プレゼンテーション蓄積検索システムにおける適合度計算の改善

岡本　拓明† 　小林　隆志†† 　横田　治夫††,†††

† 東京工業大学 工学部 情報工学科　〒 152–8552 東京都目黒区大岡山 2–12–1
†† 東京工業大学 学術国際情報センター　〒 152–8550 東京都目黒区大岡山 2–12–1
††† 東京工業大学 大学院 情報理工学研究科 計算工学専攻　〒 152–8552 東京都目黒区大岡山 2–12–1
E-mail: †okamoto@de.cs.titech.ac.jp, ††tkobaya@gsic.titech.ac.jp, †††yokota@cs.titech.ac.jp

あらまし　我々は，講義・講演資料と動画をメタデータで統合するプレゼンテーション検索蓄積システム UPRISE を提案してきた．本稿では，UPRISE での検索精度の向上を目的とし，これまでの適合度計算手法に，シーンのプレゼンテーション単位での出現頻度と，説明を伴わないスライドの出現を考慮する改良を行う．さらに本論文ではそれぞれに対する実験を行い，有効性を確認する．

キーワード　web とインターネット，e-learning，情報統合，情報検索，ユーザインタフェース

Improvement of Matching Functions for Retrieving Unified Presentation Contents

Hiroaki OKAMOTO†, Takashi KOBAYASHI††, and Haruo YOKOTA††,†††

† Department of Computer Science, Faculty of Engineering, Tokyo Institute of Technology
Ookayama 2–12–1, Meguro-ku, Tokyo, 152–8552 Japan
†† Global Scientific Information and Computing Center, Tokyo Institute of Technology
Ookayama 2–12–1, Meguro-ku, Tokyo, 152–8550 Japan
††† Department of Computer Science, Graduate School of Information Science and Engineering,
Tokyo Institute of Technology　Ookayama 2–12–1, Meguro-ku, Tokyo, 152–8552 Japan
E-mail: †okamoto@de.cs.titech.ac.jp, ††tkobaya@gsic.titech.ac.jp, †††yokota@cs.titech.ac.jp

Abstract　We have proposed UPRISE(Uified Presentation Slide Retrieval by Impression Search Engine), which unifies presentation slides used in a lecture and a video of the lecture, using metadata. In this paper, to enhance accuracy of query, we improve functions with considering scene frequency in a presentation and unexplained slides. We evaluate the matching functions using actual presentation contents.

Key words　Web and the Internet, e-learning, information integration, information retrieval, user interface

1. はじめに

我々はこれまで，教育コンテンツの統合機構，及び統合された教育コンテンツに対する高度な検索機能を実現するシステムとして，UPRISE(Unified Presentation Slides Retrieval by Impression Search Engine) を提案してきた [1]〜[5].

UPRISE は，講義ビデオやプレゼンテーション資料等の教育コンテンツをメタデータによって統合することで，それらの教育コンテンツの同期表示を実現するシステムである．さらに統合された教育コンテンツに対する高度な検索機能や，検索結果である多量な資料を効率的に提供するためのユーザインタフェースを備えている．

メタデータによる結合を実現するために，UPRISE では，ストリームメディアをシーンの連続であると抽象化し，各シーンと資料の対応情報と，各シーン，資料の検索用インデクスを格納し，検索に利用している．

UPRISE で提案しているコンテンツ検索機能は，スライドの提示時間や前後関係を加味したキーワードに対する適合度を，検索指標として利用している．適合度とは，ある検索キーワードに対して動画中の各シーンごとに算出されるポイントであり，現在はシーン中で使用しているスライドの情報や，そのスライドの説明に要した時間などを利用している，スライドが複数回

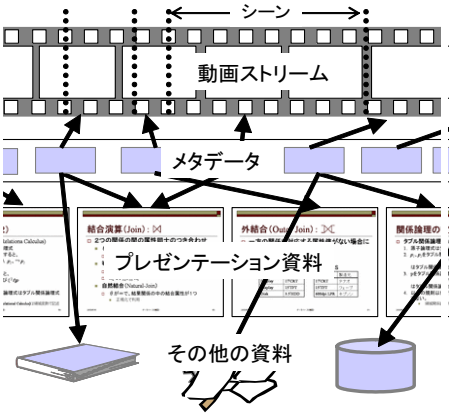


図 1　プレゼンテーション動画と資料の統合

登場する場合は，同じスライドであっても別のシーンとしてポイントが計算される．UPRISE はこの適合度を用いてポイントを算出し，ユーザに動画のシーンで用いられたスライドのサムネイルをポイントの順で表示することで，単純な文字列検索のみの従来の e-ラーニングシステムでは行うことができない，重要なシーンの効率のよい検索を可能にしている．

本論文では，検索精度の向上を目的とし，主にキーワードがどれだけプレゼンテーション内を網羅しているかを考慮していた従来の適合度に，キーワードにどれだけ目的の検索物を特定する性質があるかを考慮した手法を提案する．

解説を省いているシーンやバックトラック時の途中のシーンが含まれると，前後関係の影響が不適切な場合があった．そこでスライドの切り替え時に生じるノイズの影響を減らす手法を提案する．また，提案手法について効果を確認するために，評価実験を行う．

2. UPRISE

2.1 UPRISE の概要

メタデータを用いた，UPRISE のプレゼンテーション動画と資料の統合の概念図を図 1 に示す．メタデータには，動画のどの時刻にスライドの切り替えが起こったかというシーン情報と，その際にどのスライドを用いていたかという同期情報に加え，検索の際に使用するスライドに含まれる文字列情報に対するインデクスを含める．これによって使用されたスライドの順序とスライド毎に要した時間という情報を検索に利用することが可能になる．これらの情報を保持するメタデータによってコンテンツを緩く結合することにより，それぞれのコンテンツに修正を加えることなくコンテンツの同期表示を実現し，柔軟な統合を可能にしている．

UPRISE では，スライドが複数回出現する場合は図 2 のように別のシーンとし，それぞれについてポイントを算出する．これは，任意のプレゼンテーションの任意の地点の検索をするためである．よって，それぞれのプレゼンテーションは対応する動画のシーンの集合であり，格納コンテンツ全体は，プレゼンテーションの集合になっている，

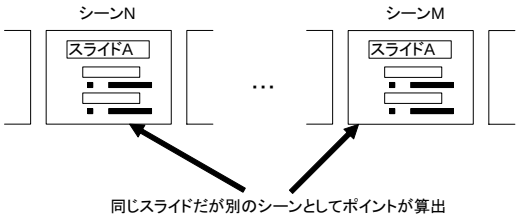


図 2　スライドとシーン

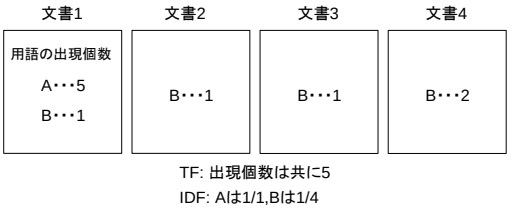


図 3　idf による頻度考慮の例

2.2 文書頻度を利用した重み付け

情報検索の分野では重み付けの方法として，キーワード頻度と共に文書頻度を考慮する手法が提案がされてきた．その代表的な手法が tf.idf [6] である．また，その拡張として widf [7] が提案されている．以下では，tf.idf と widf について説明する．

2.2.1 tf.idf

idf(文書頻度の逆数:Inverse Document Frequency) [6] は単語の出現する文書の頻度で単語の検索語としての重要度を表す．例として図 3 を考える．キーワード A, B は文書 1〜4 中に共に 5 回出現しているために、tf(キーワードの出現頻度:Term Frequency) による重みは共に等しい．しかし，文書頻度の逆数 (idf) は，文書 1 にのみ出現するキーワード A のほうが大きい．これは，A のほうが検索語として重要であることを表している．この tf と idf を組み合わせた重み付けをすることで，文書の特定性 (specificity) と網羅性 (exhaustivity) を兼ね備えた検索をすることができる．

2.2.2 widf

一方 tf.idf の拡張として，widf [7] が提案されている．widf は，tf を正規化することにより検索語の重要性を表している．複数キーワードの例として，図 4 を考える．2 つの文書が，キーワード k_1，k_2 により，図 4 のような tf のポイントを得ていたとすると，tf では文書 A のポイントが高くなる．また出現している文書数が同じであるので idf でのポイントは等しい. widf では，文書 A のポイントは 100/180 + 10/30 = 160/180 ポイントとなり，文書 B のポイントは 80/180 + 20/30 = 200/180 ポイントとなるため，文書 B のポイントが高くなる．これは，出現頻度の低い k_1 が k_2 より検索語として重要であると考え，idf で区別がつかない例でも考慮することができる．

2.2.3 文書頻度を考慮する場合の問題点

2.2 で説明したように，文書頻度を考慮することにより，検索精度が向上する場合がある. しかし，UPRISE では 2.1 で述べたように，スライドは複数回登場する可能性があり，それぞれは別のシーンとしてポイントを算出する．しかし，キーワー

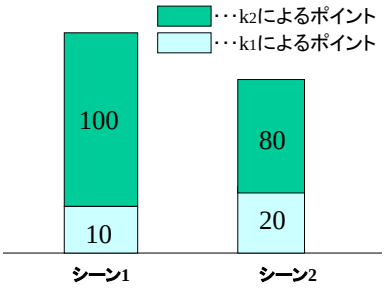


図 4　widf による頻度考慮の例

ドの出現頻度と文書頻度を考慮した場合では，スライドに登場する用語に重み付けをするため，同じスライドであれば，ポイントが等しくなり，シーンに対する順序付けをすることはできなかった．そこで、今までの UPRISE では以下の適合度を考え，シーンの順序付けを行い．tf.idf に比べて高い精度を得ていた [4]. その適合度について詳しく説明する．

2.3　従来の適合度

UPRISE が現在検索に用いている以下の適合度について説明する [4].

- I_p：スライドの文書構造 (インデント等) を利用した適合度
- I_d : I_p に時間情報を考慮した適合度
- I_c : I_d にスライドの前後関係を考慮した適合度
- I_{and} : I_c の複数キーワードの AND 検索の場合の適合度
- I_{or} : I_c の複数キーワードの OR 検索の場合の適合度
- I_t^n：複数キーワードの AND 検索と OR 検索の混合している場合の適合度

2.3.1　適 合 度 I_p

適合度 I_p はスライドの文書構造を考慮した適合度であり，以下の式によって定義される．

$$I_p(s,k) = \sum_{l=1}^{L} P(s,l) \cdot C(s,k,l)$$

ここで，s はスライド，k はキーワード，l は行数，$P(s,l)$ はスライド s で行 l に与えられるポイント，$C(s,k,l)$ はスライド s でライン l にキーワード k が含まれる個数を表している．

この適合度によってキーワードの出現個数に加えて，インデント毎の重みを考慮できる．

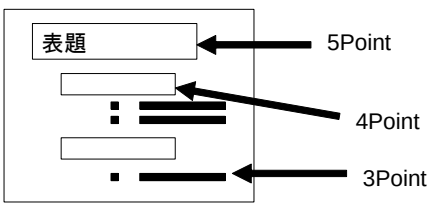


図 5　適合度 I_p の例

例えば図 5 のように，表題に 5 ポイント，1 行目に 4 ポイント，2 行目に 3 ポイントとなるような関数 $P(s,l)$ が与えられていて，表題に 1 個，1 行目に 1 個，2 行目に 2 個現れているキーワードの，そのスライドでの I_p は，$5 \cdot 1 + 4 \cdot 1 + 3 \cdot 2 = 15$ となる．

2.3.2　適 合 度 I_d

適合度 I_d はスライドの時間情報を付加した適合度であり，以下の式によって定義される．

$$I_d(s,k,\theta) = \sum_{l=1}^{L} T(s)^{\theta} \cdot P(s,l) \cdot C(s,k,l)$$

ここで，$T(s)$ はスライド s の説明に要した時間，θ は時間の影響度を決めるパラメータを表している．これによって，長く説明したシーンを重要視することができる. 例えば $\theta = 1$ とすると，図 6 のように，I_p が 20 ポイントのスライド A が 30 秒説明されているシーン A1 の I_d は，$20 \cdot 30 = 600$ となる．一方，スライド A が 15 秒説明されているシーン A2 の I_d は，$20 \cdot 15 = 300$ となる．この適合度によって同じスライドで別のシーンである場合に対して，順序付けすることができる．

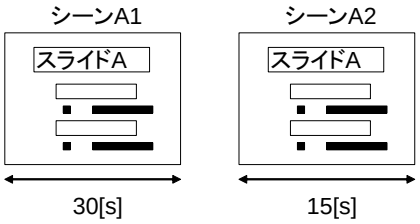


図 6　適合度 I_d の例

以下の説明では，表記の簡単化のため，

$$I_l(s,k,l,\theta) = T(s)^{\theta} \cdot P(s,l) \cdot C(s,k,l)$$

とする．

2.3.3　適 合 度 I_c

適合度 I_c はスライドの前後関係を付加した適合度で，

$$I_c(s,k,\theta,\delta,\varepsilon_1,\varepsilon_2) = \sum_{\gamma=s-\delta}^{s+\delta} \sum_{l=1}^{L} E(\gamma - s, \varepsilon_1, \varepsilon_2) \cdot I_l(s,k,l,\theta)$$

と表される．$E(\gamma - s, \varepsilon_1, \varepsilon_2)$ は，前後関係の影響の強弱を決める関数で，δ は影響の範囲を決めるパラメータである．

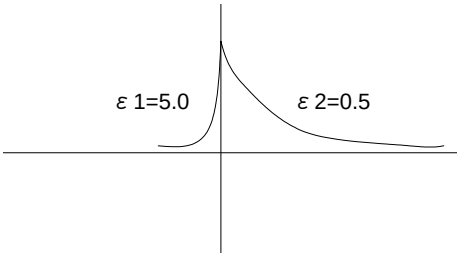


図 7　適合度 I_c の例

この適合度によって，δ の範囲内のシーンのポイントの影響をうける．$E(\gamma - s, \varepsilon_1, \varepsilon_2)$ は以下のように定義され，

$$E(x,\varepsilon_1,\varepsilon_2) = \begin{cases} exp(\varepsilon_1 x) & (x < 0) \\ exp(-\varepsilon_2 x) & (x \geqq 0) \end{cases}$$

ε が小さければ小さいほど，影響をうけやすくなる．例えば $\varepsilon_1 = 5.0, \varepsilon_2 = 0.5$ の時，図 7 のように，そのシーンの後に出てくるシーンのポイントの影響を強く受ける．

2.3.4 適合度 I_{and}

適合度 I_{and} は，複数キーワードの AND 検索のための適合度で，以下のように各キーワードの I_c の積として定義される．

$$I_{and}(s,K,\theta,\delta,\varepsilon_1,\varepsilon_2)=\prod_{l=1}^{m} I_c(s,k_l,\theta,\delta,\varepsilon_1,\varepsilon_2)$$

但し， $K=(k_1,k_2,\cdots,k_m)$ とし，キーワードの集合を表す．

2.3.5 適合度 I_{or}

適合度 I_{or} は，複数キーワードの OR 検索のための適合度で，以下のように各キーワードの I_c の和として定義される．

$$I_{or}(s,K,\theta,\delta,\varepsilon_1,\varepsilon_2)=\sum_{l=1}^{m} I_c(s,k_l,\theta,\delta,\varepsilon_1,\varepsilon_2)$$

但し， $K=(k_1,k_2,\cdots,k_m)$ とし，キーワードの集合を表す．

2.3.6 適合度 I_t^n

複数キーワードで AND と OR の混合した検索を行う時に，適合度 I_t^n を用いる．手順としては，以下のように行う．

（1） まずキーワードの集合が AND と OR で混合している式を，論理積標準形 (CNF) に直す．論理積標準形とは，キーワードの和集合が，積でつながっている形のことを指す．

（2） キーワードの積になっている部分を正規化する．

（3） 和の部分を正規化する．

正規化の方法は，[4] を参照されたい．

3. 適合度改善手法

以下では，主に考慮する適合度は I_c とし，複数キーワードの場合は，I_{and}，I_{or}，I_t^n と同様に算出するものとする．

3.1 従来の適合度の問題点

従来の適合度 I_c は，I_p でキーワードの出現頻度 (Term Frequency) を考慮していた．これは，キーワードの多く現れる文書を抽出すれば、目的のシーンを漏れなく抽出できるという性質 (網羅性:exhaustivity) を利用したものである．

しかし，キーワードの出現するスライドのシーンが多ければ多いほど，キーワードが目的のシーンを特定する性質 (特定性:specificity) がないという問題がある．例として,「UPRISE の出現頻度」≪「スライドの出現頻度」の場合を考える．「UPRISE AND スライド」で検索した時に,「UPRISE」で得た評価ポイントと,「スライド」で得た評価ポイントは，前者のほうがより検索物を特定しているポイントであるが，同等に考えてしまうことで，よりキーワード分布の多い「スライド」でのポイントが大きいシーンが上位にきてしまう．つまり、特定性のないキーワードは，検索語として適していないが，特定性のあるキーワードより出現数が多いため，目的のシーンに適合しない場合であっても，上位のシーンに含まれる数が多くなる．

また，[4] では従来の適合度 I_c と tf.idf を比較し，I_c の精度が高いことを示していたが，tf.idf が従来の I_c より良い場合があると言うことを確かめている．つまり UPRISE にとって，tf.idf のような文書頻度を考慮することは有用である．また，一般的には網羅性と特定性はトレードオフの関係にあり、その 2 つの性質の割合を適切にすることが、効率のよい検索につながると言える．

また，UPRISE ではシーンの自動抽出を画像認識により行っているため，シーンとして説明を伴わないシーンも抽出されてしまう，このため 2.3.3 の適合度において，シーンの前後関係のポイントを考慮していたが，説明を伴わないシーンがポイントの与え方に影響を与えるという問題がある．例として，$\delta=1$ の時を考える．図 8 のように，実際の話の流れは C → A → C

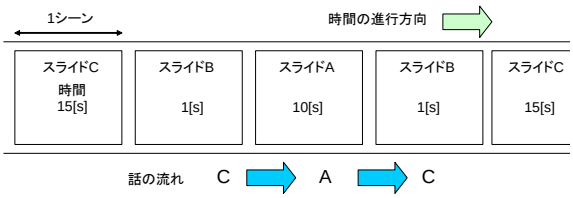


図 8 Ic のポイントに影響を与える例

であっても，解説を省いていたり，バックトラック時の途中のシーンなど，話の流れから外れるシーンが存在する，この場合、I_c は前後 1 枚のスライドの影響を加味するので，2 つ離れたスライド AC 間でのポイントの影響より，AB と BC 間の影響の方が大きい．これではプレゼンテーションの前後の流れを考慮にいれた適合度 I_c の意味が、薄れている．これを以下では「ノイズシーン」と呼ぶことにする．

3.2 文書頻度を取り入れた適合度の改善

情報検索の分野では，単語の特定する性質に重点をおいた，さまざまな提案がされている．ここでは，2.2 で説明した手法を取り入れた手法の提案を行う．

3.2.1 適合度 $I_{c.isfp}$

IDF に対して，単一プレゼンテーションにおけるキーワードの頻度を考慮する適合度を提案する．スライド s の出てくるプレゼンテーション $P(s)$，$P(s)$ に含まれるシーン数 $N(s)$，$P(s)$ にキーワード k が出てくるシーンの数を $pf(k,s)$ とする．これらのパラメータを用いてプレゼンテーション頻度 sfp(Scene Frequency in a Presentation) を次のように定義する．

$$sfp(s,k)=\log\left(\frac{pf(s,k)}{N(s)}\right)$$

例として図 9 を考える. プレゼンテーションが複数個あり，その中のスライドには A, B, C, D, E の 5 種類のキーワードが出現している. スライド中に表示されている文字は，そのスライドに出現している文字とする，全体ではキーワード A は 4/9 の割

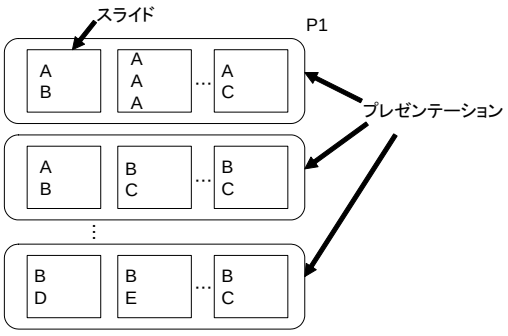


図 9 プレゼンテーション頻度の例

合で出現しているため，キーワード C と文書頻度は等しい. しかし，P1 においてはキーワード A は 3/3 の割合で出現するため特定性がない. 一方，キーワード B は全体では 7/9 の割合出現するため特定性がないが，P1 においてはキーワード B は 1/3 の割合で出現するため特定性がある. よって sfp を考えることにより，プレゼンテーション毎の頻度を考えることの特定性を考慮できる. これによって複数キーワードで検索した時に，プレゼンテーションのキーワード分布の特徴に応じてポイントをつけることができる.

この逆数 isfp に，従来の適合度 I_c との積を適合度 $I_{c.isfp}$ として提案する．つまり，$I_{c.isfp}$ を，

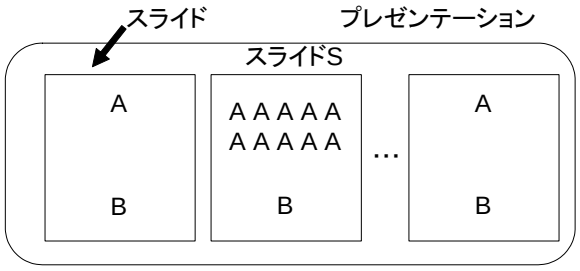
$$I_{c.isfp}(s,k,\theta,\delta,\varepsilon_1,\varepsilon_2) = I_c(s,k,\theta,\delta,\varepsilon_1,\varepsilon_2)\cdot isfp(s,k)$$

と定義する.

3.2.2 適合度 $I_{c.wisfp}$，$I_{c.widf}$

ポイントを正規化することで特定性と網羅性の両方を表す手法として，2.2.2 で WIDF[7] を説明した. これは，網羅性の関数を全体の値で正規化することにより，特定性も考慮していた.

WIDF のように特定性と網羅性の両方を考慮し，かつ，計算対象とする集合をプレゼンテーションとする手法を提案する. つまり，これは widf の計算をプレゼンテーションで行うことに対応している.



sfp: 共に3/3
wisfp: Aは1/12，Bは1/5

図 10　$I_{c.wisfp}$ の例

例として図 10 を考える. sfp ではキーワード A，B 共に 3 文書に出現しているため pf/N = 3/3 である. これは P1 のスライド S では A が 10 回出現していることを，考慮にいれていない. これに対して，wisfp では，キーワード A では 1/12，B は 1/3 となる．よって単一プレゼンテーション全体における網羅性の値で正規化することにより，文書毎の出現数の違いを区別することができる. これにより目的の検索シーンを探す時に，検索語の特徴が表れると予想する.

式は以下のように定義する.

$$I_{c.wisfp}(s,k,\theta,\delta,\varepsilon_1,\varepsilon_2) = \frac{Ic(s,k,\theta,\delta,\varepsilon_1,\varepsilon_2)}{\sum_{i\in P(s)} Ic(i,k,\theta,\delta,\varepsilon_1,\varepsilon_2)}$$

網羅性の関数に従来の適合度 I_c を用いることにより，UPRISE のシーンの検索に対応している，この手法はまた，プレゼンテーションにおける割合の計算のみであるため，検索コストが低いという利点がある.

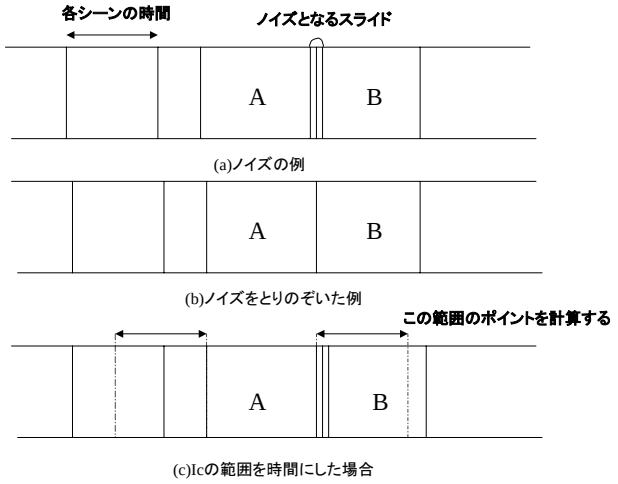


図 11　説明の無いシーンの例

この比較として，計算対象とする集合をすべてのプレゼンテーションで行う手法も定義する. これは，各シーンを文書と扱った widf[7] に対応している. これを，

$$I_{c.widf}(s,k,\theta,\delta,\varepsilon_1,\varepsilon_2) = \frac{Ic(s,k,\theta,\delta,\varepsilon_1,\varepsilon_2)}{\sum_{i\in allP} Ic(i,k,\theta,\delta,\varepsilon_1,\varepsilon_2)}$$

と定義する．これについても従来の適合度 I_c を用いることにより，UPRISE のシーンの検索に対応した. この方法はプレゼンテーション毎の違いは考慮できないが，全体での出現数の区別をすることができる. ただし，すべてのプレゼンテーションにおいて計算しなければならないため，検索コストが高い.

3.3　ノイズシーンの除去

3.1 において述べたように，前後関係のポイントに不適切な影響を与える説明の無いシーンが存在する．そして、このシーンを「ノイズシーン」と呼ぶことにした.

例として図 11(a) を考える．シーン A とシーン B は密接に繋がり合っている．そしてその間には 2 つのノイズシーンが存在している．δ = 2 の場合には既存の計算方法では A と B の関係を正しく計算できない．そこで以下の 2 手法を用いてノイズシーンを除去する.

3.4　ノイズシーンを除去する手法

この手法は図 11(b) のように，ノイズシーンを除去する方法である．実際には説明時間の閾値を設定し，それに満たないシーンを省いてポイントを与えることにより，実質的に話の流れから外れるシーンを除去する手法である．この閾値を大きくした場合，説明時間の長い重要なシーンを重点的に検索することができる．この手法により，同じスライドであっても別のスライドに挟まれている場合が多くなり，シーンの順序付けをより正確に行うことができる．その結果，ユーザの検索対象となりやすい重要なシーンを上位に表示することができると予想する．この手法の効果については実験で有効性を確認するが，適切な閾値の設定については今後の課題とする.

3.5　時間を軸とした範囲にする手法

ノイズシーンの影響の除去方法を，3.4 とは違った視点から捉えていく．そもそも問題は，I_c のポイントのパラメータである

δ に，ノイズシーンもカウントされていることである．図 11(c) のように，時間を軸とした範囲にすれば，ノイズシーンの時間は短いため，ノイズシーンの影響を最低限にすることができる．また，時間を軸とした範囲にすることにより，前後のシーンの時間の差まで考慮できるという利点もある．この手法により，時間の短いシーンは軽視され，一方，時間の長いシーンの影響が大きくなる，これにより，そのプレゼンテーションのテーマが, 検索に適合する確率が大きくなると予想する．この手法の評価実験は今後の課題とする．

4. 評 価 実 験

4.1 文書頻度を取り入れた改善の実験

I_c，$I_{c.wisfp}$，$I_{c.widf}$，$I_{c.isfp}$，$tf.wisfp$，$tf.isfp$ について次の条件で検索実験を行った．

- 検索は 10 人で実施した．
- 正解シーンは検索実施者による判断により決定した．
- 各適合度毎に 78 種類のキーワードを検索した．
- 格納プレゼンテーション数: 20
- 総シーン数: 849 (スライド枚数: 599)
- パラメタは $\theta = 0.5$，$\delta = 4$，$\varepsilon 1 = 5.0$，$\varepsilon 2 = 0.5$ で固定した．
- 格納されたコンテンツの全てのシーンに対して，適合度の種類に応じてポイント付けをした．
- 順序付けされたシーンを上から順番に表示した．
- 検索者は上から順番にシーンが正解に適合するか判断した．
- 最も適合すると思われるシーンを正解シーンとして，そのシーンが各適合度で表示された順位を記録した．

$tf.wisfp$，$tf.isfp$ は，3.2.1 と 3.2.2 の手法において，網羅性の関数として tf(キーワードの出現数) を用いたものであり，I_c の値を考慮していない．計算式は以下で定義される．

$$tf.wisfp(s,k) = \frac{tf(s,k)}{\sum_{i \in P} tf(i,k)}$$

$$tf.isfp(s,k) = tf(s,k) \cdot isfp(s,k)$$

ここで，$tf(s,k)$ はスライド s に出てくるキーワード k の個数である．

実験結果として，横軸に表示順位を取り，縦軸に件数を取ったグラフを，図 12， 13 に示す．このグラフは x 座標の小さい順位の件数が多いほうが精度の高い適合度のグラフであると言える．

図 14 と図 15 は，それぞれ図 12 と図 13 を積算したグラフである．横軸は正解シーン表示順位であり，縦軸はその順位までに表示された検索結果の件数である．このグラフはいずれも単調増加のグラフになっており，x 座標が小さい内に全検索結果数に達したグラフが精度の高い適合度であると言える．ただし，$tf.isfp$ と $tf.wisfp$ については，表示されないものが 5 件含まれていたため，合計が 36 にならない．しかし，これは I_c が，スライドの前後関係を考慮するため，検索語の一部しか含まれないシーンであっても，近傍に残りを含めば検索に適合するとしているために，正解シーンが表示されている．一方，この 5 件の正解シーンには検索語が全て含まれていないため，表示しないことが一概に不正解とは言えない．

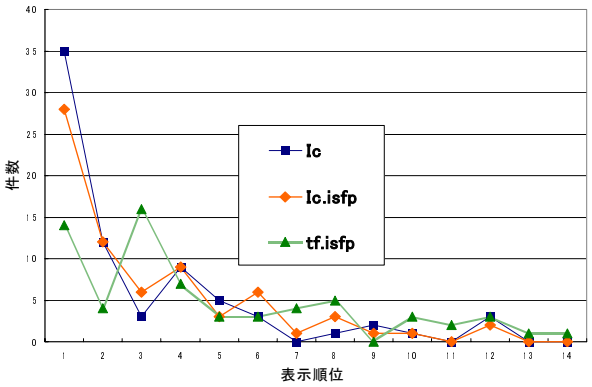


図 12　各適合度の正解シーン表示順位

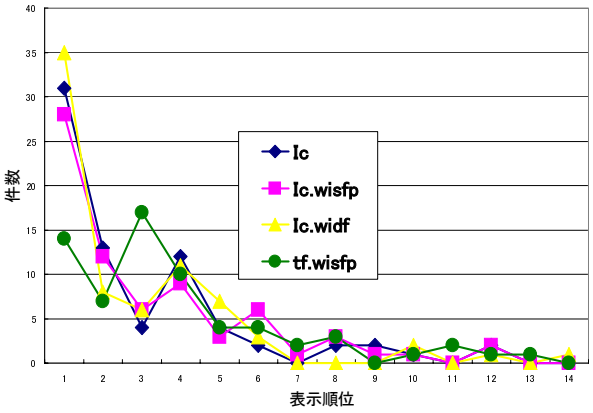


図 13　各適合度の正解シーン表示順位

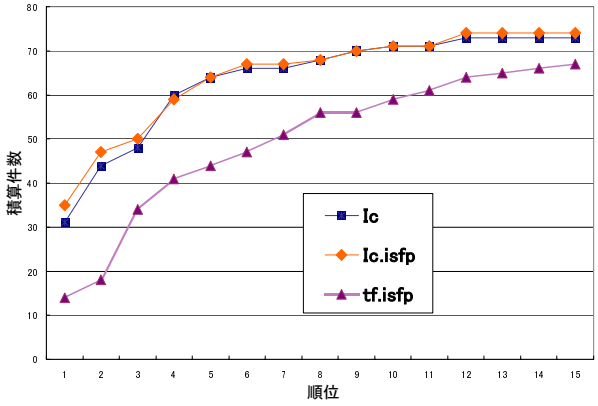


図 14　各適合度の順位と積算件数

また，本研究では，再現率と適合率を以下のように定義し，各手法ごとに計算を行った．

$$\text{再現率} = \frac{\text{検索で得られた正解シーン数}}{\text{全シーン中の正解シーン数}}$$

$$\text{適合率} = \frac{\text{検索で得られた正解シーン数}}{\text{検索の結果として得られたシーン数}}$$

この実験では正解シーンを各検索に対して 1 つとしているため，

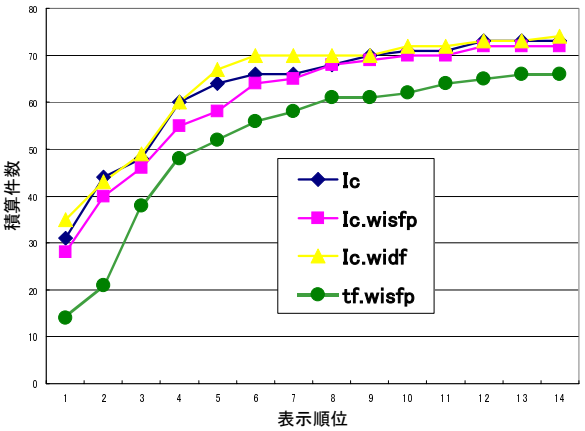


図 15　各適合度の順位と積算件数

再現率は 1.0 となる．また，適合率は以下の式で算出した．式中の n は総検索回数である．

$$\text{適合率} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{(i\text{ 番目の検索での表示順位})}$$

表 1　各手法での適合率

I_c	$I_{c.isfp}$	$I_{c.wisfp}$	$I_{c.widf}$	$tf.isfp$	$tf.wisfp$
0.564	0.598	0.526	0.593	0.364	0.392

表 1 は各手法での適合率である．$I_{c.isfp}$ は，I_c より上位に表示される数が多く，適合率でも 0.034 ポイント上回った．このことから従来の I_c にくらべて精度が向上したと言える．これは，複数キーワードの場合において，検索語として適しているキーワードをより重視した結果であると考える．

4.2　ノイズ対策実験

ノイズ対策の有効性評価実験を，4.1 と同様の条件で行った．ノイズ対策の適用対象として，従来の適合度 I_c と 4.1 において最も精度が高かった $I_{c.isfp}$ を用いた．今回の実験ではノイズ対策の閾値は 3 秒に設定した．この閾値の設定によっては，精度が低下，あるいは改善の可能性があるが，適切な閾値については今後の課題とする．

実験結果として，I_c とノイズシーン対策を加えた I_c の比較を図 16 と図 17 に，$I_{c.isfp}$ とノイズシーン対策を加えた $I_{c.isfp}$ の比較を図 18 と図 19 に示す．グラフについては 4.1 と同じ様式のグラフである．ノイズシーン対策を加えたものは，いずれも 5 件表示されないものが存在した，これらの正解シーンは共に，検索語を含んでいないか，一部しか含んでいなかったため，4.1 と同様の理由で不正解とは扱っていない，また，表 2 も，4.1 同様に算出している．

表 2　ノイズ対策前後の適合率の比較

I_c	I_c+ノイズ対策	$I_{c.isfp}$	$I_{c.isfp}$+ノイズ対策
0.564	0.594	0.598	0.601

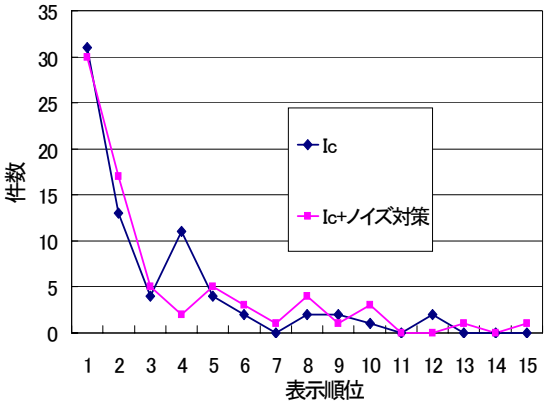


図 16　ノイズ対策を加えた適合度 Ic の表示順位別件数

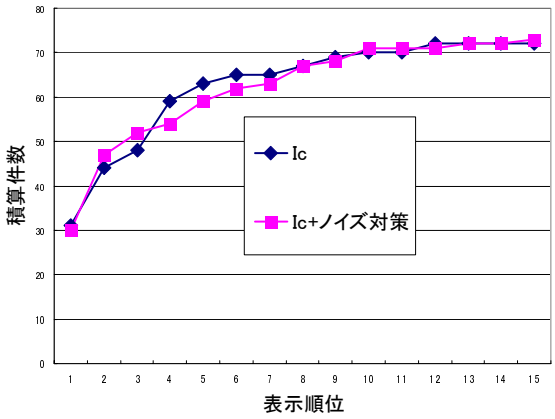


図 17　ノイズ対策を加えた適合度 Ic の表示順位別件数 (積算値)

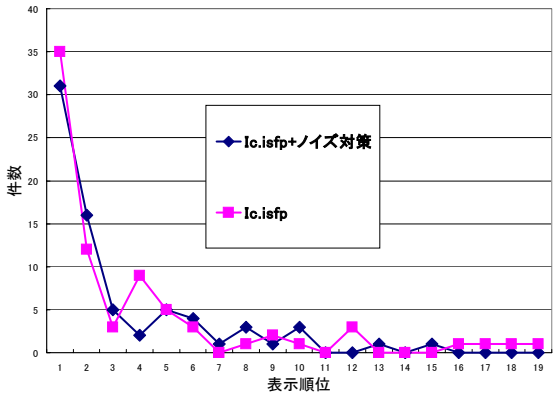


図 18　ノイズ対策を加えた適合度 $I_{c.isfp}$ の表示順位別件数

4.3　考　　察

まず 4.1 の実験について考察する．今回は特定性の関数に，log によって正規化したものを用いたが，この特定性の関数の強弱が精度に影響する．よって，この特定性の関数を検討することが，今後の課題としてあげられる．

$I_{c.wisfp}$ は，I_c より上位に表示される数が少なく，適合率でも 0.038 ポイント下回った．このことから単一プレゼンテーションにおいて計算した $I_{c.wisfp}$ では，従来の I_c より精度が下がってしまったと言える．しかし，全てのプレゼンテーションを範囲として計算した $I_{c.widf}$ では，再現率で 0.029 ポイント向上した．

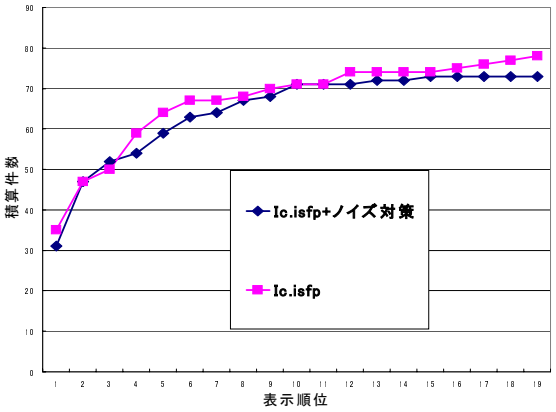


図 19　ノイズ対策を加えた適合度 $I_{c\cdot isfp}$ の表示順位別件数 (積算値)

この原因として，単一プレゼンテーションごとに計算すると，プレゼンテーション間のポイントの格差が現れてしまうことが挙げられる. 例えばあるキーワードがあるプレゼンテーションで 1 回ずつ 5 シーンに出現しており，別のプレゼンテーションでは 1 回 1 シーンのみ出現していたとする．この場合，後者のほうが 5 倍重視されることになる．これに対して全てのプレゼンテーションを範囲として計算した場合，プレゼンテーションごとの出現数には影響せずに，キーワードの特定性が考慮できる．しかし，全てのプレゼンテーションを範囲として計算すると，検索のコストが高い．よって $I_{c.wisfp}$ に，プレゼンテーション毎の傾向によってプレゼンテーション自体の評価を行い，そのポイントを反映することで，プレゼンテーションの出現格差を緩和する改善が考えられる．

図 16 と図 17 から，I_c にノイズ対策を加えることによって 1 位の表示件数が減少してはいるものの，ノイズ対策以前は 19 位まで分布していたものが，全て 15 位以内に検索されたことがわかる．そして適合度も表 2 からも，0.03 ポイント上昇している．このことから，1 位の表示件数は下がってしまったが，ノイズシーンを除去することによって，全体としては検索精度が向上したことがわかる．$I_{c.isfp}$ にノイズ対策を加えた場合についても，I_c の場合と同様な傾向が得られた．このことからノイズ対策は全体としての検索精度を向上させることが可能であることがわかった．

1 位の表示件数が減った理由としては，前後がノイズシーンであった 1 位のシーンを，前後を説明していた 2 位以下のシーンが逆転したことがあげられる．これはパラメータセットの改良によって改善の余地があると考える．また，適切な閾値を決定することにより，より精度の高い検索が可能になると考える．この点は今後の課題である．

5. まとめに

5.1 まとめ

本論文では，プレゼンテーション蓄積検索システム UPRISE において，従来の適合度 I_c に，プレゼンテーションにおける検索キーワードを含むシーンの頻度を特定性として加えた $I_{c.isfp}$ を提案し，従来の適合度 I_c とくらべて精度が向上することを実験によって確認した．

さらに適合度 I_c にシーンの頻度ではなくキーワードごとの適合度の割合による適合度である $I_{c.wisfp}$ と $I_{c.widf}$ を提案し，前者については改良点を考察し，後者については精度が向上したことを同様に実験によって確認した．

また，ノイズ対策として閾値を設定しノイズシーンを取り除く手法を提案し，I_c と $I_{c.isfp}$ に適用する実験を行い全体として精度が向上することを確認した．

5.2 今後の課題

今後の課題として，ポイントの割合を考慮した手法については，検索のコストを減らすために，プレゼンテーションごとの評価を導入し，より近似的な値を使用する方法があげられる．

また，プレゼンテーションにおけるシーン頻度を考慮した手法については，特定性の強弱を決める関数の検討があげられる．

ノイズ対策の手法に関しては，3.4 の手法における最適な閾値の検討，3.5 の手法については有効性の確認があげられる．

また，UPRISE においてユーザーが効率よく検索できるように，ユーザーの問い合わせの種類別に検索方法やパラメータ種類を検討することがあげられる. その例として, 目的のプレゼンテーションがわかっているユーザに対して, 対象のプレゼンテーションを検索してから，プレゼンテーションの中で検索するといった検索方法も考えられる．

謝　辞

本研究の一部は，文部科学省科学研究費補助金特定領域研究 (15017233)，独立行政法人科学技術振興機構 CREST，および 21 世紀 COE プログラム「大規模知識資源の体系化と活用基盤構築」の助成により行なわれた．

文　献

[1] 横田治夫. 東工大学術国際センターの情報蓄積・活用―教育コンテンツの統合とその手法―. 研究会報告 dbs-125-58, 情報処理学会, 2001.

[2] 村木太一, 吉田誠, 小林隆志, 直井聡, 横田治夫. メタデータによる講演資料と動画の統合と検索. In *Proc. of DBWeb2002*, pp. 97–104. 情報処理学会, 2002.

[3] 村木太一, 吉田誠, 小林隆志, 直井聡, 横田治夫. メタデータによる教育資料の統合における検索絞り込み指標の評価. Issn 1347-4413, DEWS2003, 5-c, 電子情報通信学会データ工学ワークショップ, 3 2003.

[4] Haruo Yokota, Takashi Kobayashi, Taichi Muraki, and Satoshi Naoi. UPRISE: Unified Presentation Slide Retrieval by Impression Search Engine. *IEICE Transactions on Information and Systems*, Vol. E87-D, No. 2, February 2004.

[5] 小林隆志, 村木太一, 直井聡, 横田治夫. 統合プレゼンテーションコンテンツ蓄積検索システムの試作. In *Proc. of DBWeb2003*, pp. 61–68. 情報処理学会, 11 2003.

[6] G. Salton. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley, 1988.

[7] 徳永健伸, 岩山真. 重み付き idf を用いた文書の自動分類について. 研究会報告 自然言語処理 No100-007, 情報処理学会, 1994.