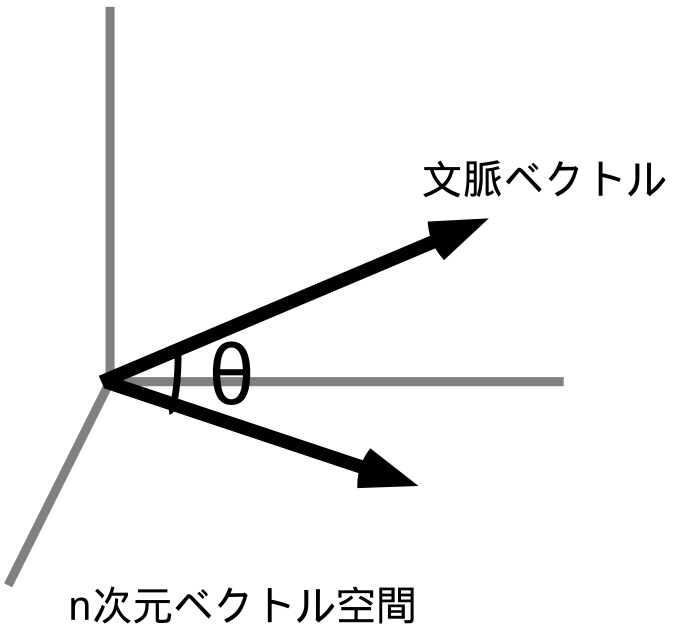


図 3.1 文脈ベクトル

クトルを正規化したベクトルに文章 i 内のキーワード kの出現頻度を掛けたベクトルを全文章で足し合わせたものと考えることができる．

今，文章が m ありキーワードが n あるとする．ある文章 i の単語出現頻度ベクトルを V_iとする．文章 i でのキーワード kの出現頻度を v_{ik}とすると，キーワード kの特徴ベクトルは

$$
\begin{aligned}
W_k &= (v_{1k} * \frac{V_1}{|V_1} + v_{2k} * \frac{V_2}{|V_2|} + \ldots + v_{mk} * \frac{V_m}{|V_m|} \\
&= \sum_{i=1}^{m} v_{ik} \frac{V_i}{|V_i|}
\end{aligned}
$$

とすることができる．

これを用いてある文書 i の特徴ベクトル A_iを記述すると，

$$
A_i = \sum_{k=1}^{n} (log \frac{m}{m_k}) v_{ij} \frac{W_k}{|W_k|}
$$

とすることができる．

ここで，m_kはキーワード kを含む文章の数である．

文章 a と文章 b の類似度は前述と同じ方法により，類似度を S_{a-b}とすれば

$$
S_{a-b} = \frac{A_a \cdot A_b}{|A_a||A_b|}
$$

とすることができる．

3.5 新しい空間ベクトル法の提案

3.5.1 従来の空間ベクトル法

これまでに説明した文脈の類似度の評価は，特徴ベクトル A_xと A_yを使って，

$$
S_{x-y} = \frac{A_x \cdot A_y}{|A_x||A_y|}
$$

によるベクトル空間における $\cos \theta_{x-y}$の値によるものであった．

この方法の利点は

1. 文章の類似度をスカラー値で表すことができる

2. 計算が簡便である

3. 直観的に理解しやすい

ことが挙げられる．

一方，単語出現頻度を基に生成した特徴ベクトルの $\cos\theta$ の値による類似度の評価は以下の点で問題がある．

1. 文章 a と文章 b の類似度と文章 c と文章 d の類似度を比較するとき $\cos\theta_{a-b}$ と $\cos\theta_{c-d}$ の値で評価することが困難である．

2. 全く類似性がないと考えられる文章間の類似度でも正の値をとることから，類似度がない文脈を抜き出すことが困難である．

3. 文章中の単語の出現頻度を基に文章の特徴ベクトルを生成するが，類似度の値に及ぼすベクトルの長さの影響が不明である．

4. 出現頻度を重視すると頻度 0 のものの情報が抜け落ちてしまう．すなわち，ある単語が目立って出現しないという情報が $\cos\theta$ の値に明確に反映しない．

ところで，ベクトル空間における二つのベクトルの $\cos\theta$ の値は一方のベクトルの要素を一つの変量の値と見て，2 つの変量間の単相関係数を求めることに似ている．

ここで変数 x_1 と変数 x_2 の間 (x_1, x_2) の単相関係数 r_{12} は，s_1 を x_1 の標準偏差，s_2 を x_2 の標準偏差，s_{12} を x_1, x_2 の共分散とすると

$$r_{12} = \frac{s_{12}}{s_1 s_2}$$

で定義される．

すなわち，2 変数の単相関係数は変量から変量平均を引いて作った 2 つのベクトルの $\cos\theta$ の値を求めたものと同じである．言い換えると，ベクトルの要素を標準化したベクトル間の $\cos\theta$ の値が単相関係数である．なお，単相関係数は 2 変量の相関関係を-1 ~ 1 の値で示したもので，2 変数の線形回帰の程度を表す．

ベクトル空間においてベクトルの各要素は独立で重みは同じである．しかし，単語出現頻度を用いた特徴ベクトルはベクトルの各要素の重みが異なってしまっている．特徴ベクトルの各要素を変数とみて，2 変数間の単相関係数を求めることは 2 変数の関係から適当でないと考えられる．

3.5.2 単語出現頻度データの標準化

単語の出現頻度を用いた文章の特徴ベクトルの類似性の比較は，出現頻度が 0 以上の値であることにより問題が生じることがわかる．そこで，統計解析における変量標準化の手法を使って単語出現頻度の正規化を行うと，単語の出現頻度をより統計的に見積もることができる．

ここで標準化とは，変量の平均を 0，分散を 1 にすることで，確率変数 Y，Y の平均をμ，分散をσとすると，標準化された変数 Zは，

$$Z = \frac{Y - \mu}{\sigma}$$

となる．

正規分布に従う変量を標準化した変量は標準化後も正規分布に従う．したがって，通常は正規分布に従う変量に対して標準化処理を行う．

ある文章における単語 i の出現頻度 v_iを変量と見て，文章集合 Dにおける v_iの分布を考えると，v_iの分布は 2 項分布に近似した分布となると想像される．2 項分布と正規分布の関係から，単語の出現頻度が正規分布に従うと考え標準化を行う．

標準化した単語出現頻度により文章の特徴ベクトルを作ると，文章 i でのキーワード kの出現頻度を正規化したものを v_{ik}とすると，文章 i の特徴ベクトル A_iは

$$A_i = (v_{i1}, v_{i2}, v_{i3}, \ldots, v_{in})$$

とすることができ，文章間の類似度はベクトル間の $\cos\theta$の値で表すことができる．

標準化により，個々の単語の出現頻度に関して統計的に正規化された頻度を得ることができる．たとえば，単語 kの文章 i での出現頻度を v_iであるとき，

$$K = (v_1, v_2, v_3, \ldots v_n)$$

として，今仮に

$$K = (3, 3, 2, 3, 3, 4, 5, 0, 2, 1)$$

とすれば，標準化により

$$K = (0.294, 0.294, -0.441, 0.294, 0.294, 1.03, 1.76, -1.91, -0.441, -1.12)$$

のようになる．

標準化により実際の単語出現頻度の大きさに関わらず，出現頻度データの重みを同じにすることができる．また出現しないということが負の出現頻度で表現される．

3.5.3 プロトコル解析文の文脈の解析例

設計実験におけるプロトコル解析で得られた設計情報について，ベクトル空間法による文脈解析実験を行った．

プロトコル解析文の特徴は

- 構造化されていない．
- 会話の記録のため一文が短い
- 一文自体に含まれる意味が少なく，前後の複数文の文章群で意味をなす

ことがあげられる．解析に用いたプロトコル解析文の一部を図 3.2 に示す．

このプロトコルは，3 人の設計者 (あまり自転車設計に詳しくない) にテーマと時間と資料を与えて議論してもらい，そこで行われた設計に関する会話をテキストデータ (約 2500 行) としたものである．テーマはリュック (backpack) をマウンテンバイクに取り付けることが可能な荷台または金具の設計である．

このプロトコル解析文は段落や話の区切れなどの情報を表層的に含んでいない．従って tf-idf 法やχ^2法を使ってキーワードを抽出し文脈の類似性を計算するには人手により文章を区切る必要がある．ここでは文章の区切りがない文書から文書構造を意味する文の区切りを見つけることを目標に，文脈類似度を求めた．また，単語の出現頻度による特徴ベクトルを使った文脈類似度の値と頻度を標準化した特徴ベクトルを使った文脈類似度の値を比較した．

```
661 G because you have .. your backpack frame now
has the fitting on it for the seat tubing so you
don 't have everything in this bracket part of the
backpack frame has like a a metal clip y'know wire
em .. that type of thing
662 H ahha ahha
663 G y'know just thin gauge metal that snaps .. a
broomhandle holder
664 H yeah ahha
665 F ohh
666 G any justice here
667 H and that will live permanently with your
backpack
668 G yeah .. and you can make that wide enough
669 F and that to be just a holder while you're ( ?)
or part of the integral fastening system
670 G it keeps it attached to your seat tube
671 F at all times .. structural (inaudible) OK
(inaudible)
672 G maybe there's a little hook here that you
fasten after you snap it on just to make sure it
doesn 't come off
673 F ahha yeah yeah that sounds
674 G then you ..these wires come down here right
clip on .. clip on
675 X sorry could I interrupt could I ask you to put
the bike on the table when you're talking specifically
about the bike
676 G sure yeah
677 X OK sorry to interrupt
678 G so .. clip on those tubes and that pivoting
over here
679 H mm mm
680 H and they actually don't have to pivot there ..
and it rocks up .. and clicks in
```

図 3.2 自転車荷台の設計に関するプロトコルの一部

特徴ベクトルの求め方は以下の通りである．

1. 文書全体を第 1 行目から 50 行ごとに区切る．図 3.5.3参照．
2. 出現するすべての単語について，それぞれの文章群で単語の出現頻度を計算する．また，それぞれの単語について数値の標準化を行う．
3. 出現頻度または標準化された出現頻度により特徴ベクトルを作る．
4. 隣合う特徴ベクトルの $\cos\theta$を計算する．
5. 次に第 2 行目から 50 行ごとに区切り，同様に単語の出現頻度と標準化を行い特徴ベクトルを作成する．隣合う特徴ベクトルの $\cos\theta$を計算する．
6. 以上を繰り返す．

得られた文脈類似度の値をグラフにした図を図 3.4，図 3.5に示す．図で横軸は行番号，縦軸は $\cos\theta$の値である．行番号 i のデータは $i-49$ ~ i 行の文章と $i+1$ ~ $i+50$ 行の文章との類似度を表す．

3.5.4 解析実験の考察

実験結果から標準化した単語出現頻度による特徴ベクトルによる手法に以下の特徴が見られた．

利点

- 統計的手法のみに依存するため，キーワードの選別などに人手が必要ではない．
- 単語の重み付けを考慮する必要がなく，重要な情報は標準化により自動的に重みを付けることができる．
- 文脈の類似度が正であると類似性があり，負であると類似性がないと数値的に判断できる．

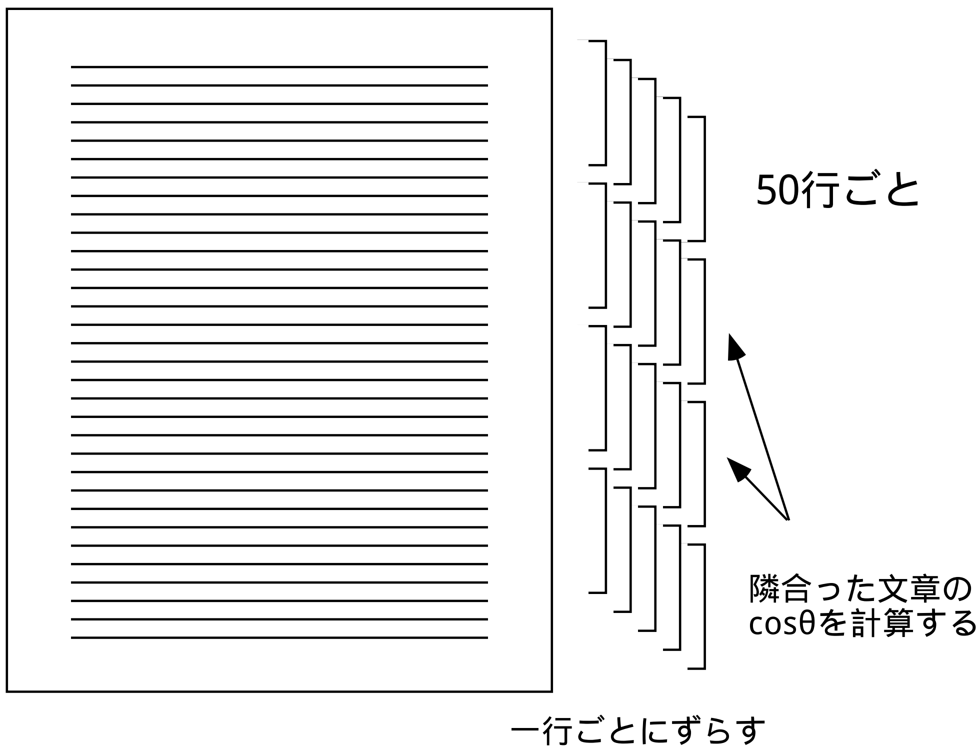


図 3.3 文章群の切り分け

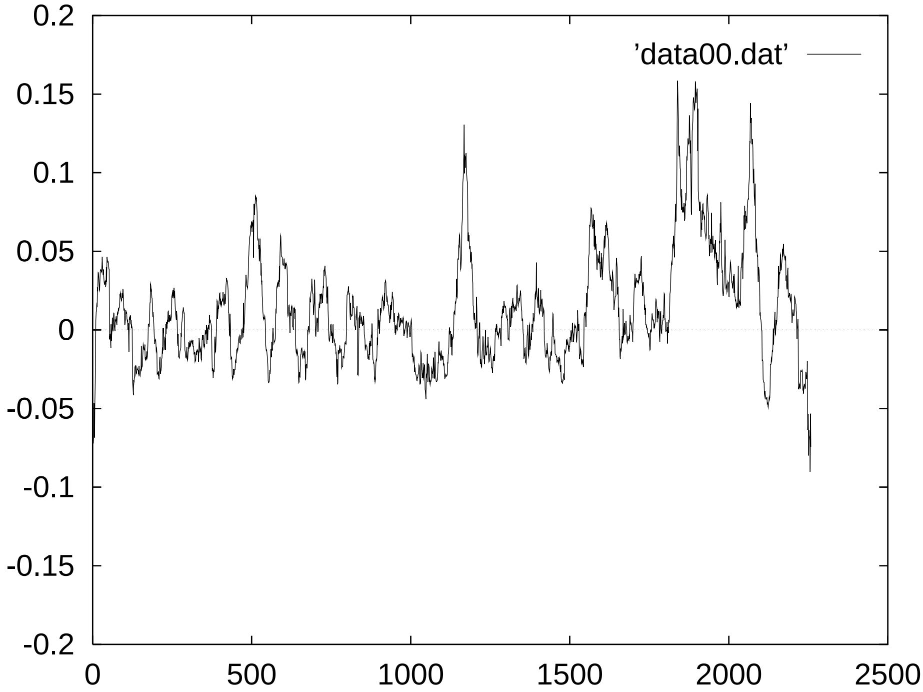


図 3.4 データ標準化した後の文章類似度

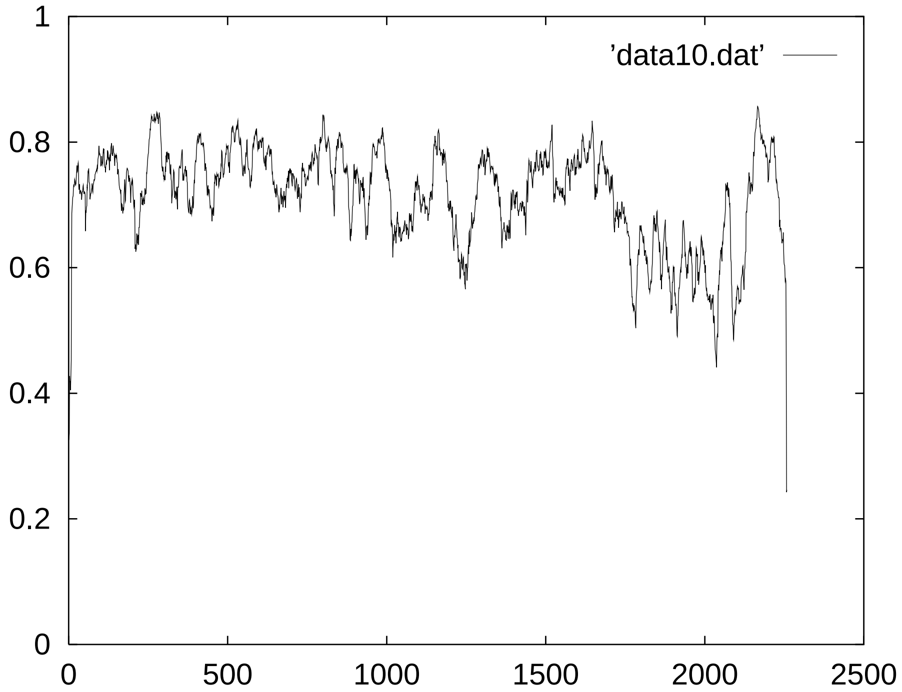


図 3.5 単語出現頻度を用いた文章類似度

- 標準化処理において単語出現頻度の平均と分散を求めるが，一般的な文書についてある単語の出現頻度の平均と分散を求めておくことにより，未知の文章について疑似的に標準化を行うことができ，未知文章と既知文章の類似度を評価することができると考えられる．

問題点

- ある文章における単語の出現頻度は 2 項分布に近似した分布を示すと考えられるが，出現頻度が極端に小さいと標準化が適切でないと考えられる．すなわち，平均μが十分に大きく，標本数 n が極限に近ずくと正規分布に近似されるが，平均が小さく標本数がそれほど大きくない 2 項分布は正規分布に近似できない．

- 類似度の絶対値が小さくなってしまい，数値の比較をすることに精度の疑問が生じる．

本実験では出現するすべての単語をキーワードとしたが，類似度の数値が極端に小さくなってしまうのはキーワードの数が多すぎることがその一因と思われる．しかし，キーワードの抽出に出現頻度の平均と分散の値を用いることが考えられ，類似度の精度を良くすることが可能であると思われる．

3.5.5 文脈解析実験のまとめ

本実験はプロトコルデータのように構造のない文書の文脈解析において，文書の構造 (区切り) の発見および文脈の類似度の評価に効果的であると考えられる．

第 4 章

文書構造化の表現としての弱情報構造

4.1 弱情報構造

文書を構造化したときの情報構造として HTML や SGML があるが，これらは強い情報構造ということができる．

HTML (HyperText Markup Language) や SGML (Standard Generalized Markup Language) は，文書の体裁を整えかつ論理的な構造を表現する言語である．HTML や SGML には以下の特徴がある．

- 文書構成要素である単語や文章にマークアップと呼ばれるタグ付けが施されており，システムに依存しない文書の体裁を整えることができる．
- 文書構成要素に論理的な構造を示すタグ付けが施されており，文書の構造を明確に指示することができる．

このように，HTML や SGML は文書の体裁，構造を明確に記述し，情報の受け取り側における意味の曖昧さの発生をなくすことを目的としている．

一方，弱い情報構造の例に連想構造がある．連想構造は情報の構成要素の間に，key と呼ぶいくつかの要素と value と呼ぶいくつかの要素の間に定義され,「key が与えられると value が想起される」というゆるやかな関連を表すものである．key と value の間の連想関係は厳密に定義されるのではなく, 多分に主観的であることを許している．一般に弱情報構造とは関連の意味を定義することなく関連の有無のみを記述する情報表現である．

弱情報構造による情報の記述は，述語論理や意味ネットワークに比べて概念や関係の記述に制約が少ないことが特徴である．従って，文書の構造化のために 2

表 4.1 文書要素

作業計画書	
題目	ガタ付き検出用アクチュエータ駆動回路の開発
作業期間	'95 年 11 月 1 日 ~ '96 年 3 月 31 日
Body	paragraph を含んだ主テキスト
Paragraph	テキストもしくは引用句を含んだパラグラフ

次情報を付加するとき，SGML や HTML では文書内容を熟知した人間が記入する必要があるのに対し，弱情報構造による文書では文書の内容を詳細に理解しなくても情報付加ができる可能性がある．

そこで，本章では弱情報構造を使うことにより，構造化されていない文書を如何に人に負担なく構造化し，有益な情報を引き出すことのできる情報形態に変換することの可能性について考察する．

4.2 文書の弱構造化

一般的に文書の論理構造は，論理要素の木構造により表現できるとされている [2][3][8]．文書に木構造で表される論理構造が存在することは Electronic Publishing の分野では前提となっている事項である．

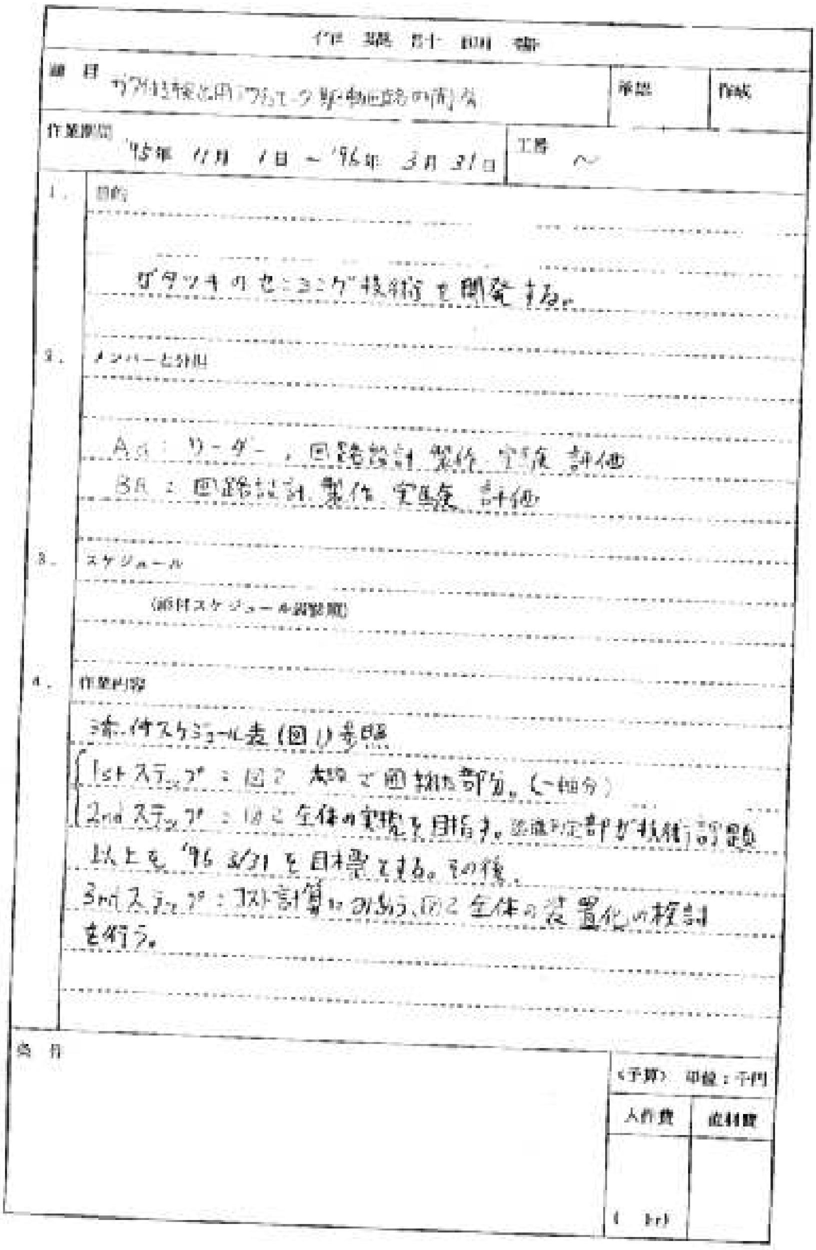
今，図 4.2の文書があったとき，この文書の論理構造を木構造で表すと，その一例として図 4.2により表現できる．ここで文書の要素は表 4.1に示す．

ある文書を木構造にする作業を複数の被験者が行うことを考えると，個々の被験者によって作成される木の構造が異なることが考えられる．

ここで，文書木構造 4.2を見ると，上のノードを key，下のノードを value と考えた簡略な連想構造と見ることができる．すなわち弱情報構造を用いて文書に 2 次情報を付加するとこの木構造を表現することが可能であることがわかる．

このような方法を用いて文書の構造化を行った場合，利点として

- 情報構造化した文書の作成が簡便である．
- 文書の論理的内容を 2 次情報として付加することができる



作業計画書

題目 ガタツキ検出用ブラシモータ駆動回路の開発

承認 作成

作業期間 '95年 11月 1日 ～ '96年 3月 31日

工番 ～

1. 目的

ガタツキのセンシング技術を開発する。

2. メンバーと分担

A氏：リーダー，回路設計 製作 実験 評価

B氏：回路設計 製作 実験 評価

3. スケジュール

（添付スケジュール表参照）

4. 作業内容

添付スケジュール表（図1）参照

1st ステップ：図2 太線で囲まれた部分。（一相分）

2nd ステップ：図2 全体の実現を目指す。…部が技術課題

以上を '96 3/31 を目標とする。その後、

3rd ステップ：コスト計算から、図2 全体の装置化の検討を行う。

備考

（予算） 単位：千円

人件費	直材費
（ hr）	

図 4.1 文書例

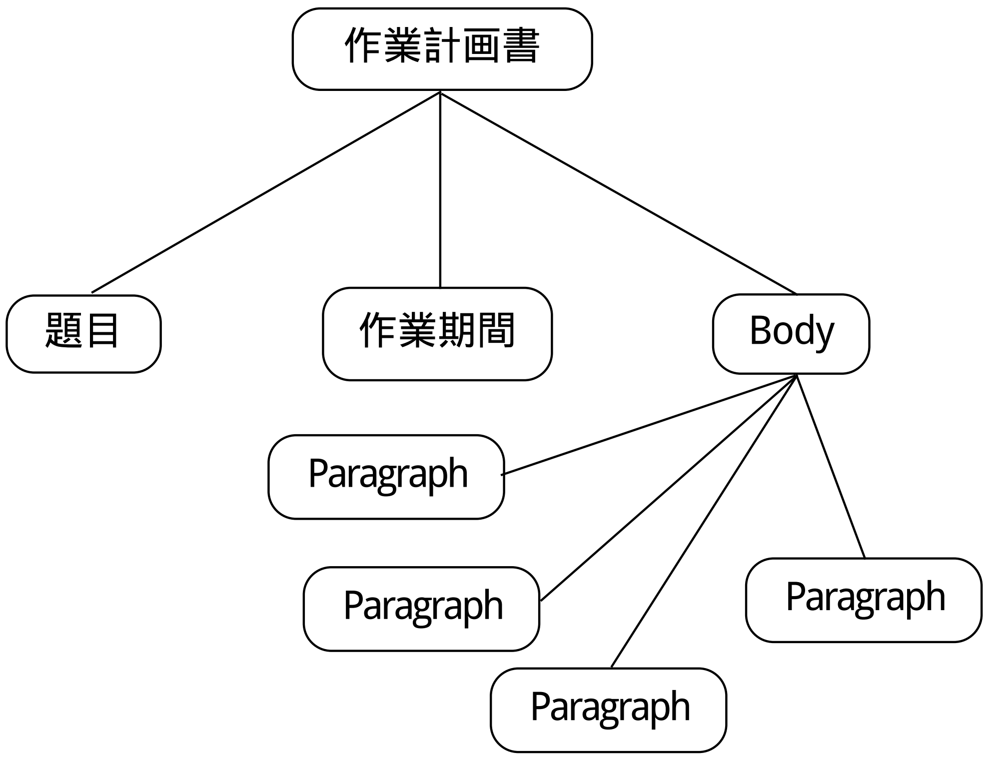

作業計画書
題目
作業期間
Body
Paragraph
Paragraph
Paragraph
Paragraph

図 4.2 文書木構造

ということが挙げられる．
一方，

- 2 次情報が多くなりすぎると構造に意味がなくなってしまう．
- 文書を解析する場合に新たな情報がどの程度得られるか不明である．

ことが欠点としてあげられる．

第 5 章

設計情報可視化システムの構築

実際の設計現場における設計情報を記録した記述形態に，設計会議の議事録や設計者のメモ書きがある．これらの情報はテキスト，図，グラフ，表など多様なドキュメントで記述され，意味的に構造化されていない非定型性を特徴とする．しかし，それらを知識メディア化することにより，設計作業の過程，設計対象の情報，問題解決過程など有益な「設計知識」[11] を取得することができれば，議事録を設計情報として再利用することが可能となり，設計者の負担軽減など設計の効率化に貢献すると考えられる．

設計作業中に残される議事録は，設計プロセスを詳細かつ忠実に表現するものではなく，その時点における問題，確認した事実，目標などが記され，言わば時系列的に整理されたメモ書きといった体裁を示し，前述の局面よりさらに小さな場面を記述している．すなわち，議事録はその場面における問題の対処を記述しており，全体で問題解決過程を記録していることになる．以下では，問題解決過程を設計知識として抽出することを主眼にアクチュエータ駆動回路設計における議事録の解析をし，設計情報の弱構造化実験を行った．議事録はある機械設計のための設計会議の記録で，Ａ４用紙２４枚のテキストとＡ４用紙２９枚の図表からなる．

5.1 議事録の解析

図 5.1 に，今回システム試作において例題として用いた設計会議議事録の一部を示す．議事録を観察すると，次のような特徴を見出すことができた．

- 一つの意味を表す文などの情報片が比較的小さい．

（1）2 ．ブリッジ回路と差動増幅回路との比較
（2）　（図 2 参照）
（3）　・ブリッジ回路と差動増幅回路の出力に位相差が生じていた。
又、波形もきれいな正弦波でなかった．
（ゆがみが時間によって変化していた。ファンを
回していた時だけおこった）
（4）　→原因を考える
（5）　・ファンなしで時間（１５分位）たって
出力波形は安定してきた。
（6）　→問題ないと判断した
（7）　→使い始めるためには準備時間が必要でないか？

図 5.1 設計会議議事録例

- 情報片間の意味的な流れが自明でないことが多々ある(熟練者ないし設計者の頭にある前提知識が入る)．
- 電子回路における概念は，is-a 関係や has-some 関係以外に明確な説明の付けにくい関係が重要であることがある．
- 図やグラフなどの重要性が高い．

これらの特徴をふまえて議事録の情報から

1. 文章の固まり (ブロック)
2. 図，グラフ
3. 電子回路用語，一般用語

を取り出し，それらの関係を分析し弱い関係付け (弱構造化) を行うことが有効であると判断した．

5.2 設計情報構造化の過程

アクチュエータ駆動回路設計の議事録の構造化を以下の過程で行った．

1. 議事録の内容理解
   議事録の理解と議事録上の電子回路特有の前提知識や用語を取り出すため電子回路の知識を持つ者と、持たない者の 2 名によって議事録から議事の再現を行った。ここでは議事録に以下の作業をおこなった。

   - 前提知識の付加、議論の流れの明確化：電子回路の知識を持たないと理解できない部分や議論の流れが明確でない部分に情報を付加した．
   - 用語の取り出し：電子回路における重要用語を取り出した．
   - 教科書の参照：前提知識や取り出された用語を説明する教科書上の知識をさがし、参照場所を付加した．

2. 議事録のブロック化
   上の作業を基にブロックの種類としては以下のような種類を用意して、各々の種類のブロックを作成した．

   - 議事録文ブロック：議事録に出現する文章で 1 ブロックで 1 つの意味を持つ。議事録の文章部分からテキストデータとして作成する．
   - 議事録図ブロック：議事録に出現する図ないしグラフ。議事録の図・グラフを画像ファイルとして作成する．
   - 用語ブロック：取り出された電子回路における重要用語。上の作業の結果を基にテキストとして作成する．
   - 用語説明ブロック：重要語に関する教科書などの説明。教科書などから該当個所を抜き出して作成する．
   - 付加説明ブロック：付加された前提知識や議論の流れを説明する文や図のブロック。上の作業の結果、参加者のコメント・メモを基に作成する．

3. ブロック間関係の抽出
上記から用語を除く４つのブロックについてそれらの間の関係を付加した．作業の手順としてはブロック間に何らかの関係がある場合に有向または双方向リンクを張った．内容的に意味の明確なリンクについてはその意味を付加した．我々のリンクに付加した意味を以下に挙げる。

   - 進行状況
   - 日時
   - 理由
   - 今後の方針
   - アドバイス
   - 新たな問題点
   - 図参照

   意味の付加しなかったリンクは自然な議論の流れとみなしたものである．

4. 設計知識の CM-2 上への実装
以上のブロックやリンクを電子化された議事録（テキスト、スキャンされた図）から、弱構造化メディア CM-2[16] へ実装した．実装にあたって行った作業を以下に挙げる．

   - 議事録図ブロックはイメージとして取り込み，固有のＩＤをラベルとした．
   - 用語は用語そのものの名前をもつ概念として取り込んだ．
   - 付加説明ブロックは議事録ブロックと同様に扱った．
   - リンクに付けた意味を概念として登録した．
   - 議事録文ブロックに関して用語が含まれる場合対応する概念に連想関係のリンクを張った．この作業はプログラムにより自動的に行った．
   - ブロック間リンクに連想関係のリンクを張った．ブロック間リンクに意味がある場合その概念をキーとして付加した．

目標	→	問題	→	事実	→	新たな問題
(1)		(1)		(3)		(4)
(1)		(1)		(5) (6)		(7)

図 5.2 ブロックの構成と文章内容の言明

- 用語や付加説明ブロックについては電子回路についての概念がより分かりやすいように情報を付加した．
- 検索や編集のしやすいように議事録，文脈 (リンクの意味)，日時に関してまとめたワークスペースを作った．

また，作成した文脈と文章の関係を観察すると互いに近傍にある文章群が図 5.2 のような典型的な構成を持つことがわかった．これらは局所的問題の解決と新しい問題の発生過程を表現していると考えられる．もちろん，全ての文章がこのような構成をとる訳でないが，その他はこの構成のサブセットと考えることができた．

ここで，各文章を目標，問題，事実，新たな問題というように言明したとき，解決した「問題」を中心にした塊と捉えることができ，そのとき，塊内の文章群を言明により意味的に定型化したことになる．今，図 5.1 のデータについてこの考えを当てはめると，図 5.2 の言明に基づき，

$$((1),(3),(4)) \text{ と } ((1),(5),(6),(7))$$

にまとめることができる．

次に，ある文章にある言明 (目標，問題，事実，新たな問題) を与える事ができたとき，文章は自然言語としてどのような特徴を示すか観察し，図 5.3 の結果を得た．

文章の言明とこれらの特徴が必ずしも一致する訳ではなく，また一つの文章が複数の意味を持つことや，一つの文章が複数のブロックに含まれることも確認した．しかし，ある一文が与えられたとき，それが議事録中でどのような言明を与えることができるか，それを推測する判定規準の一つになりえる．

目標	(a) 名詞での言いきりで文章が終る (b)「～したい」で終る (c)「～する」で終る
問題	(d) 疑問文である (e) 文章の最後に"？"の記号である (f) 最後の述語が否定されている (g) 目標で言明された文に含まれる単語に関連する単語を含む
事実	(h) 最後の述語の時制が過去形のである (i) 最後の述語が形容詞である (j) 目標，問題で言明された文に含まれる単語に関連する単語を含む
新たな問題	(k) 最後の述語が現在形で終る (l) 疑問文である (m) 目標，問題，事実で言明された文に含まれる単語に関連する単語を含む

図 5.3 言明と文章表現の関係

5.3 CM-2 上に弱構造化された設計知識

この節では前節の過程で CM2 上に実装されたアクチュエータ駆動回路設計の議事録を様々な視点からブラウズした例を挙げる．

- 専門用語から辿る例
  専門用語"差動増幅回路"から様々な情報を引き出す実験を行った．図 5.3は差動増幅回路を中心に近傍探索した結果である．ここで"y"で始まるノードは議事録文ブロック、"数字-g"で始まるブロックは議事録図ブロック，"s"で始まるブロックは付加説明ブロック，"t"で始まるブロックは用語説明ブロック，それ以外が用語ブロックである．
  概念"差分回路"や"減算回路"は"差動増幅回路"の別名であり，t48 は差動増幅回路の教科書における説明への参照である．
  y440，y441，y478 で始まるラベルをもつテキストは内部に用語"差動増幅回路"を含んでいる．すなわち，これらの議事録の段階で"差動増幅回路"が言及されたことを意味している．y440 から y441 へのリンクは議論の流れを

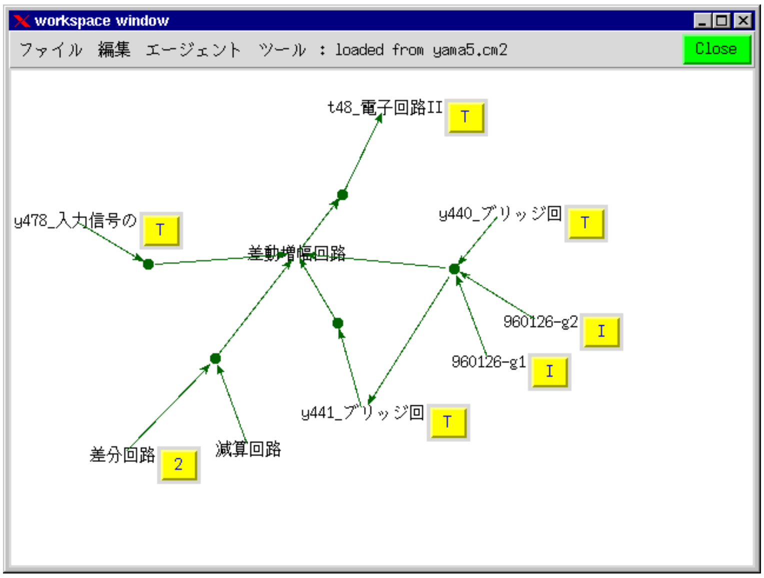


図 5.4 ”差動増幅回路”から検索した結果

示しており，議事録の上で直接関係があることがわかる (図 5.3参照)．また，これらの文中で述べられている位相やノイズはグラフ 960126-g1、960126-g2 に見つけることができる (図 5.3参照)．
このように用語を中心として他の用語，教科書知識，議事録文やグラフを得ることができる．

- 議事録中のブロックから文脈を辿る
  まず始点として図 5.3のように議事録 5 に関係あるすべてのブロックを表示させる．次にこの中の議事録文ブロック１つ (y421) を選んだ。次にそれに関係のある情報を得るため近傍探索を行った結果が図 5.3である．y415 → y421 → y422，y423 のリンクの流れが得られている．これらは y421 の前後の文脈を意味している．s06，s07 は後で付加された説明文．図を表している (図 5.3参照)．

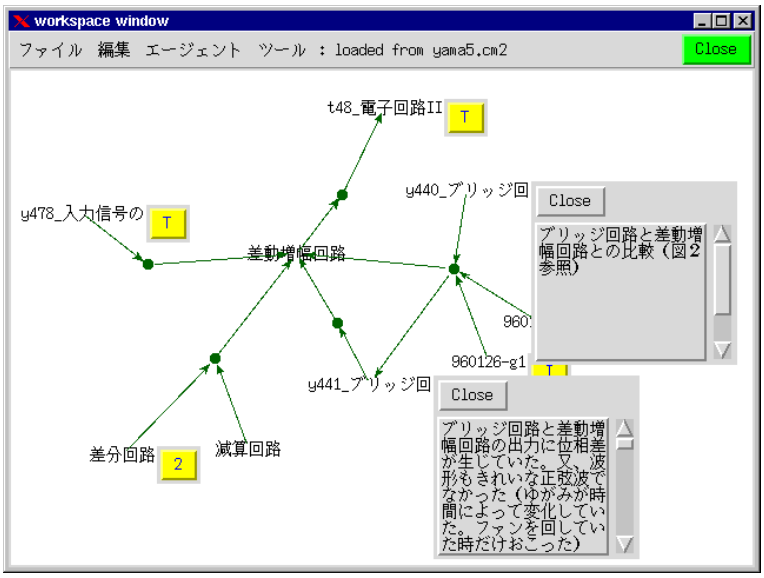


図 5.5 ”差動増幅回路”から検索した議事録文ブロックの内容

5.4 設計情報の弱構造化実験の評価

ここでは，CM2 上に設計情報を弱構造化したこの実験に関する評価を述べる．利点として，

- 議論の流れや議論間の関係が画面上で見て辿ることができる．
- 様々な視点から設計知識を見てゆくことができる．
- ユーザーの予想のしなかった思いがけないリンクを見つけることがきた．
- テキスト、図、グラフなど複数のメディアを同時に共存できる．
- 設計知識 (データ) を容易に追加，変更できる．

が挙げられる．欠点として，

- 表示が煩雑になり見にくくなることがある．

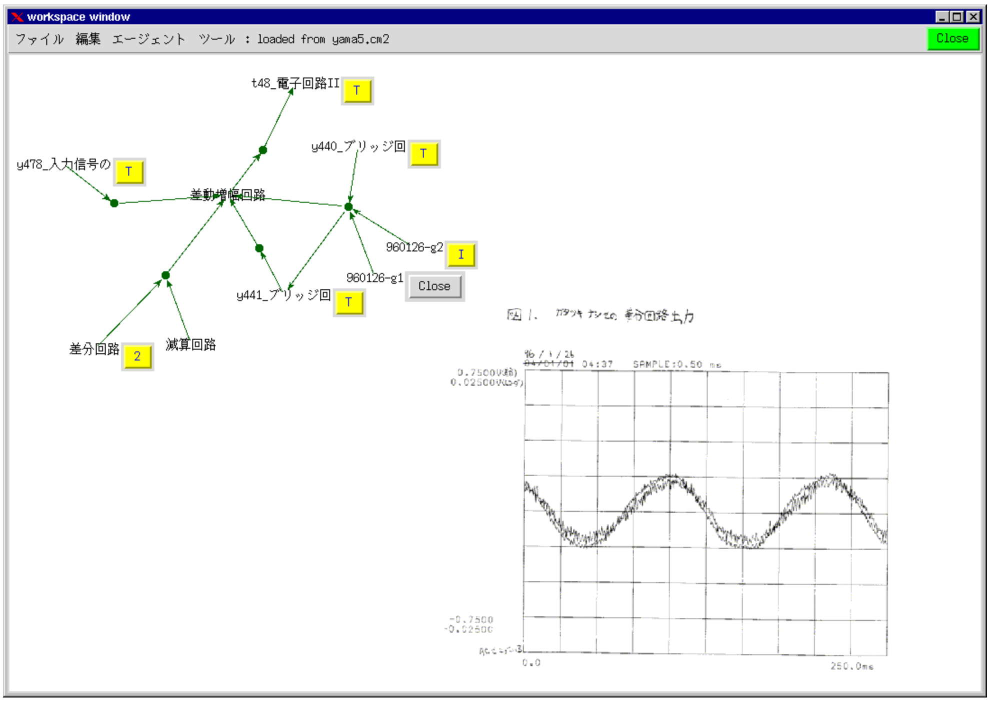


図 5.6 ”差動増幅回路”から検索した議事録図ブロックの内容

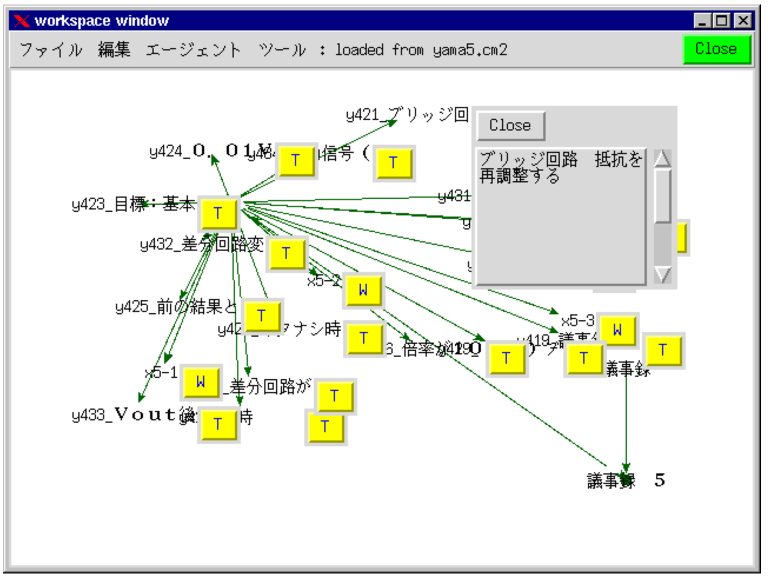


図 5.7　議事録から文脈をたどる例 (1)

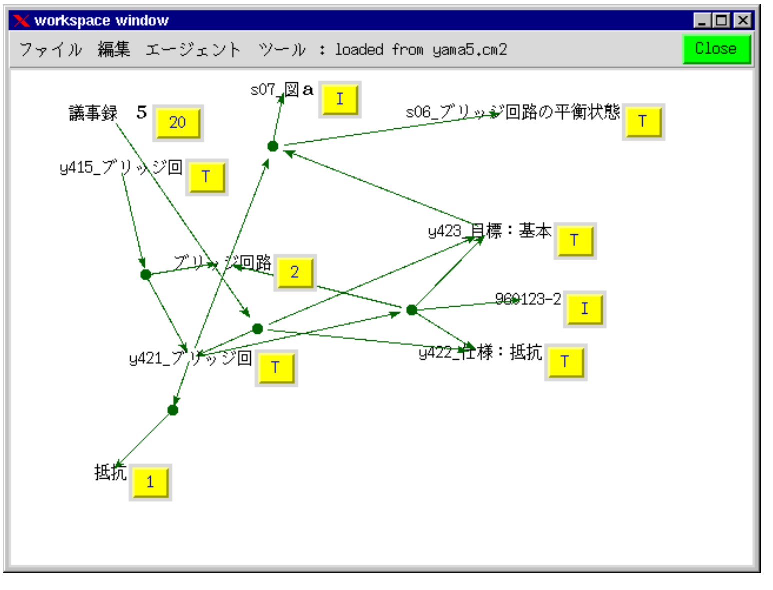


図 5.8　議事録から文脈をたどる例 (2)

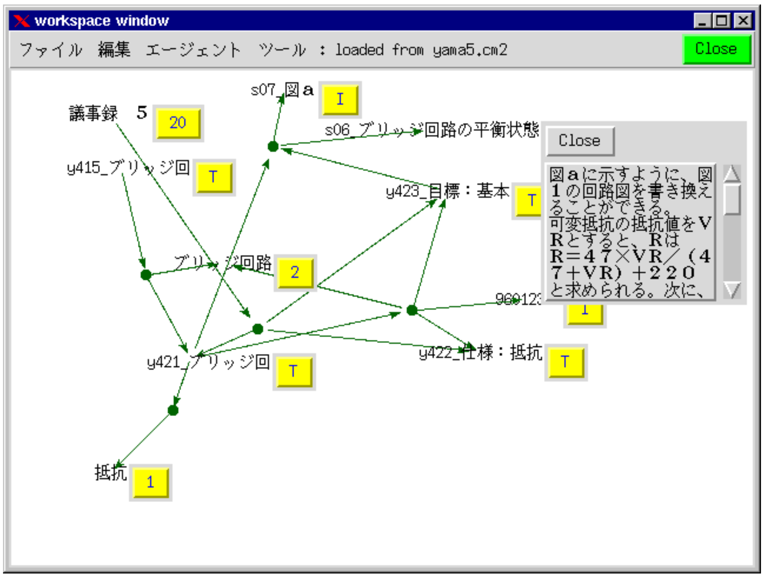


図 5.9 議事録から文脈をたどる例 (3)

- データ (文書) の入力時に内容を理解する必要がありその作業に時間がかかった．

利点とも欠点ともいえる大きな特徴は，様々な視点に基づいて知識間のリンクを付ければ付けるほど解りやすく整理された情報が得られることである．

5.5 設計情報からの問題解決抽出

前節に基づき，ある文章の言明をすることを助けるような文章間の構造を記述する情報が存在すれば，文章を自然言語処理した情報を組み合わせることにより，問題解決を意味する冗長性の少ない定型化された文章群を抽出することが可能であると判断した．

具体的に，図 5.1 の文章が，図 5.10 のように弱情報構造の情報を付加されたなら，((1),(3),(4)) のブロックは以下の手順で導くことが可能である．

- (1) は目標である–(a)

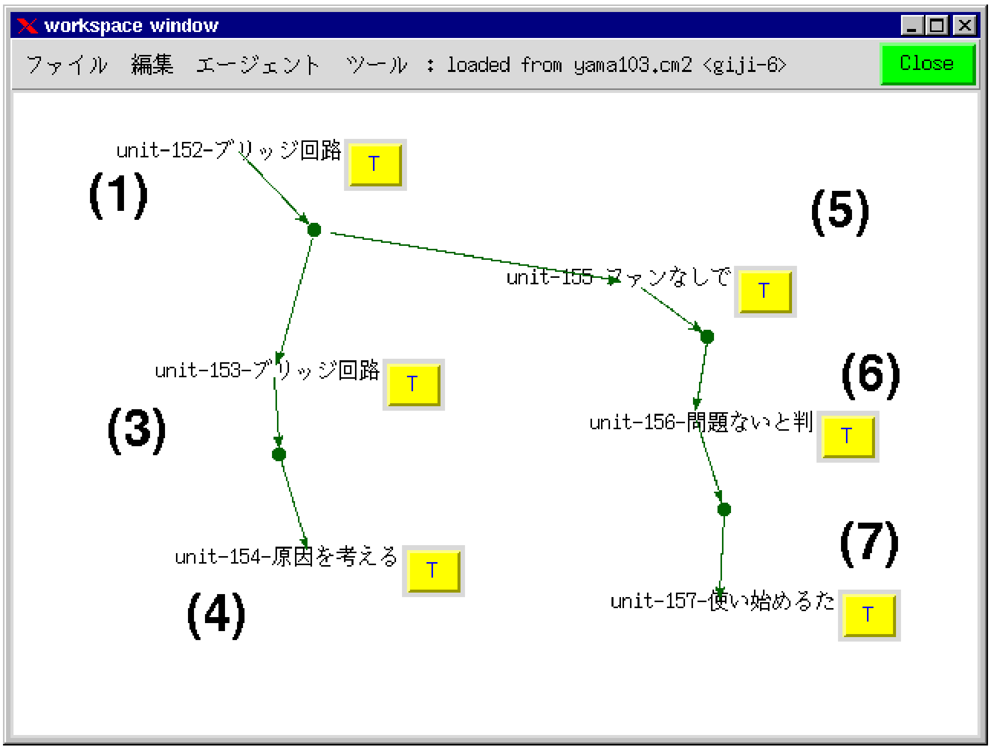


図 5.10 弱構造化された文章群

ドキュメント前処理部

↓

ユニット弱構造化部

↓

ブロック化部

↓

情報知識表現部

図 5.11 設計情報可視化システム概略

- (1) とリンクされた文を探す
- (3) は問題か事実である–(g)(h)(j)
- (3) とリンクされた文を探す
- (4) は新たな問題である–(k)

以下にシステム概要を説明する (図 5.11参照)

- ドキュメント前処理部
  非定型ドキュメントからテキストを抽出し，一文を一つのユニットとして取り扱い，CM-2 形式データ (S 式) に変換する．議事録などの非定型なテキストの場合，文の区切り文字や接続語が少ないため，意味のある一文の切り出しが困難である．そこで，一文の区切りは人の手を介在して行う．s 式にしたテキストを図 5.12 に示す (簡単のため連想関係の記述を省略した)．また，CM-2 システムは画像データなど非テキストデータも統一的に扱い，表示をすることができることから，画像データも一つのユニットとする．
- ユニット弱構造化部
  ユニット間に弱い情報構造 (連想関係) を構築する．前節で述べたブロック

```
(unit-6 (label "unit-6-ブリッジ回路") ("連想関係データ")
(data "ブリッジ回路と差動増幅回路との比較(図2参照)"))
(unit-7 (label "unit-7-ブリッジ回路") ("連想関係データ")
(data "ブリッジ回路と差動増幅回路の出力に位相差が生じていた。
       又、波形もきれいな正弦波でなかった
      (ゆがみが時間によって変化していた。ファンを回していた
       時だけおこった)"))
(unit-8 (label "unit-8-原因を考える") ("連想関係データ")
(data "原因を考える"))
(unit-9 (label "unit-9-ファンなしで") ("連想関係データ")
(data "ファンなしで時間(15分位)たって
       出力波形は安定してきた。"))
(unit-10 (label "unit-10-問題ないと判") ("連想関係データ")
(data "問題ないと判断した"))
(unit-11 (label "unit-11-使い始めるた") ("連想関係データ")
(data "使い始めるためには準備時間が必要でないか?"))
```

図 5.12 CM-2 データ

内の言明の構成を参考にし，議事録における議論の流れ，問題解決手順の明確化をするため，話の流れや文脈に沿ってユニット間に連想関係を付加する．

- ブロック化部
  ユニット間の弱情報構造とユニット内テキストの自然言語処理により，ブロックを抽出する．これらの機能を Lisp 言語で実装した．なお，自然言語処理には日本語形態素解析器茶筌 Version 1.0[13] を利用した．

- 情報知識表示部
  構築した情報ベースを CM-2 システムにより視覚表示する．なお，設計知識の可視化には，弱い情報構造における情報の収集，編集，ブラウジングを支援する CM-2 システムを利用した．

試作システムを用いた問題解決過程情報の CM-2 システムによる表示例を図 5.13 に示す．

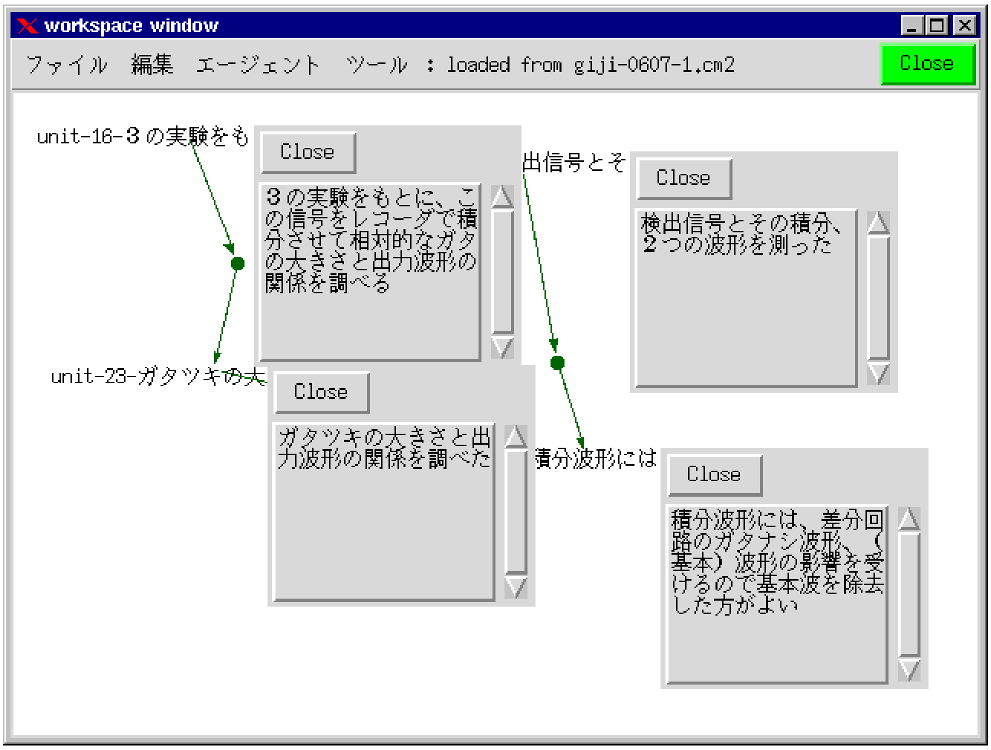


図 5.13 問題解決過程情報の表示例

5.6 問題解決仮定抽出実験結果と考察

試作システムによりブロック化されたユニット群と議事録を基に手作業でブロック化したユニット群を比較すると，試作システムではブロック化されないユニットが存在することがわかった．特に文章構造が複雑で，一つの文章にブロックが複数重なるとき，ブロック中に入って欲しい文章をブロック化できなかった．これは，試作システムでは自然言語処理の実装がまだ不完全であることに起因する．

ブロックの表示において，CM-2 システムは問題解決過程の表示に不向きであることが指摘された．問題解決過程の理解を支援するためには，設計知識の効果的な表示があることが示唆された．

議事録の自然言語処理では，文末の時制や活用形よりも関連単語のマッチングが言明の判定の要因になることが多かった．そこで，関連単語リストを使う事が効果的であることがわかった．例えば，

(”積分” ”積算”) (”ガタ” ”ガタつき” ”ガタツキ”)

は前者は厳密な意味は異なるが似ているもの，後者は意味は同じだが表記の異なるものであるが，双方の単語リストを作成することが必要である．議事録の場合，記録者が複数人であることが多いと思われるので，このような記録者間の記述の違いを受け止めるには特に重要であると思われる．

実験目的は，試作システムの設計情報可視化の効果を評価することである．したがって，議事録における設計目的を正確に理解し，設計対象に付随する基礎知識，背景知識を身に付けた専門家よりも，逆にその設計対象に素人の人物の評価が重要であろう．本実験では第三者による評価を得なかったが，議事録をブロック化し，問題解決過程の表現を部分的に行うことができ，設計プロセスの一部を提示することができた．これは第三者による設計の理解に有効であり，設計支援に繋がるものであると思われる．

会議議事録のような非定型ドキュメントは，一般的に次のような特徴を持つ．

- 照応関係の不完全さ
- 状況記述の欠如
- 文章間における論理の飛躍が多い
- 論理関係の記述がほとんどなく，一文が簡潔である
- 人によって文章記述の仕方が大きく異なる

これらの特徴から，ドキュメントの内容を自然言語処理のみにより，設計情報として知識メディア化することは困難であると考えられる．したがって，非定型ドキュメントの知識メディア化においては，予め人の手作業により弱情報構造化を行い，その後計算機を使ってさらに構造化を行うという発想に優位性があると思われる．また一般に，種々の設計知識の融合，設計知識の再利用に伴う情報の

整理などの場面において，弱構造化された情報知識は計算機による扱いが容易であることから，この発想は効果的であると思われる．一方，手作業による弱構造化は，設計情報を正確に理解した設計者でないと有意義な構造化を行えない場合があり，設計者に負担を強いることがある．

テキストを基に設計情報の取得，設計知識の構築を行う試みに鷹合，Andy の研究がある [1]．鷹合は設計対象に関する概念体系 (オントロジー) を用いてテキストから設計知識を取り出すことを行い，設計支援を試みている [17]．また，Andy らは単語の cluster と belief network を組み合わせることにより設計知識の構築を試みている．本研究においてもオントロジーを利用して単語の関連を記述することにより，テキスト情報から問題解決過程の取得を行う際の効率化を図ることなどが考えられ，これらの成果を参考にしたい．

第 6 章

結論

本研究では設計情報を利用した設計作業の支援という観点からプロトコル解析文と議事録の構造化に関する考察を行った．また議事録の構造化を用いた設計情報可視化システムの試作を行った．

プロトコル解析文の構造化においては統計的手法を用いた方法に関して検討を行い，新しい手法を提案した．本手法はすべての文書に適用することができるわけではないが文書の論理構造の発見や文書の分類に利用できる可能性を示すことができた．また，議事録の弱情報構造に関して検討を加えた．設計情報可視化システムでは問題解決過程を不十分ながら抽出し提示することができた．

設計情報を利用した設計支援という視点で研究を進めたが，一般的な人間の作業支援はこれからますます重要な研究分野になってくるであろう．というのも人間と計算機の接点を作る最も総合した研究分野であると考えられるからである．今後も設計の視点から発展し，より一般的な人間の作業支援に関心を持っている．

謝辞

主指導教官である西田豊明教授には様々な面で御指導を頂きました．研究の基礎的背景である「知識情報」に関して適切な助言，指導を頂き「知識」そのものに対する理解を深めることができました．ここに深くお礼申し上げます．

植村俊亮教授には，本研究の副指導教官となって頂き，御指導を賜わりました．ここに深く感謝致します．

副指導教官の武田英明助教授には研究目的の設定から研究の内容，結果の評価に至るまで貴重な御指導を頂き深く感謝致します．知識を科学的に捉える見方，設計という視点から知識を考察する見方の厳密性かつ重要性を学ぶ事ができました．

また，設計情報の例として貴重な設計議事録の使用を許可したいただいた山武ハネウェル株式会社に御礼申し上げます．

参考文献

[1] Alice M. Agogino Andy Dong. Text analysis for constructing design representations. *Journal of Artificial Intelligence in Engineering (Previously published in Artificial Intelligence in Design '96*, 1:21–38, 1996.

[2] Tao Hu and Rolf Ingold. A mixed approach toward an efficient logical structure recognitio from document images. *Electronic Publishing*, 6(4):457–468, 1993.

[3] Rolf Ingold. *Text structure recognition in optical reading. In Jacques André, Richard Furuta, and Vincent Quint, editors, Structured Documents.* Cambrigde University Press, 1989.

[4] G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

[5] G. Salton and C.S. Yang. On the specificatio of term values in automatic idexing. *Journal of Documentation*, 29(4):351–372, 1973.

[6] Gerard Salton. *Automatic Text Processing.* Addison-Wessley Publishing Company, Inc., 1989.

[7] NAM P. SUH. *The Principles of Design.* Oxford University Press, Inc., 1990.

[8] Kristen summers. Toward a toxonomy of logical document structures. *DAG95:Electronic Publishing and the Information Superhighway*, http://www.cs.cornell.edu/Inof/People/summers/summers.html, 1995.

[9] Y. Nitta Y. Niwa. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. *Proc. COLING 94*, 1:304–309, 1994.

[10] 河合敦夫. 意味属性の学習結果にもとづく文書自動分類方式. In 情報処理学会論文誌 *Vol.33, No.9*, 1992.

[11] 吉川 弘之，冨山 哲男. インテリジェント *CAD*. 朝倉書店, 1991.

[12] 山本和英，増山繁，内藤昭三. 分類体系相互の関係を利用したテキストの自動分類. In 情報処理学会自然言語処理研究会 *Vol.106, No.2*, 1995.

[13] 松本 裕治，北内 啓，山下 達雄，今一修，今村 友明. 日本語形態素解析システム『茶筌』version 1.0 使用説明 書. Technical Report NAIST-IS-TR97007, Nara Institute of Science and Technology, 1997.

[14] 西田 豊明，武田 英明. 知識コミュニティプロジェクト (第 3 報). In *1996* 年度人工知能学会全国大会 *(*第 *10* 回*)*, 1996.

[15] 西田 豊明，武田 英明. 知識コミュニティプロジェクト (第 4 報). In *1997* 年度人工知能学会全国大会 *(*第 *11* 回*)*, 1997.

[16] 前田 晴美，糀谷 和人，西田 豊明. 連想構造を用いた情報収集・整理支援. In *1996* 年度人工知能学会全国大会 *(*第 *10* 回*)*, 1996.

[17] 鷹合 基行，足立 秀和，武田 英明，西田 豊明. オントロジーを用いた設計者の統合的支援環境. In *1997* 年度人工知能学会全国大会 *(*第 *11* 回*)*, 1997.

[18] 長尾真，水野幹男，池田浩之. 日本語文献における重要語の自動抽出. In 情報処理 *Vol.17, No.2*, 1976.

[19] 湯浅夏樹，上田徹，外川文雄. 大量文書データ中の単語間共起を利用した文書分類. In 情報処理学会論文誌 *Vol.36, No.8*, 1995.

[20] 武田英明. 設計過程表現論. 東京大学工学部博士論文, 1991.

[21] 崔 進，小松英二，安原 宏. Edr 電子化辞書を用いた単語類似度計算法. In 情報処理学会自然言語処理研究会報告，*NL93-1*, 1993.