2F3-4

設計議事録からの主題遷移構造の抽出と利用

Topic change extranction and utilization from design records

田中 克明　赤石 美奈　堀 浩一
Katsuaki Tanaka　Mina Akaishi　Koichi Hori

東京大学先端科学技術研究センター
Research Center for Advanced Science and Technology, The University of Tokyo

In artifact design process, designers make changes based on their intention. The result of changes affects to designers' intention and new changes are made. Therefore chronological change is important to handle knowledge of design process. In this paper, we discuss how to extract the topic change stcucture in chronological order from design records and how to reorganize the structure.

1. はじめに

人工物の設計，運用過程では，人間の意図とそれに基づく操作によって状態が変化し，さらにその結果を人間が得ることで，新たな意図が生じ，新たな操作が行われる．そのため，これらに関する知識を計算機の上で扱うためには，時間経過を考慮し，ある時点の情報とそれ以前の情報とのつながりをとらえることが重要である．

本稿では，東京大学大学院工学系研究科航空宇宙工学専攻中須賀研究室において行われている，超小型衛星 CubeSat "XI-IV" [*1] の設計を対象とし，設計過程において作成された議事録を情報源として，図 1(b) に示す主題遷移構造の抽出と，抽出した構造をユーザが持つ背景知識と組み合わせて利用する手法について，検討する．

時間経過と情報を扱う研究としては，設計支援システム操作からの設計プロセス抽出 [武内 07]，ニュース記事を対象とした文書からの主題とその時間変化の抽出 [Allan 02]，大まかな遷移の可視化 [Havre 00]，統計的な情報の通時的変化の抽出と可視化 [加藤 04] などが行われている．また，議事録を対象として [松村 03] などの研究も行われている．

これらに対し本研究では，特別な背景知識を用いない時間経過にともなう情報遷移構造の抽出と，抽出された構造と利用者の背景知識を組み合わせによる利用者の求める形への再構造化を行った．

2. 主題遷移構造の抽出

設計過程のある時点で設計者が議論している主題は，設計者が共有する問題構造において，その時点に焦点となった部分である．つまり，主題の遷移は設計者が持つ設計対象の問題構造変化を反映していると考えられる．よって，主題遷移の構造を取り出すことにより，設計過程でどのような意図が働き操作が行われたのか，を知ることができると考えられる．

2.1 抽出手順

主題遷移構造の抽出を，以下の手順により行う [Tanaka 06].

1. 文書の作成時間を元にした文書集合の定義
2. 文書集合からの主題の抽出

連絡先: 田中克明，東京大学先端科学技術研究センター知能工学研究室，〒 153-8904 東京都目黒区駒場 4-6-1，(03)5452-5289，(03)5452-5312，jsai2007@katsuaki-tanaka.net

*1 http://www.space.t.u-tokyo.ac.jp/cubesat/

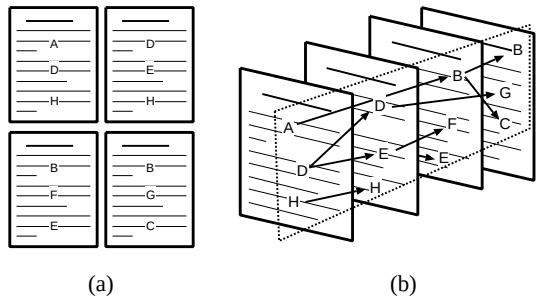

図 1: 主題抽出 (a) と主題構造変化抽出 (b)

3. 文書集合間の主題の関連度計算

文書群 D に対して，もっとも古い文書の作成時刻と最新の文書の作成時刻の間を N 等分し，文書集合の時間間隔 T を求め，N 個の文書集合 $D_1, D_2, \cdots, D_N$ ，$D_i \equiv \{\mathbf{d} \mid c(\mathbf{d}) \leq E(D) + i \cdot T\}$ を定義する．$c(\mathbf{d})$ は文書 $\mathbf{d}$ の作成時刻を表す．この結果，各文書集合は，$D_1 \subseteq D_2 \subseteq \cdots \subseteq D_N = D$ となる．

各文書は 1 つ以上の話題を含んでいることが，ほとんどである．例えば，CubeSat のある設計議事には，全体の進捗報告の他，各設計担当ごとに課題が記述されており，文書からこれらの主題を取り出す必要がある．本研究では，各文書を一定の長さに断片化し，これらをクラスタリングして類似したものを集めることにより，主題の取り出しを行う．

ここでは，文書を W bytes ごとに，$\frac{W}{3}$ の重なりを設けて断片化する．次に，各文書集合の各断片について形態素解析を行い，単語とその出現回数により単語ベクトルを作成し，GETA[高野 02] により M 個のクラスタにクラスタリングを行い，これらを主題とする．

D_{i+1} に属する文書断片のクラスタリングの終了後，D_i に存在せず D_{i+1} には存在する文書断片（すなわち新たな断片）を含まないクラスタ $C_{i+1,k}$ に属する文書断片に対し，D_{i+2} 以降のクラスタリングに用いる文書ベクトル作成の際，単語の重みを R 倍 $(R < 1)$ した．このように新たな言及がない主題の重みが減らすことにより，古い主題を風化させる．

文書集合数 N，断片化サイズ W，クラスタリング手法，クラスタ数 M，風化させる定数 R は，データの特性と後述する利用の目的に応じて設定を変更する．

クラスタリング後，隣接する文書集合間の主題，すなわちクラスタ同士の類似度を求めるため，$sim(C_{n,i}, C_{m,j})$ を以下の

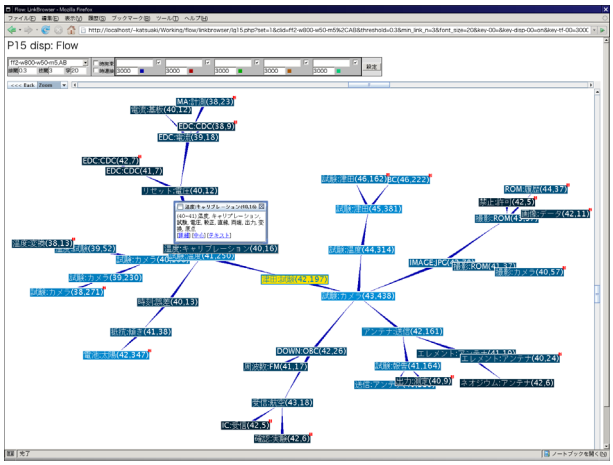

図 2: 主題遷移構造表示例

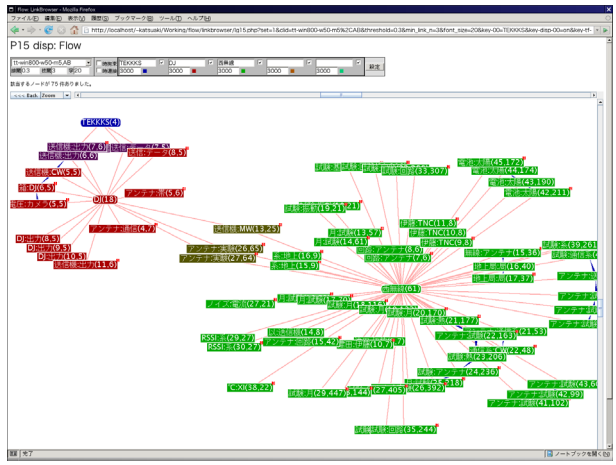

図 3: ユーザ視点に基づく再構造化例（「TEKKKS」「DJ」「西無線」を指定）

ように定義した．

$$sim(C_{n,i}, C_{m,j}) = \frac{|C_{n,i} \cap C_{m,j}|}{|C_{n,i}|}$$

n，m は，$1 \leq n \leq m \leq N$ となる文書集合番号，$C_{n,i}$ は文書集合 n の i 番目のクラスタを示す．

2.2 主題構造変化の表示

類似度関数 $sim(C_{n,i}, C_{m,j})$ により，隣接した文書集合のクラスタ間の関係をグラフ化し表示する．表示には TouchGraph LinkBrowser*2 を用いた（図 2）．類似度が 0.3 以上のクラスタ間にリンクを持たせ，クラスタ $C_{n,i}$，$C_{n+1,j}$ 間の距離を類似度の逆数と比例させて定義した．各クラスタをグラフ上のひとつのノードとし，出現頻度の高い上位 2 語，文書集合番号，およびクラスタに含まれる断片の数をラベルとした．図 2 の矢印の向きが，時間の経過方向を示す．

3. 主題遷移構造の利用

以上の手法により，議事録から主題遷移構造を特別な背景知識なしに抽出する．次に，抽出された構造を，ユーザの背景知識を組み合わせて再構造する手法について述べる．

3.1 ユーザ視点に基づく再構造化

得られた主題遷移構造からユーザが指定した語に関連する主題のみを選択することにより，全体の俯瞰ではなく，ユーザの視点に基づいた主題遷移構造を表示する [Tanaka 06]．ユーザが指定した語を，主題を構成する単語ベクトル中から検索し，該当する主題のみを選択表示する．CubeSat 設計中に議論された 3 つの無線機，DJ シリーズ，TEKK-KS，西無線無線機を指定した例を図 3 に示す．3 つの無線機が検討された順にとりだされ，それぞれをつなぐ主題には無線機を変更した理由が含まれていた．

3.2 作業過程の抽出

主題遷移構造からユーザが指示した対象へどのような作業が行われたかの一連の作業過程を抽出する [田中 07]．前項と同様にユーザが指定した語を含む主題を取り出し，そこからサ変接続可能名詞を新しいもの順に選択し，ノードのラベルとし，

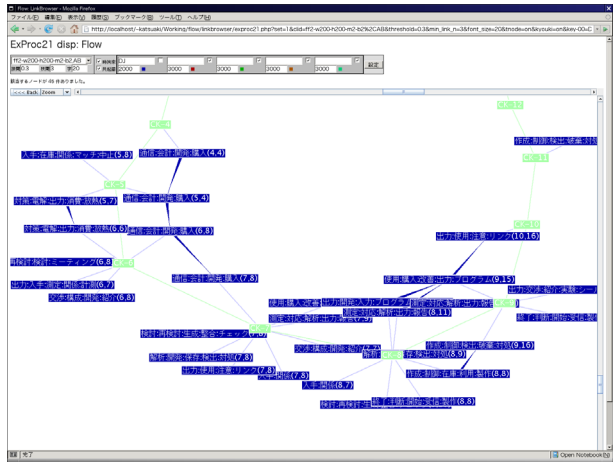

図 4: 作業過程抽出例（「DJ」を指定）

時刻を表すノードと結びつけてグラフ化する．無線機 DJ シリーズを指定した例を図 4 に示す．記述内容が順次取り出されるだけでなく，宇宙対策と購入手続きが並行して検討されていたことなどを抽出できた．

4. まとめ

本稿では，設計議事録から背景知識を用いず主題遷移構造を抽出する手法，抽出した構造をユーザが持つ背景知識と組み合わせた利用手法について検討した．これにより，時間の経過にともなって設計がどのように進んだかを知ることができ，設計過程における設計者の意図を知る手がかりを得ることができた．

現在，主題は単純なクラスタリングによって構成し単一の粒度を持つものとしているが，今後は，クラスタリング手法が持つ階層性と組み合わせることにより，ユーザの視点に応じて粒度を切り替える手法について検討を進める予定である．

*2 http://www.touchgraph.com/

参考文献

[Allan 02] Allan, J.: *Topic Detection and Tracking: Event-based Information Organization*, Kluwer Academic Publishers (2002)

[Havre 00] Havre, S., Hetzler, B., and Nowell, L.: ThemeRiver: Visualizing Theme Changes over Time, in *Proc. of IEEE Symposium on Information Visualization* (2000)

[Tanaka 06] Tanaka, K., Akaishi, M., and Hori, K.: Topic Change Extraction and Reorganization from Problem-solving Records, in *Proceedings of International Conference on Software Knowledge Information Management and Applications*, pp. 153–158 (2006)

[加藤 04] 加藤 恒昭, 松下 光範, 平尾 努：動向情報の要約と可視化に関するワークショップの提案, 情報処理学会自然言語処理研究会 2004-NL-164 (15), pp. 89–94 (2004)

[高野 02] 高野 明彦, 丹羽 芳樹, 西岡 真吾, 岩山 真, 今一 修, 久光 徹：汎用連想計算エンジン GETA, *http://geta.ex.nii.ac.jp/* (2002)

[松村 03] 松村 真宏, 加藤 優, 大澤 幸生, 石塚 満:議論構造の可視化による論点の発見と理解, 日本ファジィ学会誌, Vol. 15, pp. 554–564 (2003)

[田中 07] 田中 克明, 赤石 美奈, 堀 浩一：設計議事録からの設計プロセス抽出の試み, 電子情報通信学会技術研究報告 知能ソフトウェア工学, 第 106 巻, pp. 43–48 (2007)

[武内 07] 武内 雅宇, 小路 悠介, 來村 徳信, 林 雄介, 池田 満, 溝口 理一郎：知識成長過程を指向した設計意図知識管理システムの構築, 人工知能学会論文誌, Vol. 22, pp. 263–275 (2007)