話題構造および文意味構造に基づく文書可視提示方式の提案

4R－10

稲垣 博人　早川 和宏　田中 一男

NTTヒューマンインタフェース研究所

1 はじめに

近年、webの流行や種々の通信手段の発展によりデジタルコンテンツを大量に流通することができるようになってきた．そのため，大量の情報の中から，いかに効率的に，かつ効果的に必要な情報をアクセスできるようにする枠組が必要となってきた．その中で，流通されている情報の中から重要な情報をピックアップして可視的に提示する文書可視化提示による支援技術を我々は研究してきた[1),2)]．その中で，文書の意味構造として，事象構造や話題構造などの文書の意味構造を提案してきた．本稿では，文書の意味構造の1つである話題構造を中心に，文書中の重要なフレーズを抽出し，コントラストをつけて表示するインタフェースを提案する．特に，クライアント側での使用を考慮して，話題構造や文書の意味構造を短時間で抽出する快速覧エンジンを作成し，その快速覧エンジンを利用して，カラーリングによりコントラスト付けする文書提示方式を提案する．

2 話題構造と文意味構造の抽出

文書中の内容を一覧するために，文書の論理的構造や意味的構造の把握と，それぞれの文書の構造内における意味の抽出が重要となる．

文書の構造については，一般的にいわれている，章，節，段落等の論理的文書構造がある．これは，従来から文書構造を明確にするために使われているテクニックである．文書構造的には，章・節・段落が最も上位の構造として用いられている．必要に応じて，箇条書や表組みによる構造化も行なわれる．さらに，センタリング等により，明示的にある意味付けを行なった段落を構成する場合がある（ここでは，そのような段落を"意味段落"と呼ぶ）．

ここでは，表層的手法により文書の論理的構造を抽出すると共に，意味段落等の意味的に重要な構造を抽出する．さらに，抽出した文書中の各ブロックの意味内容を抽出する．ブロックの意味内容としては，竹下ら[2)]が定義した話題を利用した．つまり，個々のブロックにおける話題が，文書内のブロックの意味内容を表す．

クライアント向けに実時間程度の応答性を要求する場合，複雑な意味解析を行なうことはできない．そのため，形態素解析程度の処理時間で処理する必要がある．そこで，話題抽出処理としては，竹下ら[2)]のテキスト向け話題抽出システム"テキスト速覧"を改良した，形態素解析および文書構造解析によるテキスト向け話題抽出システム"快速覧"を使用した．

2.1 文意味構造の抽出

文書の論理的構造を抽出するために，対象テキストを行単位で読み込み，章，節，段落，箇条書等を抽出する．章，節，段落に関しては，章，節，段落番号を文意味構造解析用辞書に登録し，文意味構造解析時に該辞書を用いて解析を行なう．箇条書も同様に，箇条書番号や箇条書記号を文意味構造解析用辞書に登録する．

意味段落は，意味段落を構成する行頭の表現を対象とするドキュメントごとに抽出し，その表現と一致する場合，後続する段落を意味段落として認識する．そのため，対象とするドキュメントに対して，意味段落の行頭表現を抽出し，文意味構造解析用辞書に登録を行なった．表1に医療関係の文書から抽出した意味段落の行頭表現例を示す．

2.2 快速覧による話題構造の抽出

テキスト向け話題抽出システム"快速覧"とは，対象となる文書の形態素解析情報，文書構造情報を用いて，文書中の話題を抽出するシステムである．実時間程度の処理速度を実現するために，係り受け解析や意味解析などは行なわず，形態素解析のみを実行する．形態素解析としては，InfoBee形態素解析エンジン[3)]を用い，Infobee形態素解析エンジンに組み込んだ形とした．

快速覧は，話題構造を高精度で抽出することよりも，文書可視提示等の文書中の重要な表現である話題を抽出することが主であるため，以下の制限を設けている．

- 形態素解析情報（単語，品詞，活用）を利用する．
- 話題階層は2階層とする．

表1: 意味段落の行頭表現例

トピックス，トピック，話題
質問，回答，参考，Q&A
メモ，ノート
相互作用，副作用，メカニズム

A document visualization technique utilizing document structure and topic/semantic-structure of sentences .
Hirohito INAGAKI, Kazuhiro HAYAKAWA, and Kazuo TANAKA.
NTT Human Interface Laboratories

これらの制限により，係り受け解析や，話題構造の解析に時間を必要とせず，短時間に話題構造の抽出が可能となる．

話題構造には，大局話題と呼ぶ文書中で表現される大きな話題と，その話題を転換したり，広げたりする意味での局所的に発生する局所話題とがある．

大局話題を導入する文は，「第一に」「最初に」などの話題導入手掛かり句を持つ文である．次に，その文の中で「に関する」「について」などの明示マーカーと呼ばれる，話題を導入するマーカーが後置される名詞句や，「は」「が」などの助詞により弱く明示される弱明示マーカーと呼ばれる語句に後置される名詞句を顕著名詞句として抽出する．その顕著名詞句の中で，話題の重要度や，品詞情報により，最も重要な顕著名詞句を話題語として選出する．

局所話題は，大局話題の導入部がない文で，かつ上記のような顕著名詞句のうち最も重要な顕著名詞句を，話題語として選出する．

2.3 話題構造と文意味構造の抽出精度評価

評価対象とする文書は，医療分野で閲覧されている各種の医薬品関係情報や，各省庁における議事録などを対象とした．

話題構造および文意味構造における正解は，被験者に対象となる文書を読ませ，その文書の中で話題，またはキーとなる語を抽出させた．人間の正解作成作業では，話題の階層については抽出しなかったので，話題の階層については評価しなかった．表2に，抽出した文意味構造および話題構造に対する適合率と再現率をもとめた．

表をみればわかるように，平均で再現率86%を得た．また，快速覧に基づく話題構造を中心として重要語を抽出しているため，快速覧の得意な議事録等では，再現率94%と非常に高い精度で話題構造が抽出できた．

3 文書可視提示方式

通常行なわれる文書の閲覧・回覧方式では，webサービスによる情報伝言板や，電子メールによる通知などが考えられる．これら電子メールやwebによる文書の回覧を考えると，原文の内容を変更することなくコントラストを表現できる，HTMLタグによるコンテンツのカラーリングが適していると考えられる．

表2: 話題構造と文意味構造の抽出精度

文書分類	再現率	適合率
医薬品情報	79%	72%
議事録	94%	61%
平均	86%	66%

HTMLタグによるコンテンツカラーリングは，多くの人がみるwebブラウザにおける標準的な手法として用いられており，ユーザは非常に馴染み易い．さらに，カラーリングなどの単純なHTMLタグを解釈できるHTMLメーラーもあり，文書の閲覧・回覧のインタフェースとしては，HTMLタグによるカラーリングは適していると考えられる．そこで，話題構造や文意味構造により抽出されたフレーズをカラーリングにより提示する方式を検討した．

カラーリングについては，赤色，黄色，緑色などの普遍的意味づけが行なわれている色がある．赤色は特に注意を引いたり，停止するなど，webのデザインでもあまり使ってはいけない色となっている．一方，標準的なwebの世界では，青は，ハイパーリンクを示す色となっており，webを使うユーザの中では，比較的認知度が高い色となっている．そのため，話題等の文書構造を示すには最適と考えられる．

一方，段落，箇条書等の文意味構造における色付けでは，普遍的な色というよりも，各段落の機能をユーザに明示する色付けが必要となる．そのため，段落，箇条書等は，他の文と区別できるように濃い色を用いた．また，意味段落の行頭表現は，その段落の意味付けを行なっているだけであるから，輝度の高い色を用いることとした．

4 まとめ

本稿では，文書の速読や，短時間に情報を閲覧するためのインタフェースとして，文書中の話題構造や，文意味構造に基づき，文書の内容を表す重要なフレーズをカラーリングすることにより，文書を可視的に提示する手法について提案を行なった．本提示方式をInfoBeeエンジンに組み込み評価を行なった．医療関係の文書に対して再現率86%，適合率66%の非常に高い精度で自動的に重要なフレーズを抽出できることがわかった．

今後は，本提案方式を用い，種々の文書閲覧系に組み込み文書可視提示方式の評価を行なう．

参考文献

1) Hirohito Inagaki and Tohru Nakagawa. An abstraction method using a semantic engine based on language information structure. *Coling-92*, 1992.

2) 竹下敦, 井上孝史, 田中一男. テキストの概要把握支援のための話題構造抽出. 情報処理, 1996.

3) 井上孝史, 大久保雅且, 杉崎正之. Infobeeテキスト検索情報検索技術. NTT R&Dジャーナル,Vol.46,No.10,pp.1103-1108, 1997.