1C4-03

設計議事録からの主題構造変化の抽出

Topic change extranction 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, University of Tokyo

In this paper, we discuss the function of an information sharing system from the chronological viewpoint. In the artifact design process, the chronological change of designers' focal points is important information because it reflects their view of the problem structure. Designers' focal points are embedded in documents created during the design process. By detecting and tracking topics in the documents, we will be able to trace their focal points. We propose a user interface system to support users as they search for topic changes from design records. We also propose a design process management system based on topic changes.

1. はじめに

システムが大きく複雑になるにつれ，多くの人間がひとつのシステムの設計と運用・管理にかかわるようになり，人間にとって，システム全体を把握することが難しくなる．このような状況に対応するため，多くの情報管理，獲得，構造化技術が研究され [Baeza-Yates 99]，情報の共有が試みられている．

多くの情報共有技術では，情報をひとかたまりのデータとして扱っており，時間経過に伴う変化は，考慮されていない（図 1(a)）．しかしながら，実際の設計，運用過程では，システムに対する人間の意図，操作によってシステムの状態が変化し，さらにその結果を人間が得ることで，新たな意図が生じたり，新たな操作が行われたりする．すなわち，時間経過を考慮し，各時点の情報を，それ以前の情報とのつながりを踏まえた形で，獲得，共有することが必要である．

そこで，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] の設計過程を対象とし，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) に示すような主題構造変化の，抽出を行うことを目的として，本研究を行った．

2. 主題構造変化の抽出

設計過程にて設計者の議論の対象となる主題とは，設計過程のある時点において，設計者が共有する問題構造の中で，設計者が焦点をあてた部分である．そのため，主題構造の変化は，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．すなわち，主題構造の変化を取り出すことにより，設計過程でどのような意図が働いたのか，を知ることができると考えられる．

2.1 抽出手順

主題変化の抽出を，以下の手順により行う [Tanaka 04].

1. 文書の作成時間を元にした文書集合の定義
2. 各文書の断片化，および断片のクラスタリング処理
3. 各文書集合間のクラスタの関連度計算

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，東京都目黒区駒場 4-6-1，(03)5452-5289，(03)5452-5312，jsai05@katsuaki-tanaka.net

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

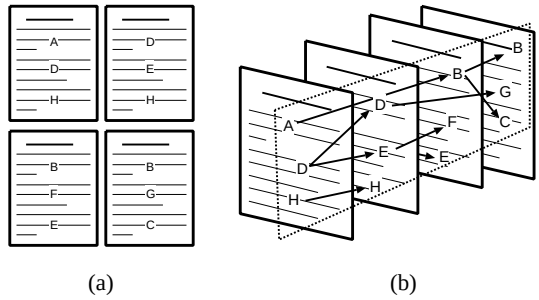

図 1: 主題抽出 (a) と主題構造変化抽出 (b)

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 を定義する．T にもとづき，N 個の文書集合 $D_1, D_2, \cdots, D_N$ を，$D_i \equiv \{\mathbf{d} \mid c(\mathbf{d}) \leq E(D) + i \cdot T\}$ と定義する．$c(\mathbf{d})$ 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．なお，$N = 50$ とした．

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 1 つの設計議事録文書中には，全体の進捗報告の他，各設計担当ごとに課題があげられており，それぞれ別の主題として取り扱う必要がある．そのため，文書を主題単位に分割する必要がある．本論文では，各文書を一定の長さで区切ることによって断片化を行い，これら断片をクラスタリングすることで主題の取り出しを行う．

はじめに，各文書の形態素解析を，茶筌 [松本 00] を用いて行った．議事録中では専門用語が数多く使われているため，連続して出現する名詞をひとつの名詞として扱うように，形態素解析を行った．形態素解析の結果を踏まえ，200 文字ごとに文書を断片化した．境界が単語の途中に存在する場合には，断片が単語全てを含むように，断片の長さを延ばした．また，各断片間には 67 文字分の重なりを設けた．

次に，各文書集合の各断片について，形態素解析結果を踏まえ，単語とその出現回数により単語ベクトルを作成した．ここでは，茶筌によって，名詞と品詞分類された単語のみを用いた．その後，GETA[高野 02] を用いて，Ward 法によりクラスタリング処理を行い，各文書集合中の各断片を 50 個のクラスタに分割した．

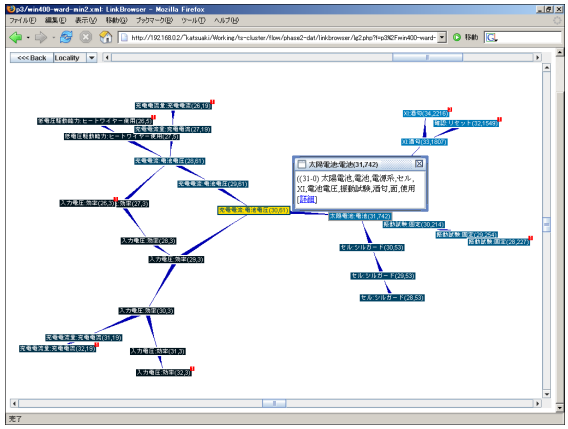

図 2: 主題構造変化表示例

クラスタ数の上限を設定することにより，新しいクラスタが既存の要素から離れた要素により形成されると，これまで存在していたクラスタは，その数を減らすために，再構成される必要が生じる．その結果，類似したクラスタは，ひとつのクラスタにまとめあげられる．このように，時間経過とともに，似た要素が同じクラスタを，新たな要素が新しいクラスタ形成することを目的としクラスタリングを行わせる．

クラスタリング後，隣接する文書集合間の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であり，$C_{n,i}$ は文書集合 n の i 番目のクラスタである．

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可視化する．表示には TouchGraph LinkBrowser*2 を用いた（図 2）．この際，類似度が 0.3 以上のクラスタ間にリンクを持たせることとし，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，クラスタの特徴語 1 つと，文書集合番号，およびクラスタに含まれる断片の数を，ノードのラベルとして表示させた．特徴語は [Yang 00] にて定義された TFIDF 値に基づき，値が大きい単語を選択した．図 2 の矢印の向きが，時間の経過方向を示す．ノードを選択すると，クラスタの特徴語 10 個とクラスタ詳細表示画面へのリンクが，ポップアップウィンドウとして表示される．

3. 主題構造変化の蓄積と再構造化

これまで述べた手法により，設計議事録に記録された主題構造の変化の概要が，ユーザに提示される．このような各段階の変化の積み重ねにより，設計は順次進んでいく [Tanaka 05]．そこで，ユーザが選択した各段階の変化構造を蓄積・管理することにより，問題全体の管理を行うシステムを，あわせて構築した．1 段階分の変化を管理する画面例を図 3 に示す．各段階をユーザが組み合わせることにより，問題全体の管理を行う．このシステムでは，設計中に交換される主題にともなう，詳細な情報の管理も行う．

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

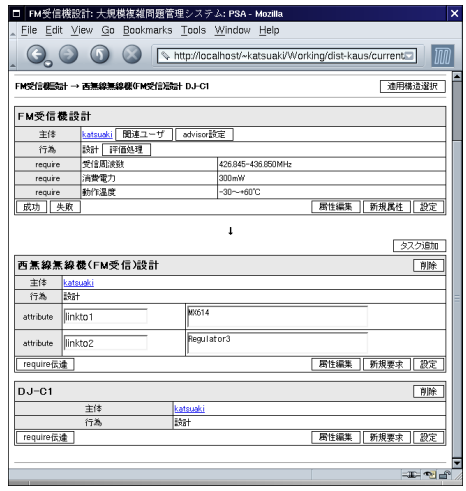

図 3: 問題構造管理システム

4. まとめ

本稿では，設計過程における設計者にとっての問題構造の変化を把握することを目的に，人工衛星設計過程で作成される議事録より，時間経過に沿った主題変化の抽出を試みた．また，主題変化を積み上げることで，設計過程の管理を行うシステムを構築した．

抽出過程における，クラスタリング手法，クラスタ間の関連度計算手法は一般的な手法を用いているので，これらを主題構造変化抽出により適した手法とすることを考えている．また，表示された結果からの主題の読み取りと再構造化処理は，全てユーザに任せているが，これらを計算機が支援する仕組みを組み込むことも，検討している．

参考文献

[Baeza-Yates 99] Baeza-Yates, R. and Ribeiro-Neto, B.: *Modern Infromation Retrieval*, Addison Wesley (1999)

[Tanaka 04] Tanaka, K. and Takasu, A.: Topic Change Extraction from Problem Solving Records, in *Proc. of the 8th World Multi-Conference on Systemics, Cybernetics and Informatics* (2004)

[Tanaka 05] Tanaka, K., Akaishi, M., and Hori, K.: Semantic Structure Transition with Elapsed Time, in *Proc. of the Semantic Computing Initiative* (2005)

[Yang 00] Yang, Y., Ault, T., Pierce, T., and Lattimer, C. W.: Improving Text Categorization Methods for Event Tracking, in *Proc. of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 65–72 (2000)

[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)

[松本 00] 松本 裕治, 北内 啓, 山下 達雄, 平野 善隆, 松田 寛, 高岡 一馬, 浅原 正幸：日本語形態素解析システム『茶筌』version 2.2.1 使用説明書 (2000)