

再利用部分の抽出による プレゼンテーションストーリーの変遷の可視化

Visualization of transitions of presentation stories by extracting reused parts

中沢 拓磨*¹
Takuma Nakazawa

久保田 秀和*^{1*2}
Hidekazu Kubota

角 康之*¹
Yasuyuki Sumi

西田 豊明*¹
Toyoaki Nishida

*¹京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University

*²日本学術振興会特別研究員 PD
JSPS Research Fellow PD

It can be said that presentation stories that are continuously accumulated by people are valuable knowledge archives. We have developed the system that visualizes transitions of personal presentation stories for the purpose of making it clear how personal knowledge develops. In order to visualize presentation stories, we consider reused slides. A new story is often made with reused slides that are modified from a new viewpoint. The system can distinguish reused slides by comparing the words and images even if they are partly modified. Since the system piles up reused slides in order of time and connects stories as a graph, the transition of stories is visualized.

We applied the system to real presentation data. The system makes it clear that there exist a lot of reused slides. We have discovered various meanings from shapes of crossovers of the presentation stories. The answers of the questionnaire of the experiment of using the system tells us that the system lets users remember the background and context of compiling stories and gives them indexes of improvement of documents.

1. はじめに

プレゼンテーションを行うことは、議論を深め知識を共有する上で重要な活動である。このプレゼンテーションという知的活動は、発表者の述べたい内容を網羅的に包含するとともに、重要である部分を簡潔にまとめ上げたものである。その知識の密度は非常に高いものとなっている。近年ではスライドによる情報提示を行うプレゼンテーションソフトウェアが広く利用されているが、このようなソフトウェアで作られたプレゼンテーション資料もまた、知識を構成する重要な一翼を担っていると言える。よって、長期的に蓄積されたプレゼンテーション資料は、人々の知識アーカイブであると言える。

また、プレゼンテーション資料は単に蓄積されていくだけでなく、改訂という形で頻りに掘り返されるものである。しかしながら、それぞれのプレゼンテーションはファイルという形で別々にまとめられ、本来あるはずのプレゼンテーション間の関係が失われている。また、大量のプレゼンテーションファイルを所有する状況下では、最新のものだけが顧みられるようになり、古いプレゼンテーションが行われた際の文脈は次々と失われることになる。

このような現状を省みると、個人の知識が失われないようにするためには、新しいプレゼンテーションと古いプレゼンテーションの関連性を示す手法が必要であると考えられる。プレゼンテーション間の関係を見つめなおすことで、これまで積み上げてきた知の文脈の広がりを認識するようになり、重要な部分の抽出や、不要な部分の発見が可能となる。本研究ではこのような要求を可能とする可視化手法を提案する。

2. スライド再利用への注目

ここでは事前の仮説として、新しいプレゼンテーション資料を作成する際の行動モデルを立てることによって、可視化システムをデザインする。

ある話題に関するプレゼンテーションの内容が、プレゼンテーションを行うたびに全く異なるということは稀である。多くのプレゼンテーションは以前に行ったプレゼンテーションと関連があり、したがってプレゼンテーション資料の内容も、以前の内容を含むようになる。このように内容の関連するプレゼンテーションを行うとき、多くの場合は以前の資料の改訂版を用いるなど、スライドを再利用することで行われると考えられる。

ここから、プレゼンテーションストーリーは再利用という活動を通して変遷していくものと予想される。この再利用に対して2つのレベルを考える。ひとつはスライドレベルの再利用であり、もうひとつはスライド要素レベルの再利用である。スライドレベルの再利用とは、あるストーリーのスライドが、他のストーリー上に再登場することである。スライド要素レベルの再利用とは図1のように、スライド中の文章や画像を他のスライドでも利用することである。再利用が行われる様は、文脈に従って多種多様なバリエーションがある。プレゼンテーション資料から再利用の文脈を抽出することは、本研究の狙いの一つである。

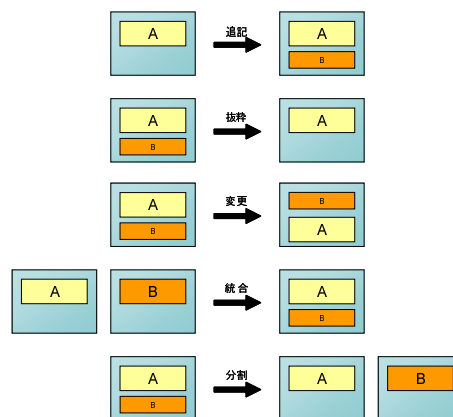


図1: スライド要素レベルの再利用

連絡先: 中沢拓磨, 京都大学大学院情報学研究科, 〒606-8501
京都市左京区吉田本町 工学部 10 号館 214 号室,
TEL:075-753-5371, nakazawa@ii.ist.i.kyoto-u.ac.jp

3. 再利用部分の抽出

我々はこのような再利用を抽出するための方法を考案し、実装した。プレゼンテーション用のスライドは、主に箇条書きの文章と画像によって構成されているため、これらの特徴を手がかりに再利用部分を抽出することにした。これに先立って、スライドを、文章はテキスト形式に、画像はPNG形式とJPEG形式に変換する処理をあらかじめ行った。

3.1 文章に基づく判別

まず文章に基づく判別手法を述べる。スライド中の文章をMeCab[工藤 06]を用いて形態素に切り分け、名詞と動詞のみを抽出する。抽出されたものをリスト化したものを単語リストと呼ぶ。この単語リストから、出現単語の成分を1、非出現単語の成分を0とする特徴ベクトルをスライドに与える。ここで本手法における2スライド間の再利用距離の評価式を式1に与える。変数は、 $p = (2 \text{ つのスライドに共通して現れる単語数})$ 、 $s = (\text{規模の小さい方のスライドの単語数})$ である。 c は定数である。これは $[0, 1)$ 上の距離であり、値が大きいほど再利用の度合いが高いことを意味する。

$$D_{\text{reuse}} = \frac{p}{s}(1 - e^{-sc}) \quad (1)$$

3.2 画像に基づく判別

再利用において画像がどう取り扱われているかを、いくつかのプレゼンテーション資料について調べると、再利用とともに拡大縮小などの変形が行われている様子が見られた。よって再利用の追跡のためには、変形による影響を考慮する必要がある。ここでは変形前の素材画像どうしが一致するかどうかと、スライドに表示されたままの形の画像どうしが一致するかどうかという2つの基準を用いた。変形前の素材画像どうしが一致するとは、すなわち画像に加えられた変形を無視することを意味する。また、スライドに表示されたままの形の画像どうしが一致するとは、共通する素材画像を使用し、さらに変形方法も同じであることを意味する。素材画像については単語と同様に再利用距離を用いる。

ところで画像の一致を判別する方法として、この他に画像処理によるものが考えられるが、画像処理の場合、スライド全体から文字部分と画像部分に切り分けなくてはいけないとともに、画像部分に関しても、画像の意味ごとの部分に分けなくてはならない。これは大変困難であるため、今回は画像処理は使用せず、PowerPoint形式のファイルが内部的に持つ画像情報を使用した。ただしこの方法では、素材画像と表示画像の取得方法が異なるため、素材画像の取得数のほうが表示画像の取得数よりも少ない場合がある。5.1節におけるAのデータについて画像数を比較したところ、表示画像が4324個取得できたのに対し、素材画像は1634個しか取得できなかった。

3.3 クラスタリング

再利用の度合いを表す指標を得た上で、再利用スライド群を構成する必要がある。再利用スライド群とは同種類の再利用スライドの集合である。クラスタリングアルゴリズムとして、以下のような非階層的クラスタリング手法を用いた。

最初の状態では、すべてのスライドはそれぞれ大きさが1のクラスタを構成する。次に、“少なくとも1つの表示画像を共有する”という条件をみたくスライドどうしを同じクラスタに併合する。これは、表示画像を共有するスライドは、スライドごとコピーしたものであるという強い仮説を用いている。表示画像についてはこのような強い結合が行われるため、再利用距離に基づく判別は行われない。このように粗いクラスタリング

を行った後、単語と素材画像の再利用距離を用いて細かなクラスタリングを行う。ここでクラスタ間距離を評価する必要があるが、この方法としてスライドマージによる方法をとった。これにより、スライド間の再利用距離と同じ方法でクラスタ間の再利用距離が計算可能となる。

クラスタリングのアルゴリズムは以下である。

1. クラスタを1つ選ぶ
2. そのクラスタに対して、以下のいずれかの条件を満たすクラスタをすべて併合する
 - (a) 単語に関する再利用距離が閾値 t_w 以上である
 - (b) 素材画像に関する再利用距離が閾値 t_i 以上である
 - (c) 両クラスタが素材画像をただ1つだけ持ち、その画像が一致する
3. 1-2を繰り返す
4. すべてのクラスタが併合できなくなれば終了

これによって、同一座標に重なるべきスライドが、同一クラスタに所属することになる。

3.4 ストーリーの列挙

クラスタは全体としては再利用関係にあるスライドで構成されているが、その中の具体的な2枚を取り出したとき、必ず再利用関係にあるとはいえない。また5.1節にあるように、再利用の誤検出の問題が内在する。これらの問題の解決のために、全スライドを詳細に確認する作業を行い、手動で再利用か否かの対応表を作成した上で以下の方法をとった。まずはじめに、人によって明らかに再利用と判定できる基準を見つけ、ルールとして整備した。このルールの例としては以下のようなものがある。

1. スライド間の表示画像一致率が100%なら再利用。
2. 規模の小さなスライドの単語数が10以上で、スライド間の単語一致率が90%以上なら再利用。
3. 規模の小さなスライドの単語数が10以上で、スライド間の単語一致率が50%~20%なら非再利用。

ルールを用いる方法では確実に判定できないものについては、分類器を用いた。分類器はサポートベクターマシンライブラリであるLIBSVM[Chang 01]を用いた。学習は単語数が10以上の場合と、10未満の場合に分けて行った。これは単語数が10未満の場合、単語一致率が100%であっても再利用関係にないという状況が見られたことから、この2つは識別空間の性質が異なると考えられるからである。また分類器に渡すベクトルは、単語数が10以上の場合は単語一致率、表示画像一致率、素材画像一致率とした。単語数が9以下の場合には、これに加えて単語数と画像数を与えた。訓練セットは5.1節における19329例を用い、このうちルールの適用範囲外であるものを数えると、単語数が10以上のものは2212例、単語数が10未満であるものは1772例であった。これについて学習した後、訓練セットを再度与えて学習の成果を調べた。その結果、単語数が10以上の場合は79.9%の正解率を、単語数が10未満の場合は98.5%の正解率を得た。ただし、訓練セットは特定個人に閉じたデータを用いたため、一般のデータに対してはうまく分類できない可能性がある。

4. プレゼンテーションストーリーの変遷の可視化

4.1 スライドの重ね置き

プレゼンテーションストーリーを構成するスライドには、再利用スライドと、そうでないスライドの2種類がある。このことをストーリー作成者が確認できるようにするとともに、他のどのストーリーとスライドが再利用関係にあるかを示すための手法として、重ね置きを提案する。重ね置きとは複数のストーリーを平面上に置き、そのうち同種の再利用スライドを重ねるものである。再利用スライドを同じ座標におくことで、ストーリーが互いに交差するように表現され、ストーリー間の関係を直感的に与えることが出来る。また時間の順序に従って重ねることで、変遷という時間軸に沿った要素を可視化することが出来る。

4.2 ユーザインタフェース

スライドをノード、ストーリーのつながりをエッジとみなすと、プレゼンテーションストーリーはグラフとみなすことができる。このようなグラフを自動配置するために、力指向グラフ配置手法のひとつであるバネモデル [Eades 84, Kamada 89] を用いた。この配置をユーザに提示するためのインタフェースとしては、コンテンツ空間である知球 [久保田 04] を用いた。

5. 実験とその考察

5.1 クラスタリング適用実験

これらの方法を用いて実際にクラスタを生成した。適用対象は、筆者および筆者の研究室に在籍する研究員が、それぞれこれまでに作成してきたプレゼンテーションファイルである。この様子を表1に示す。ただし最小クラスタ規模はすべて1である。

表 1: クラスタ状況

作成者	ファイル数	スライド数	クラスタ数	最大クラスタ規模	最小クラスタ数
A	193	3593	1288	68	614
B	77	1882	446	72	209
C	49	1333	530	30	281
D	120	2528	913	49	442

A についてはさらに分析を進めた。1288 クラスタの中に、重ね置きの妥当性を判定できるスライドの対は 19329 対あった。19329 対の妥当性を人手によって判定したところ、7468 対が重ね置きが妥当な対であり、11861 対が妥当でない対であった。したがって妥当でない対が 61.4%含まれているということである。ただしこれは全対について調べたものであるから、規模の大きなクラスタの影響が、規模の2乗で増加するという点に注意が必要である。たとえば最大クラスタの大きさは68であるから、その全対は4556であり、全体への影響は実に23.6%に相当する。

これらの妥当でない対の大部分は、素材画像によるクラスタ結合への影響が強すぎるのが原因であった。具体的に最大のクラスタについて分析する。このクラスタのコアは4つの画像で構成されているスライドである。さらにこのコアには複数のバリエーションがあり、「人」を表す画像を差し替えたものと、英訳版がある。しかも、そのうちアニメーションを加えたバージョンと加えていないバージョンが存在する。加えてこ

のスライドは4つのパートに分割して説明されるというバリエーションをもつ。また、概要や背景といった説明のスライドにも同じ「人」の画像が象徴的に使われている。このように多彩なバリエーションに使われる画像が、クラスタ肥大に関わっていた。一方、表示画像による結合が問題となる例はあまり見られなかった。

この実験から改善のための指標を得た。この結果を基に、ストーリー列挙の際にさらにルールと分類器を用いるように改善を行い、妥当でない対を排除するようにした。この改善を行った上で、5.2節の実験を行った。

5.2 システム利用実験

システムが提示するストーリー配置をストーリーの作成者に実際に見てもらい、ストーリーを作成したときの背景を基にストーリー交差の部分を説明してもらおうという実験を行った。実験に用いたシステムの提示例を図2に示す。束になったスライドが、ストーリーの流れに従って直線で繋がることにより表されている。ただし本稿における理解のために、ストーリーの流れを示す矢印を実際より強調表示している。

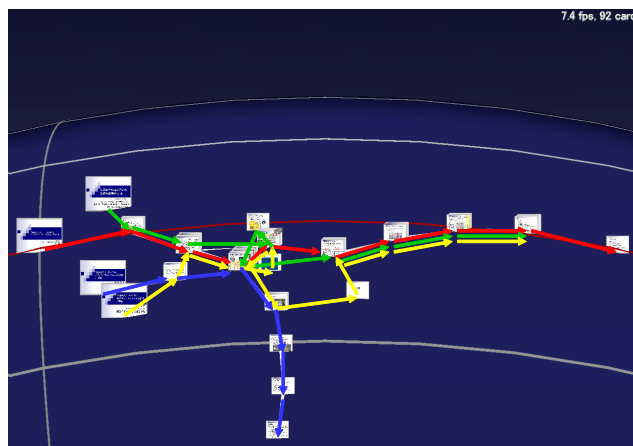


図 2: システムの提示例

実験参加者の回答から、いくつかの特徴的な交差の意味が明らかになった。実験における交差構造の現れ方を図3および表2にまとめる。ここから、ストーリー変遷の基本となるものが、ストーリーの分岐と合流による、聴衆に合わせた部分的改編や要約あるいは詳細化というものであることが示唆された。これらは異なる話題、あるいは異なるプレゼンテーション作成者であっても共通して現れている。他の特徴について考えると、例1と例2では倒置と分割という構造が共通して現れているのに対して、例3と例4では表れていない。このことから、同じ人であれば同じストーリー特徴が表れやすいという仮説が立てられる。

表 2: 交差構造の現れ方

	作成者	スライド数	ノード数	特徴
例 1	A	53	25	要約, 改編, 倒置, 分割
例 2	A	44	21	要約, 改編, 倒置, 分割, 繰返し
例 3	D	76	53	要約, 中核
例 4	D	79	58	要約, 改編, その他

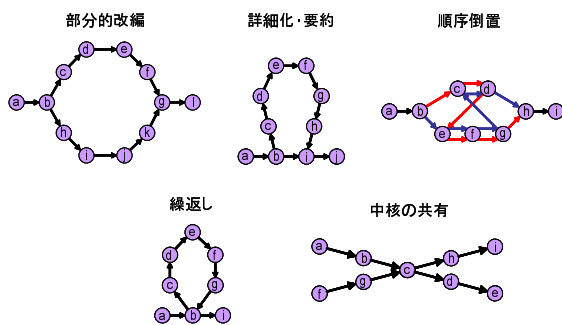


図 3: ストーリー構造のトポロジ

5.3 システムの効果

実験に伴うアンケートより、本研究の提案するシステムはユーザに対して以下のような効果があることが示唆された。まずプレゼンテーションストーリーの変遷が可視化されることによって、自分のプレゼンテーション作成の行動パターンに気がつくというものである。例えば、貴重な実験データは何度も同じものを使用するということが気がついたり、同じ主張を、全く違う切り口からストーリー展開をしていることに気がついたりする様子が見られた。また、プレゼンテーションを作ったときの背景や文脈を思い出しやすいということが示唆された。スライドの順序についてかなり悩んだことを思い出したという回答や、説明不足で分かりにくいストーリーを、詳細化して改善したことを思い出したという回答を得た。さらにプレゼンテーションの改善の指標としても役立つ可能性がある。例えばループの存在する部分では、同じ内容を同じ観点から何度も説明していることになるが、そのようなやり方は聴衆にとっては退屈であるため、同じ内容を繰返し説明する場合は違う観点から切り込んだ方がよいといったような、改善の指針を与えることが出来る。

総合的な観点からは、複数のストーリーの中から出来の悪いストーリーと出来の悪いストーリーを発見することが出来るので、このシステムを用いることで新たに最善のプレゼンテーションが作れそうだという意見があった。新しいストーリーが古いストーリーより必ずしも良くなっているとは限らないため、古いストーリーも合わせて全体の良い部分を選びすぎると、最善のストーリーが作れるのではないかと考えられる。

5.4 問題点

素材画像による結合力が強すぎることを原因とする、再利用の誤検出の問題が提示された。補正関数の定数 c と閾値 t_i を見直す必要があると考えられる。

またスライド内容の分割という再利用状況においては、再利用スライドを重ねることに問題があることが示された。同じ内容を重複して説明していると誤解される恐れがあるからである。

加えて、開発段階からの既知の問題としてはスケラビリティが挙げられる。これはパネモデルを用いていることが原因であるが、次数の高いノードの周辺が過密になるという現象が起こってしまう。このような状況では、ユーザは過密部分の構造を読み取ることが出来なくなるため、問題となる。

6. 関連研究

ドキュメントのバージョン管理は、富士ゼロックスの DocuWorks や、サイボウズのサイボウズ Office など、各種の商用ドキュメント管理システムにおいて広く行われている。しかしこれらのシステムでは、ドキュメントを改訂したことを人がシステムに知らせる必要があり、過去のドキュメント集合から自動的にドキュメントの変遷を発見することはできない。

ドキュメント間の関連性をグラフで表現する手法も数多く提案されている [杉山 93]。本研究の特色は、複数枚のスライドから成るプレゼンテーションストーリーを対象とすること、および複数の共通スライドをひとつのノードとして扱うことにより、ストーリーのグラフを生成する点にある。

7. おわりに

本研究では、プレゼンテーションストーリーにおける再利用スライドの存在を明らかにし、その状況を重ね置きの手法を用いることで可視化するシステムを開発した。その際、再利用距離という尺度を導入することで、再利用スライドの判別を可能にした。システムは、再利用スライドのクラスタリングと、パネモデルによる自動配置を主要技術として用いた。ユーザインタフェースとして知球を用いることで、自動配置後にユーザによる構造の修正が可能となるインタラクティブ性を備えた。

配置後のストーリーを実験によって確認した結果、ストーリーの交差の構造が様々にあることが分かった。しかし主要な構造は、分岐と合流を繰り返すことによる、改訂と呼ばれる変遷であった。また、本システムで可視化したストーリー変遷は、ユーザにとってプレゼンテーション資料作成の際の行動パターンを明確にし、ストーリー作成の背景・文脈を想起させたり、改善の指標となることが示唆された。今後はストーリー分岐の分析を進めるとともに、口頭によるプレゼンテーションの場を含めて研究を行って行きたい。

参考文献

- [Chang 01] Chang, C.-C. and Lin, C.-J.: *LIBSVM: a library for support vector machines* (2001), <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>
- [Eades 84] Eades, P.: A heuristic for graph drawing, *Congressus Numerantium*, Vol. 42, pp. 149-160 (1984)
- [Kamada 89] Kamada, T. and Kawai, S.: An algorithm for drawing general undirected graphs, *Information Processing Letters*, Vol. 31, No. 12, pp. 7-15 (1989)
- [久保田 04] 久保田 秀和, 角 康之, 西田 豊明: 「知球」: 持続的に発展可能な時空間記憶の構築, 情報処理学会研究報告「ヒューマンインタフェース」, Vol. 2004, No. 90, pp. 1-8 (2004)
- [工藤 06] 工藤 拓: MeCab: Yet Another Part-of-Speech and Morphological Analyzer (2006), <http://mecab.sourceforge.net/>
- [杉山 93] 杉山 公造: グラフ自動描画法とその応用, 計測自動制御学会 (1993)