2E1-05

ドキュメントスキーミング支援環境の構築とその評価

Development of Skimming Support Enviroment

羽山 徹彩[*1] 金井 貴[*1] 國藤 進[*1]
Tessai Hayama Takashi Kanai Susumu Kunifuji

[*1] 北陸先端科学技術大学院大学 知識科学研究科
Graduate School of Knowlge Science, Japan Advanced Institute of Science and Technology

We developed a skimming support system to support skimming documents from computer screen and a recommend system which generates personalized summaries. Skimming support system has the interface combined the fisheye effect and the overview+detail effect. Focus points of the fisheye in the skimming support system are the sentences selected by the original sentence extraction algorithm and the overview is used the automatic generated the table of contens. On each experiment, our system performs efficiently as same as reading paper; our sentence extraction algorithm by using the value of standard distribution is higher value of F-measure than the sentence extraction algorithm by using TF or Japanese lexical chaining method, our skimming supported system is the same effective to skim documents as reading from the paper, and summaries which our recommend system generates reflect the user's preference.

1. はじめに

教育や研究などの知的活動において，文書を読み，適切に理解をすることは重要である．近年の情報化社会にともない，ネットワーク上から電子化されたドキュメントの入手が容易となった．そのため，コンピュータ画面上からドキュメントを読むことが一般的であると考えるが，多くの人は印刷し紙面上から文書を読むことをおこなっている．この理由として，現代の人間の読み方であるスキーミングがコンピュータ画面において不得意であるためと考える [5][12].

本研究の目的は，コンピュータ画面からのドキュメントスキーミング支援環境を構築することである．そのために，スキーミング支援システムと要約提供システムを構築した．スキーミング支援システムはコンピュータ画面からのドキュメントスキーミングを支援するインタフェースを持ち，要約提供システムはユーザの嗜好を考慮した要約を提供するシステムである．本研究でのドキュメントスキーミングの定義は短時間で正確に文書の意図を理解する行為とし，対象ドキュメントを日本語の論文とした．

2. システムの実装

ドキュメントスキーミング環境はクライアント/サーバーシステムであり，クライアント側ではスキーミング支援システム (2.1 節参照)，サーバー側では要約提供システム (2.2 節参照) が主に処理をする．

2.1 スキーミング支援システム

紙面に対しコンピュータ画面から読み難さの原因は，一般的に物理的制約と物理的特性の問題であるといわれている [11]. 物理的制約とは紙にあったフォーマットのドキュメントにとってコンピュータ画面への表示が難しく，論文などの長いドキュメントに対して一覧性を確保できないという問題である．そのため，ディスプレイ上では一般的にスクロールバーを用いた表示方法をおこなっているが，ページングに比べ内容の理解が難しい [15]．この理由として，読み返しが困難であるためと考える．物理的特性とは現在の読み位置を認知的に知ることが可能であることであり，コンピュータ画面からの読みでは読み位置を把握し難いという問題がある．本スキーミング支援システムでは，以上の２つの問題に対し以下のアプローチをおこなう．

- 物理的制約
  - ドキュメントのセグメント単位表示
  - 重要な文に対して Fisheye 効果の適用
  - インタラクティブに文単位での Fisheye 効果の切替え
- 物理的特性
  - 目次型インタフェースを用いた Overview+Detail 効果の適用

これらアプローチをもとに構築したスキーミング支援システムのインタフェースを，図 1 に示す．

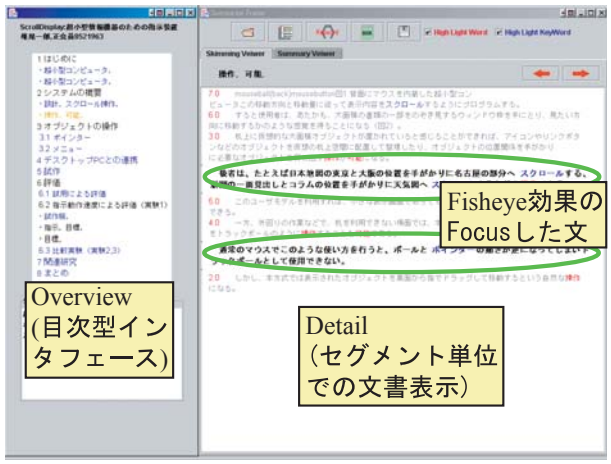

図 1: スキーミング支援システムのインタフェース

ドキュメントのセグメント単位表示はスクロール表示を減らすことを目的とし，TextTiling アルゴリズム [7] に基づく手法よってセグメント分割をおこなった．重要な文に対しての Fisheye 効果の適用は，重要な文に対するアクセス速度向上によりドキュメントの内容を短時間で理解する支援を目的

連絡先: 羽山徹彩，北陸先端科学技術大学院大学知識科学研究科，〒 923-1292 石川県能美郡辰口町旭台 1 ー 1，0761(51)1699(内線 1850)，0761(51)1775，t-hayama@jaist.ac.jp

とした．また，ユーザが文単位での Fisheye 効果の Focus の切替えを可能にすることでリーディングポイントの明示化をおこない，読み返しを支援する．Fisheye 効果の Focus は，提案する重要文抽出手法 (2.1.2 節参照) により抽出した文とした．Overview+Detail 効果はドキュメント全体に対する現在の読み位置の把握を支援するために使用し，Overview には実世界において全体の内容を知るために使用している目次を適用した．

2.1.1 セグメンテーション手法

本研究のセグメンテーションは，TextTiling アルゴリズム [7] に基づく McDonald の手法 [4] を用いた．McDonald の手法は，以下の手順でおこなう．

1 トークン分割

ドキュメントを形態素解析*1をおこない，ストップワードの除去と形容詞、形容動詞，動詞を原形へ変換する．

2 結束スコアの決定

2.1 トークンシーケンスを L づつ区切り、ブロックを作る．

2.2 隣接するブロックの類似度を計算する．

前にあるブロックはそのブロックより前の k トークンシーケンスを加え，後ろにあるブロックはそのブロックより後ろの k トークンシーケンスを加え，Jaccard 係数 (式（1）) によりブロック間の類似度を計算する．

$$S_{i,j} = \frac{\sum_{k=1}^{L}(w_{ik}, w_{jk})}{\sum_{k=1}^{L} w_{ik}^2 + \sum_{k=1}^{L} w_{jk}^2 - \sum_{k=1}^{L} w_{ik} w_{jk}} \quad (1)$$

3 境界の決定

求められた結束スコアーをトークンシーケンスナンバー，類似度においてプロットし，値の補完*2をおこない滑らかにする．その結果，極小解となった点を境界とする．

本研究でセグメント分割手法に用いたパラメータは，Hearst と同様に L の値を 20，k の値を 10 とし，ストップワードは助詞，助動詞と Hearst が用いたワードを日本語化したものを用いた．境界の判定には極小解に近いパラグラフの境目を境界とした．

2.1.2 重要文抽出手法

本研究の重要文は，話題の多様性および話題と文章全体の意味関係を考慮する必要があると考えた．そのため,『全体に分散する単語は，全体を表現するため重要である．特定の部分にだけ集中する単語は，その部分を表現するため重要である．それら２つのタイプの単語が含まれる文は，全体と部分を結びつけるため重要である.』という仮説をたてた．この仮説をもとに，本重要文抽出手法は，全体を章（あるいは節）とし部分をセグメントとして分散値をもとにおこない，それに加え，手掛かり語を含む文と項目の最初の一文が重要であるというヒューリスティックなルールを適用した．その計算手順を以下に示す．

1. 手掛かり語「従って，すなわち，提案する」を含む文と項目の最初の一文を重要文として抽出する

2. 章（あるいは節）とそれに含まれるセグメントとの分散値によって，単語の重みを計算する

$$\text{単語 } x \text{ の重み} = \left| \sum_{k=0}^{N} \frac{(x_k - \bar{x})^2}{N} - \bar{\sigma^2} \right| \quad (2)$$

k はセグメントナンバー，x_k はセグメントナンバー k の単語の頻度，$\bar{x}$ は章（或いは節）内の単語の総頻度，N はセグメント数，$\bar{\sigma^2}$ は全ての単語の分散値の平均である．この式から分散する単語と集中する単語は高い重みとなる．

3. 各文の重要度を文に含まれる単語の重みの平均値として計算する．

4. 重要度が上位の文を全体の 20 %となるまで重要文として抽出する．

以上から抽出された文は，Fisheye 効果の Focus として表示する．

2.1.3 目次自動生成手法

目次型インタフェースは，ドキュメントに含まれる論文中の章のタイトル，節のタイトル，およびセグメントに含まれるキーワードを順に抜き出し表示した．セグメントに含まれるキーワードは，出現キーワードの頻度の順番において上位にきた名詞を選択し，最も頻度の高い名詞を含む一文中のキーワードを表示した．これは，人に印象を与えることにおいて語の並びが重要であるとの考えに基づく．

2.2 要約提供システム

各ユーザごとの関心や知識によって，文章を読む観点が異なる．要約生成においても，ユーザの観点によって正解となる要約が異なる [6]．そのため，我々は，ユーザの嗜好を考慮した要約提供システムを構築する．

要約提供システムのアプローチは，各ユーザのスキーミング支援システムを使用した後の履歴をもとに協調フィルタリングをおこなう．スキーミング支援システムの履歴は，ユーザ名，論文名，マウスによって選択された文番号である．システムの要約提供方法は，以下の計算手順でおこなう．

1. システム使用後にユーザ名，論文名，マウスによって選択された文番号をサーバー側へ送信しデータベースへ登録する．

2. ユーザがデータベースに登録されている論文リストから読みたい論文を選択する．

3. サーバーにおいて選択した論文を過去に読んだユーザから選択したユーザと類似するユーザを Jaccard 係数により求める．

4. 最も類似性の高いユーザの要約を提供する．

また，論文要約のビューワーはスキーミング支援システムのタブボタンを切り替えることで表示する．

*1 Naist 松本研究室の茶筅を使用した
*2 Newton 補完法を用いた

3. 評価実験

3.1 重要文抽出手法の評価

本実験の目的は，提案した重要文抽出手法の精度を評価することである．評価方法は，人手で作成した正解データをもとに論文 3 本に対し本手法と簡易要約器 Web Posum[*3] と比較した．正解データは，評価者 7 人が個別に抽出した重要文に対し過半数 (4 人以上) の人が選択した文とした．評価尺度は，テキスト要約の評価で用いられる式 (3)，式 (4)，式 (5) の再現率，精度，F-measure を用いた．また，Posum の適用手法は，出現頻度を用いた手法と分類語彙表による単語間のつながりを利用した手法の 2 つに対しておこない，要約率は章または節ごとに 20 %とした．評価結果を表 1 に示す．

$$\text{再現率} = \frac{\text{適合した文数}}{\text{正解データの文数}} \tag{3}$$

$$\text{精度} = \frac{\text{適合した文数}}{\text{正解データに適合した文数}} \tag{4}$$

$$F - measure = \frac{\text{再現率}}{\text{精度}} \tag{5}$$

表 1: 本提案手法と簡易要約器 Posum の再現率，精度，F-measure の値

論文	システム	再現率	精度	F-measure
A	Posum（頻度）	0.25	0.18	0.21
	Posum（分類語彙表）	0.24	0.24	0.27
	本提案手法	0.52	0.45	0.48
B	Posum（頻度）	0.44	0.39	0.41
	Posum（分類語彙表）	0.44	0.39	0.41
	本提案手法	0.45	0.50	0.47
C	Posum（頻度）	0.15	0.27	0.20
	Posum（分類語彙表）	0.14	0.24	0.18
	本提案手法	0.46	0.56	0.48

表 1 の結果から，本提案手法は F-measure が平均 0.48 と，比較した他の手法より有効かつ安定した値が得た，以上から，分散値を用いた本提案手法は有効な手法であると考える．

3.2 スキーミング支援システムの評価

本実験の目的は，スキーミング支援システムがコンピュータ画面からのドキュメントスキーミングを支援しているかの評価をおこなうことである．実験は，普通に読んだ場合とスキーミングをおこなう場合に分けて，被験者 9 人づつに対し実施した．実験環境は，21 インチ CRT ディスプレイを使用した．実験方法は，論文 3 本に対し異なる読み方で読み，各論文に対しての 7 問の正誤問題を解くことをおこなった．読み方は紙，Thumnails 付き Acrobat Reader，およびスキーミング支援システムであり，論文との対応は各被験者に対しランダムに決めた．ドキュメントを対象とした実験のため，論文には，図，表，概要を取り除いたものを用いた．普通に読んだ場合とスキーミングをおこなった場合の異なる点は，普通に読んだ場合に被験者がかかった平均時間の 3 分の 2 をスキーミングをおこなった場合の制限時間としたことである．評価方法は，各場合において紙面，サムネール付き Acrobat Reader，本スキーミング支援システムの 3 つの方法をの評価指数によって比較した．評価指数は，論文を読む時間と理解度を考慮した Jackson の効率的読解度 [9] をもとにして，各論文の文書と問題の難易度を考慮し各論文ごとに時間の正規化をおこなった値に対し問題の正解数で割った値を用いた．実験結果を表 2 に示す．

表 2 が示すように普通に読んだ場合では本スキーミング支援システムが最も有効な読み方を提供していが，分散分析の結果では 0.034($< F_{0.10}(2,24)$) と 10%優位水準において優位差がなかった．スキーミングをおこなった場合では，本スキーミング支援システムが最も有効な読み方を提供しており，分散分析において結果が 2.85($> F_{0.10}(2,24)$) と 10%優位水準で優位差を確認した．

表 2: 普通に読んだ場合とスキーミングをおこなった場合の効率的読解度の結果

実験方法	読み方	効率的読解度
普通に読んだ場合	紙	6.71
	Acrobat Reader	6.78
	スキーミング支援システム	6.51
スキーミングをおこなった場合	紙	6.09
	Acrobat Reader	8.00
	スキーミング支援システム	5.91

コンピュータ画面と紙面からのスキーミングを効率的読解度により比較する研究として Paul の実験 [12] がある．本実験の普通に読んだ場合において，3 つの読み方の効率的読解度の有意差はなかった．これは，Paul の実験と同様であり本実験環境がある程度，正当性がいえる．スキーミングの場合においても Paul の実験同様，有意差を確認した．Paul の実験ではコンピュータ画面より紙面の方が効率的読解度によって 41 %有意であった結果に比べ，本システムではサムネール付き Acrobat Reader より 51 %有意，紙より 3 %有意であった．以上より，本スキーミング支援システムがコンピュータ画面からのスキーミングに対する困難さをある程度解決できたと考える．

3.3 要約提供システムの評価

本評価実験の目的は，要約提供システムの提供する要約が各ユーザの嗜好を反映しているかを検証することである．そのために，本実験では，スキーミング支援システムの履歴をもとに要約提供システムが算出したユーザ間の類似度とユーザが主観でおこなった要約の順位付けとの相関を調べる．実験の被験者は 4 人 (A から D) で，論文 1 本に対しおこなった．ユーザが主観でおこなった要約の順位付けは，被験者 A から D がスキーミング支援システムの使用履歴を論文の要約として提供し，最も好ましい要約を 4 として，順位付けをおこなってもらった．実験結果を表 3 に示す．

表 3: システムが算出したユーザ間の類似度とユーザが主観でおこなった要約の順位付け

	A(順位)	B(順位)	C(順位)	D(順位)
A	1.00(4)	0.68(3)	0.49(1)	0.65(3)
B	0.68(3)	1.00(4)	0.57(2)	0.69(2)
C	0.49(1)	0.57(1)	1.00(4)	0.55(1)
D	0.65(2)	0.69(2)	0.55(3)	1.00(4)

表 3 では，4 人の被験者 A から D とし，対応する値が要約提供システムの算出したユーザ間の類似度を表し，括弧内の値がユーザの主観によっておこなった要約の順位付けを表している．表 3 をもとにシステムが算出したユーザ間の類似度と各ユーザがおこなった順位付けとの相関係数を求めた．求めた相関係数は 0.87 であり，自由度 14 において 1 %有意水準で強

*3 http://www-cl.tufs.ac.jp/pub/tools/posum/

い相関を確認した．以上より，本要約提供システムが提供する要約は，ユーザの嗜好を反映していると考える．

4. 関連研究

複数の視覚的効果を組み合わせてドキュメントの読み易さを支援する研究として，Suh[2] と Graham[10] の研究がある．いづれのシステムも全体の縮小画像を用いた Overview+Detail 効果と Liner 効果を組み合わせたインタフェースを持つ．本スキーミング支援システムは，Overview+Detail 効果と Fisheye 効果を組み合わせたインタフェースを適用している．この２つの効果を組み合わせた研究はまだないと思われる．

多くの対話的テキスト要約は，重要文抽出手法により提示された要約をもとにして，ユーザに首尾一貫性の判断を委ねたアプローチをおこなっている [3][13]．TXTRACTOR[4] は，単語の頻度，手がかり語，固有名詞と文の位置のパラメータをユーザが調節することで首尾一貫性のある要約を提供する．Saggion[8] のシステムは，ユーザが論文のアブストラクトに対し読みたい個所を選択することでユーザの嗜好を反映したアブストラクトを提供する．本研究では，スキーミング支援システムをインタラクティブに使用した履歴をもとに協調フィルタリングをおこなうことで，ユーザの嗜好に近いユーザの要約を提供する．これにより，首尾一貫性と嗜好を反映した要約提供の自動化に対する問題をある程度解決できたと考える．

協調フィルタリングは，ユーザ間の関係を求めるために各ユーザプロファイルを使用する．Smart Courier[14] は，アノテーション情報をユーザプロファイルとする論文推薦システムである．Concept Index[1] は，各ユーザがドキュメント内で指定した興味のある個所のキーワードをユーザプロファイルとして使用したリーディングポイント推薦システムである．本要約提供システムは，スキーミング支援システムにおいてユーザが選択した文をユーザプロファイルとしている．

5. さいごに

本研究では，ドキュメントスキーミング支援環境としてスキーミング支援システムと要約提供システムを構築した．スキーミング支援システムは，コンピュータ画面からのスキーミングを支援するために Overview+Detail 効果と Fisheye 効果を組み合わせたインタフェースを適用した．本システムの Overview には自動生成した目次型インタフェースを使用し，Fisheye 効果の Focus には提案する重要文抽出手法を使用した．要約提供システムは，スキーミング支援システムの履歴をもとに協調フィルタリングによってユーザの嗜好を考慮した要約を提供した．

評価実験では，重要文抽出手法，スキーミング支援システム，および要約提供システムの評価をおこなった．重要文抽出手法では，提案手法の方が他の手法に比べ F-measure が平均 0.48 と高く，安定した値を得た．スキーミング支援システムでは，既存研究においてコンピュータ画面からのスキーミングが困難であったのに比べ，紙面からのスキーミングと同等より優位な効率的読解度の値を得た．要約提供システムでは，ユーザが主観的に選んだ要約の順位とシステムが各ユーザの履歴から計算したユーザの類似度の順位との間に 0.87 と強い相関があることを確認した．

今後の課題は，本スキーミング支援環境において画像や動画などの他のメディアの組み込みと，各機能の制度と支援機能の拡張などが挙げられる．

参考文献

[1] A.Voss, K.Nakata, M.Juhnke; Concept Indexing. In: Hayne,S.C;Proc. Interna tional ACM SIGGROUP Conference on Supporting Group Work, pp14-17, 1999

[2] B.Suh,A.Woodruff,R.Rosenholtz,A.Glass;Popout Prism; Adding Perceptual Principles to Overview+Detail Document Interfaces, Proc. CHI 2002, 2002

[3] Boguraev,B.K.B,Wong,Y.Y.,Kennedy,C.Bellamy, R.K.E, Brawer,S.and Swartz, J; Dynamic presentation of document content for rapid on-line skimming.,Proc of AAAI Spring Symposium on Intelligent Text Summarisation,1998

[4] D.McDonald,H,Chen.Using Sentence Selection Heuristics to Rank Text Segments in TXTRACTOR, In Proc of the 2nd ACM/IEEE Joint Conference on Digital Libraries, pp25-38, 2002

[5] David M. Levy; I read the news today, oh boy: reading and attention in digital libraries, Proceedings of the second ACM international conference on Digital libraries, pp202-211, 1997

[6] 難波英嗣，奥村学;" 観点に基づいた新聞記事の重要文選択に関する心理実験と考察", 言語処理学会第４回年次大会併設ワークショップ,「テキスト要約の現状と課題」，pp30-35, 1998

[7] Hearst,M.A; Segmenting Text into Multi-Paragraph Subtopic Passages, Computational Linguistics, 1997

[8] H.Saggion, G.Lapalme; The Generation of Abstracts by Selective Analysis, AAAI'98 Spring Symposium, 1998

[9] Jackson,M.D., McClelland,J.L.;Processing determinants of reading speed, Journal of Verbal Learnig and Verbal Behavior, pp151-181,1976

[10] J.Graham; The Reader's Helper, A personalized Document Reading Environment,CHI'99,1999

[11] Mills,C.B., Weldon, L.J.; Reading text from computer screens, ACM Computing Surveys, pp329-358,1987

[12] M.Paul, M.Paula; Reading and skimming from computer screens and books:The paperless office revisited?, Behaviour & Information Technology, P257-266.1991

[13] 奥村学，難波英嗣;" テキスト自動要約に関する研究動向", 自然言語処理,「テキスト要約のための言語処理」特集号 Vol6 No.6，1999

[14] S.Ito,Y.Sumi,K.Mase; Supporting Knowledge Sharing by Document Annotation at an Exhibition Site, Proc. of 15th annual conference of JSAI,2001

[15] Schwarz.E.,Beldie.I.P,Pastoor S; A comparison of paging and scrolling for changing screen contens by inexperienced users,Hum Factors 25,pp279-282,1983