単語の出現頻度を用いたドキュメントデータベースのメタデータ自動生成

河本 穣[†], 図子 泰三[††],
清木 康[‡]

[†]慶應義塾大学総合政策学部
[††]慶應義塾大学大学院政策・メディア研究科
[‡]慶應義塾大学環境情報学部

背景

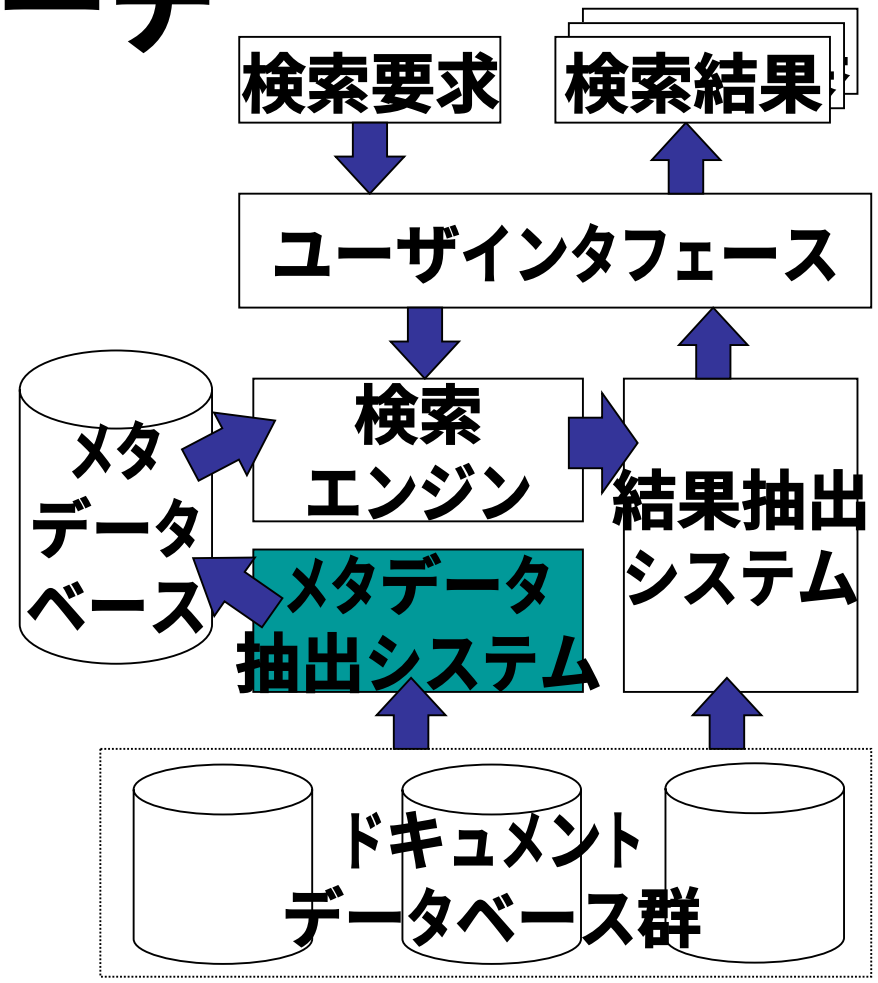
- ドキュメントを対象とする検索は、表記のみではなく意味に着目して行うべきである。
- ドキュメントは本来、検索者の視点(文脈)に応じて意味が確定する。
- 文脈依存型の情報検索の実現方式である、意味の数学モデル[1][2]による意味的連想検索が提案されている。

	形式的	意味的
静的な文脈	パターンマッチング	LSI
動的な文脈		意味の数学モデル

[1]清木 康,金子 昌史,北川 高嗣: 意味の数学モデルによる画像データベース探索方式とその学習機構,電子情報通信学会論文誌,D-II,Vol.J79-D-II,No. 4,pp. 509-519, 1996.
[2] Kiyoki, Y., Kitagawa, T. and Hayama, T.:
A metadatabase system for semantic image search by a mathematical model of meaning, ACM SIGMOD Record, vol. 23, no. 4, pp. 34-41, 1994.

アプローチ

- 大規模ドキュメントデータベースを対象として意味的な解釈が可能な検索を行うには、メタデータの生成が必要である。
- 本研究は、意味的連想検索を対象とした、単語の出現頻度を用いたメタデータ生成方式を実現する。



ドキュメント
の間接検索

意味的連想検索

--- 意味の数学モデル[1][2]---

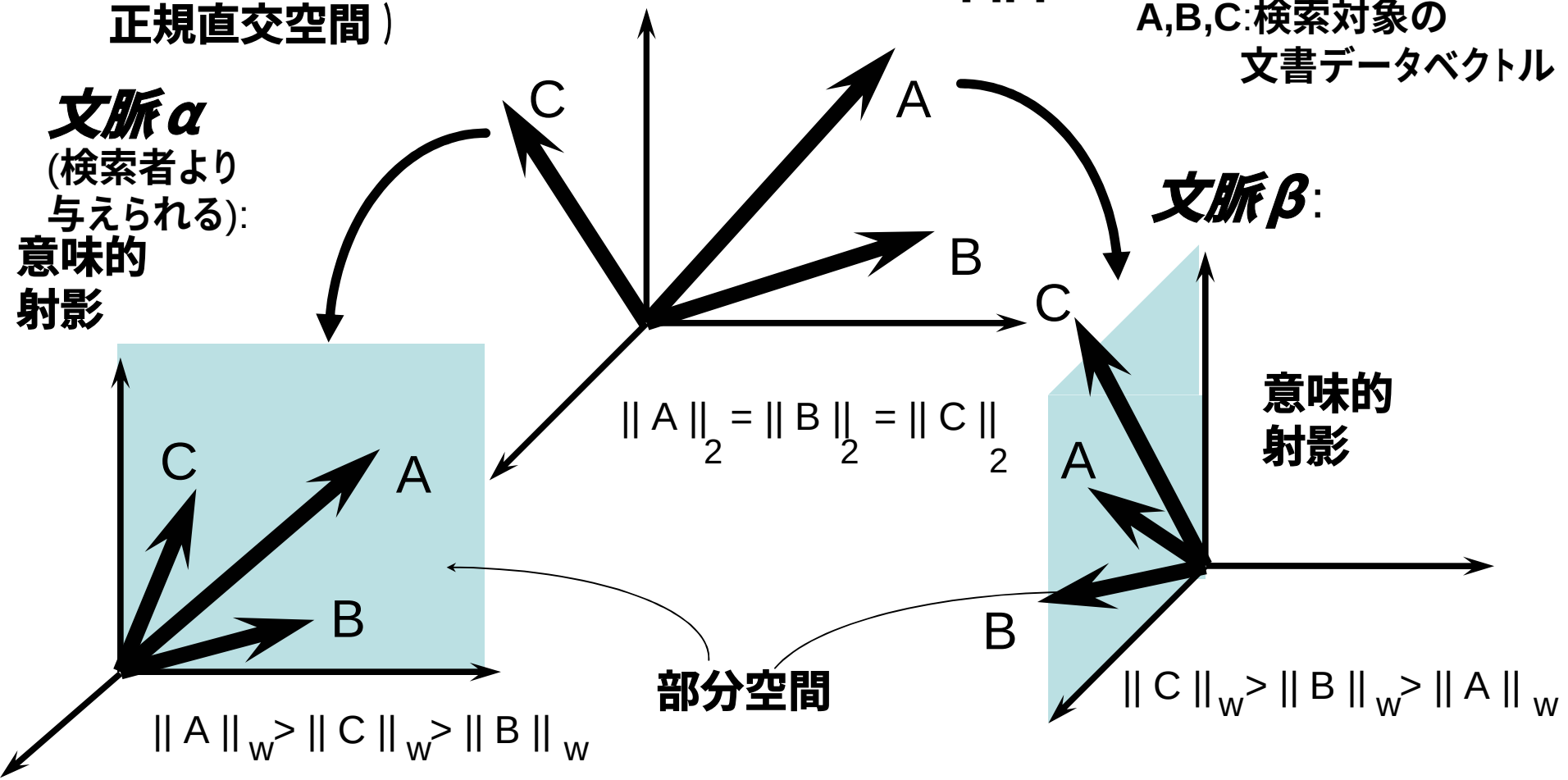
意味空間:
(270**次元の**
正規直交空間)

A,B,C:**検索対象の**
文書データベクトル

文脈α
(検索者より
与えられる):
意味的
射影

文脈β:

意味的
射影



[1]**清木 康**,**金子 昌史**,**北川 高嗣**: **意味の数学モデルによる画像データベース探索方式とその学習機構**,**電子情報通信学会論文誌**,D-II,Vol.J79-D-II,No. 4,pp. 509-519, 1996.

[2] Kiyoki, Y., Kitagawa, T. and Hayama, T.:
A metadatabase system for semantic image search by a mathematical model of meaning, ACM SIGMOD Record, vol. 23, no. 4, pp. 34-41, 1994.

提案方式の概要

従来方式

胃がん02: 痛み 出血 胃 粘膜 ポリープ 上皮 胃壁 筋肉 細胞
胃がん10: 胃 粘膜 腹 腫瘍 内臓 白血球 吐き気 脱毛 エイズ

		悪性しゅよう	遺伝病	胃がん	消化器	虫歯	
$\bigoplus_{i=1}^{t} \mathbf{o}_i$	痛み:	…	1	0	1	0	0 …
	出血:	…	1	0	0	0	1 …
	胃:	…	0	0	0	1	1 …
胃がん02:		…	1	0	1	1	1 …
胃がん10:		…	1	0	1	1	1 …

提案方式

胃がん02: 痛み 出血 胃 粘膜 ポリープ 上皮 胃壁 筋肉 細胞
胃がん10: 胃 粘膜 腹 腫瘍 内臓 白血球 吐き気 脱毛 エイズ

↓

提案方式Step1,Step2 → 胃がん02:ポリープ 粘膜 出血 上皮 胃
胃がん10 胃 粘膜 腫瘍 エイズ 内臓

提案方式Step3

		悪性しゅよう	遺伝病	胃がん	消化器	虫歯	
$\bigoplus_{i=1}^{t} \mathbf{o}_i$	胃がん02:	…	0	0	1	1	1 …
	胃がん10:	…	1	0	1	1	0 …

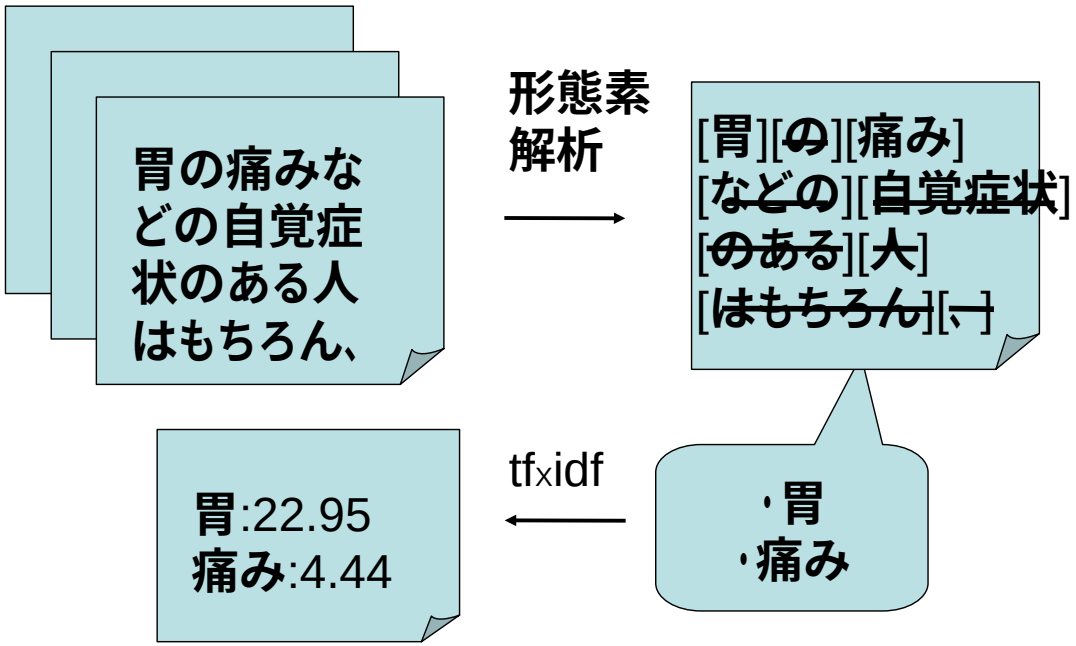
提案方式

対象とする文書群内の文書中の単語の出現頻度(tf×idf[3][4])**を用いて意味的連想検索に用いる意味的メタデータ群を選択的に抽出する。**

Step1

形態素解析した対象とするドキュメント群から、検索対象メタデータとして意味の定義されている単語のみを抜き出す。抜き出された単語群から tf×idf**の値を求める。**

$$TFIDF_{d,t} = FREQ_{d,t}(1+\log\frac{N}{DFREQ_t})$$



[3]Salton, G. and Buckley, C.: ``Term-weighting approaches in automatic text retrieval,'' Information Processing and Management, 24, pp.513-523,1988d.

[4]Salton, G. and Buckley, C.: ``Improving retrieval performance by relevance feedback,'' Journal of the American Society for Information Science, 41(4), pp.288-297,1990.

提案方式

Step2

下記の6種類のルールを用いて、ルールに適合したtf×idfの値をもつ単語のみを、メタデータを生成するための単語集合として選択する。

- [c] tf×idfの中央値を閾値とし、その値を超えた単語を選択する。
- [m] tf×idfの平均値を閾値とし、その値を超えた単語を選択する。
- [t_n] tf×idfの値nを閾値とし、その値を超えた単語を選択する。
- [u_n]ドキュメント内のtf×idfの値順で上位n件までの出現単語を選択する。
- [a] すべての出現単語を重みなしで採用する(従来方式A)。
- [w] すべての出現単語をtf×idfの値の重み付きで採用する(従来方式W)。

例:ドキュメントID メタデータ群(tfxidfの値)

がんのQOL05 [w]骨(15.1) 全身(5.4) 腰椎(5.1) 胃(4.1) がん(4.1)
腰(3.6)首(3.1 元気(3.1) 声(3.1) 頭(2.6) 胸(2.5)

[a]骨,全身,腰椎,胃,がん,腰,首,元気,声,頭, 胸

[u_5]骨,全身,腰椎,胃,がん

提案方式

Step3

オペレータ $\bigoplus_{i=1}^{t} \mathbf{o}_i$ を用いて単語群からベクトルを生成し、意味の数学モデルに適用する検索対象メタデータとする

ドキュメント:胃がん01 のメタデータ:胃 苦痛 痛み ポリープ

	胃	胃かいよう	子宮がん	疾病	虫歯	
…	1	1	0	1	0	…
…	0	0	0	1	0	…
…	0	0	1	1	0	…
…	0	0	1	1	0	…

胃
苦痛
痛み
ポリープ

オペレータ $\bigoplus_{i=1}^{t} \mathbf{o}_i$ によるドキュメントのベクトル表現

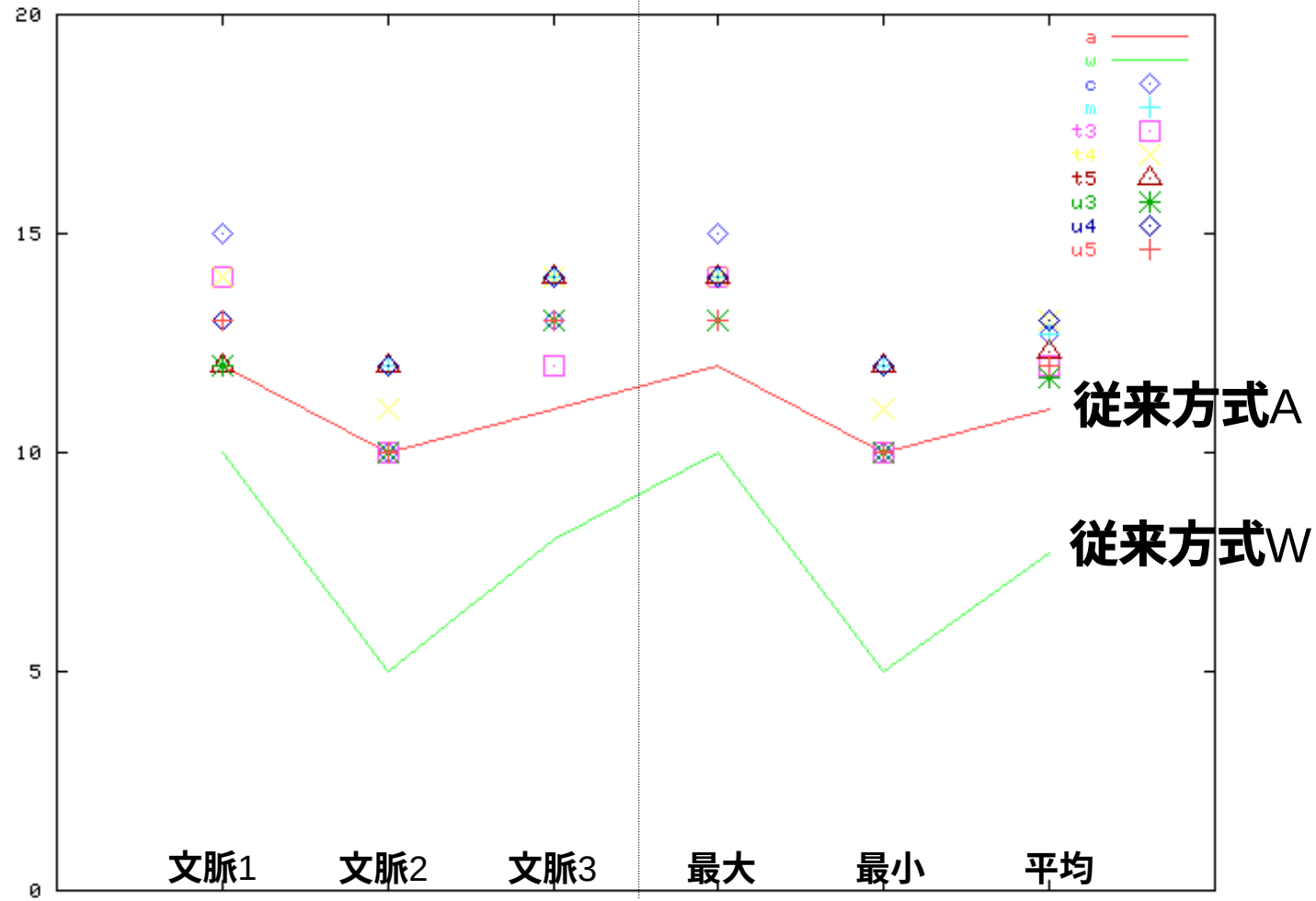
	胃	胃かいよう	子宮がん	疾病	虫歯	
…	1	1	0	1	0	…
…	1	0	1	1	0	…
…	1	1	0	1	0	…

胃がん01
胃がん02
胃がん03

実験

- 実験の目的
  - 提案方式の有効性、実現可能性の検証
- 実験対象
  - 医療分野の95件の新聞記事を対象。
- 評価方法
  - 医療分野の新聞記事編集者により生成されたメタデータを用いた場合の20件の検索結果と比較する。
  - 両者の一致件数を検索性能の尺度する。

実験の結果



実験に用いた3つの文脈において従来の単語抽出方法(従来方式A,従来方式W)と比較して、性能が向上していることが確認できた。

実験の考察

- 意味的連想検索の検索対象メタデータとして用いた場合では、実験に用いた3つの文脈において従来の単語抽出方法(従来方式A,従来方式W)と比較して、性能が向上していることが確認できた。

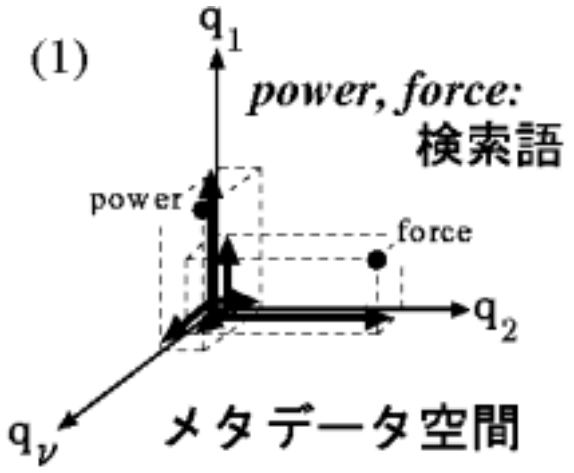
まとめと今後の課題

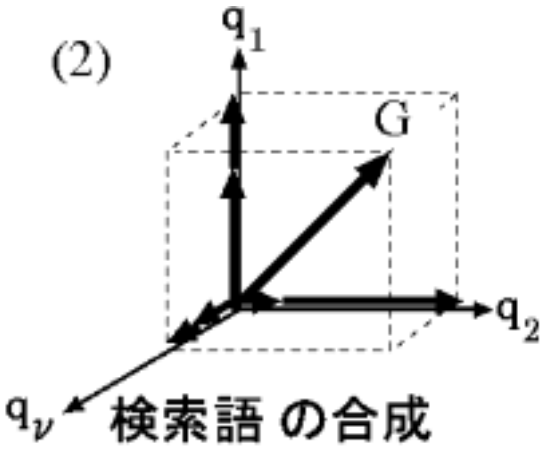
- 本方式により、意味的連想検索に用いる意味的メタデータの自動生成が行えることが確認できた。
- 提案方式は、意味的メタデータを自動的に客観的な基準で生成できることにより、大規模ドキュメントデータベースへ適用可能である。

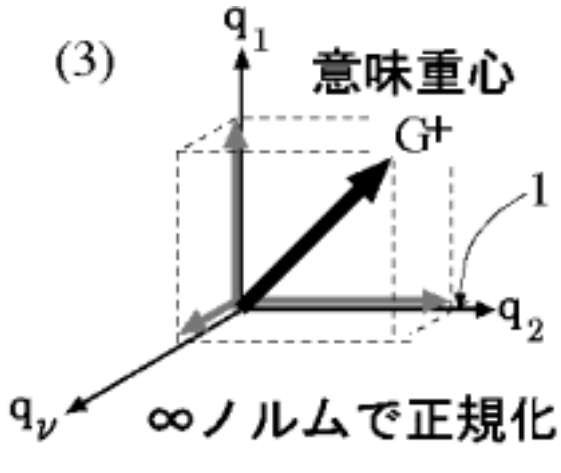
今後の課題

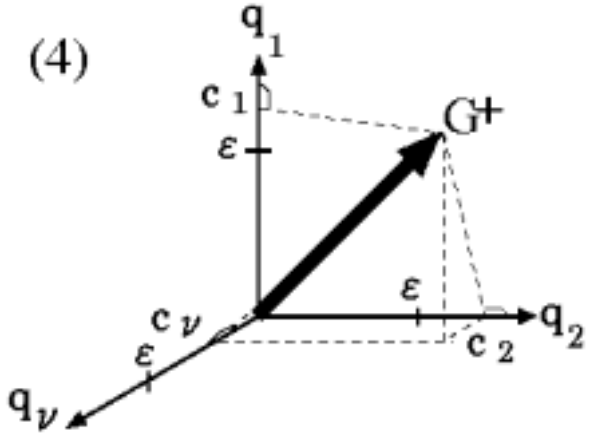
1. 大規模ドキュメントデータベースへの適用可能な方式の提案と実装
2. 提案したメタデータ選択方式相互間の優劣の検証と動的選択方式

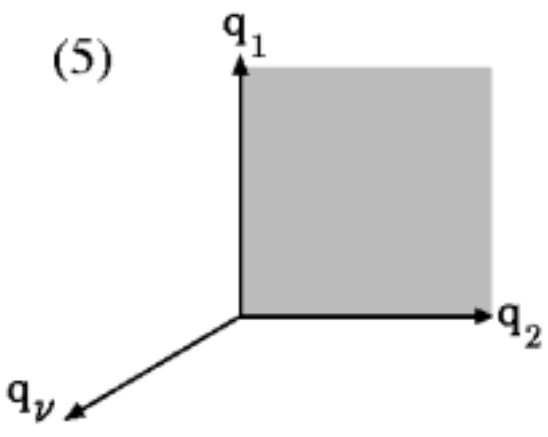
意味的連想検索(メディア検索)の概要

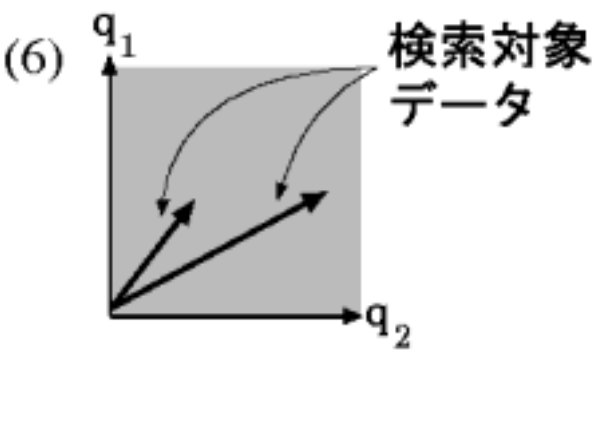






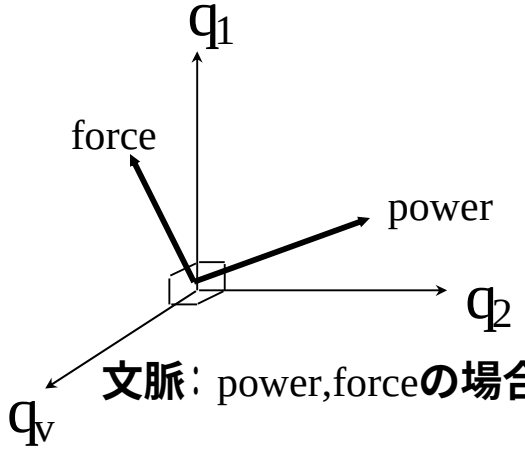






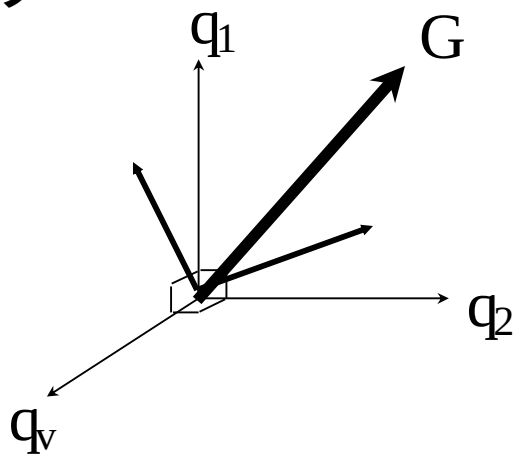
意味的連想処理機構における部分空間選択

(1) メタデータ空間への文脈語群 (検索語) のマッピング

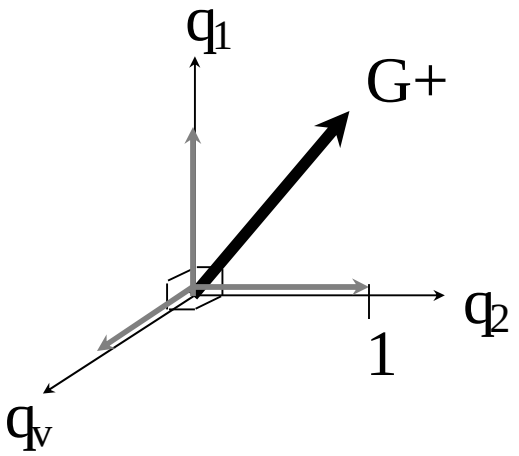


文脈：power,forceの場合

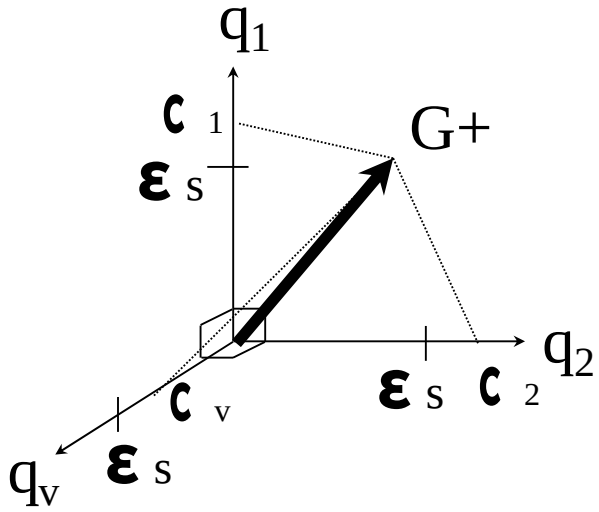
(2) 検索語の合成



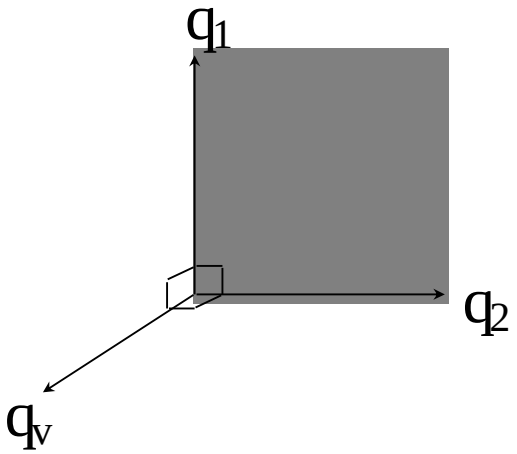
(3) ∞ノルムによる正規化



(4) 意味重心の各軸へ射影



(5) 部分空間の選択



G+:意味重心
(文脈語群ベクトル)
ε_s: 閾値
c_j: 重み $(1 \leqq j \leqq v)$

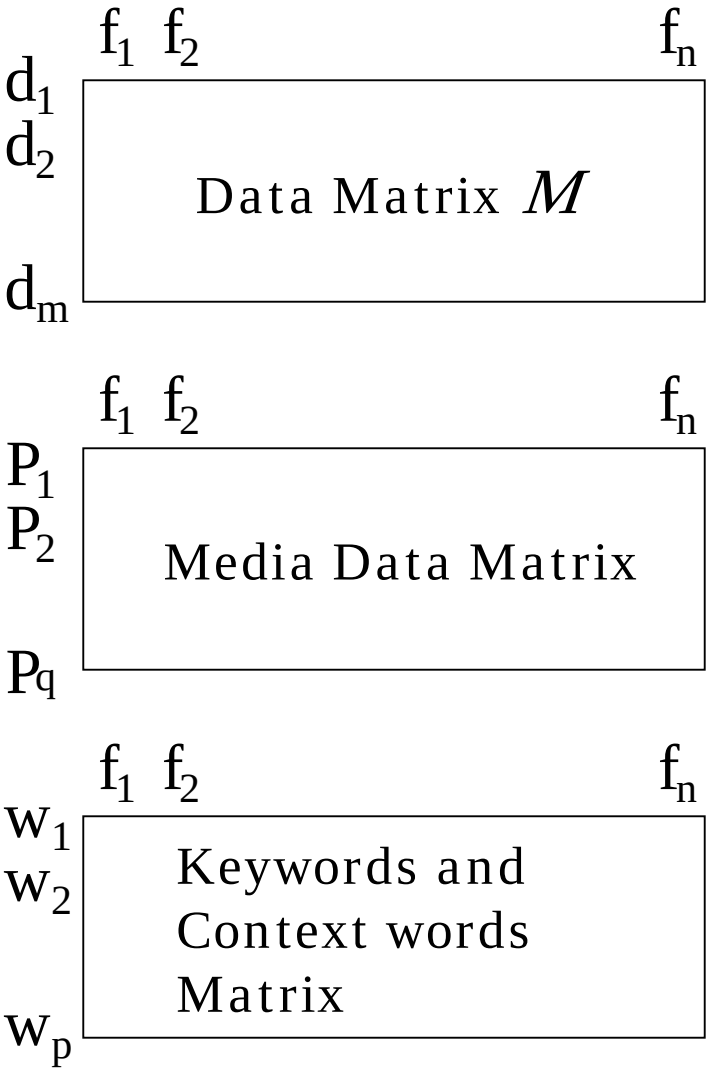
意味重心 G+ を各軸に射影し，その座標が閾値 ε_s を超えた軸のみを選択

意味の数学モデルに用いるメタデータ(メディア検索用)

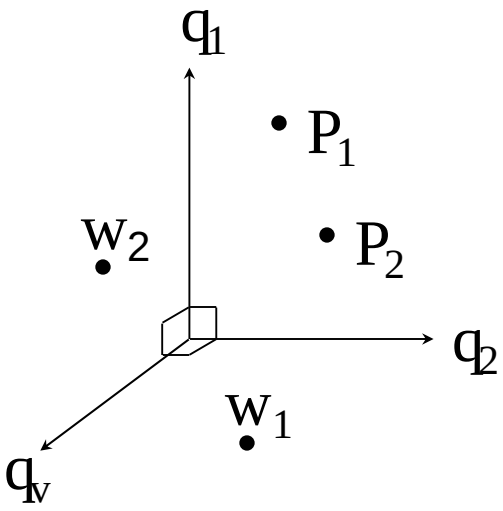
(a)空間生成のためのメタデータ

(b) メディアデータのためのメタデータ
(検索対象データ)

(c) キーワードのためのメタデータ
(検索者より与えられる文脈語群)



$M^T M$ の固有値分解



(b) メディアデータのためのメタデータ

- **Step1: メディアデータ** P1 **のメタデータ**: vivid quiet

vivid	*adj*. bright , strong
quiet	*adj*. without or with little -sound , -motion

	...	bright	form	material	motion	speed	sound	strong	...
vivid	...	1	0	0	0	0	0	1	...
quiet	...	0	0	0	-1	0	-1	0	...

- **Step2: オペレータ** $\bigoplus_{i=1}^{t} o_i$ **によるメディアデータ** P1 **のベクトル表現**

	...	bright	form	material	motion	speed	sound	strong	...
P1	...	1	0	0	-1	0	-1	1	...

目的(2/2)

- 意味的メタデータの生成はオーバーヘッドが大きく、大規模なドキュメントデータベース上の適用が困難である。
- 統一的な基準でメタデータ生成を行うことで、一貫性の高いメタデータが生成可能であり、より質の高い検索結果を得ることができるようになる.

目的(3/3)

- 意味に応じた検索のための意味的メタデータの生成
- メタデータ数を必要最小限の数にとどめる必要性
- 自動化による大規模データベースへの応用
- 統一的な基準による一貫性

↓

意味的連想検索に適した、必要最低限で、かつ十分な意味的メタデータの自動生成が有用

マルチメディアデータの検索方式[2/2]

(メタデータの表現形式の現状)

以下のようなメタデータの表現形式に関する諸提案がある。
本研究では，メタデータとして単語の集合を用いた検索手法を前提とする．
これにより，抽象度が高くかつ一般性の高い検索を実現することが可能である．

- MPEG7
- Dublin Core
- RDF(Resource Description Framework)

マルチメディアデータの検索方式[1/2]

- **検索要求と検索対象となるマルチメディアメディアデータの類似度の計算方式** (overview)
  - **検索要求と検索対象マルチメディアデータをベクトル化し，類似度を計算**
    - **ベクトル間の距離,ベクトル間の内積,ベクトルのノルム** etc.
  - **シソーラス,ファジー理論,ニューラルネットワーク**
  - LSI (Latent Semantic Index)
    - **相関量計量によって文書の持っている意味を行列表現によって数値化した**
  - **意味の数学モデル**[1][2]
    - **直交空間における部分空間選択の演算を定義し，この演算により言葉と言葉の意味的な類似性を検索者の与える文脈に応じて動的に計算することが可能である．文脈の概念を導入することにより，言葉の意味の曖昧性を排除している．**

[1]**清木 康,金子 昌史,北川 高嗣**: **意味の数学モデルによる画像データベース探索方式とその学習機構,電子情報通信学会論文誌**,D-II,Vol.J79-D-II,No. 4,pp. 509-519, 1996.

[2]Kiyoki, Y., Kitagawa, T. and Hayama, T.:
A metadatabase system for semantic image search by a mathematical model of meaning, ACM SIGMOD Record, vol. 23, no. 4, pp. 34-41, 1994.

実験2

- 提案方式によって得られたメタデータ群を用いた完全一致によるパターンマッチングの検索を行う。
- 人手により生成されたメタデータを用いて検索を行った結果を正解集合とした、再現率及び適合率を求め、検索性能の尺度とした。

実験2の結果

	熱		高血圧		ストレス		肥満		疲労		糖尿病	
	適合	再現	適合	再現	適a合	再現	適合	再現	適合	再現	適合	再現
a,w	0.46	1	0.44	0.8	0.47	0.82	0.5	0.71	0.47	1	0.58	1
c	0.57	0.89	0.25	0.2	0.58	0.64	0.6	0.43	0.55	0.75	0.94	0.89
m	0.89	0.89	1	0.2	1	0.36	0.5	0.29	0.5	0.71	1	0.78
t3	0.46	1	0.44	0.8	0.58	0.64	0.5	0.71	0.58	0.88	0.89	0.89
t4	0.64	1	0.5	0.2	0.58	0.64	0.6	0.43	0.58	0.88	0.89	0.89
t5	0.73	0.89	0.5	0.2	0.58	0.64	0.6	0.43	0.58	0.88	1	0.78
u3	0.8	0.67	0.33	0.2	1	0.18	0.8	0.57	0.5	0.57	1	0.83
u4	0.86	0.86	0.25	0.2	1	0.36	0.8	0.57	0.55	0.86	0.94	0.83
u5	0.88	0.88	0.25	0.2	0.88	0.64	0.8	0.57	0.55	0.86	0.94	0.83

実験2の考察

- 部分一致によるパターンマッチングのインデクスとして用いた場合では、性能の向上がある検索語と、そうでない検索語が存在している。

実験環境

- **実験環境**
  - Sun Microsystems Enterprise 3500, Solaris 8
  - Perl**によるメタデータ抽出システム**
  - **意味的連想検索システム**
  - Perl**による部分一致検索プログラム**
  - **医療ドキュメント意味空間**[3]

アプローチ

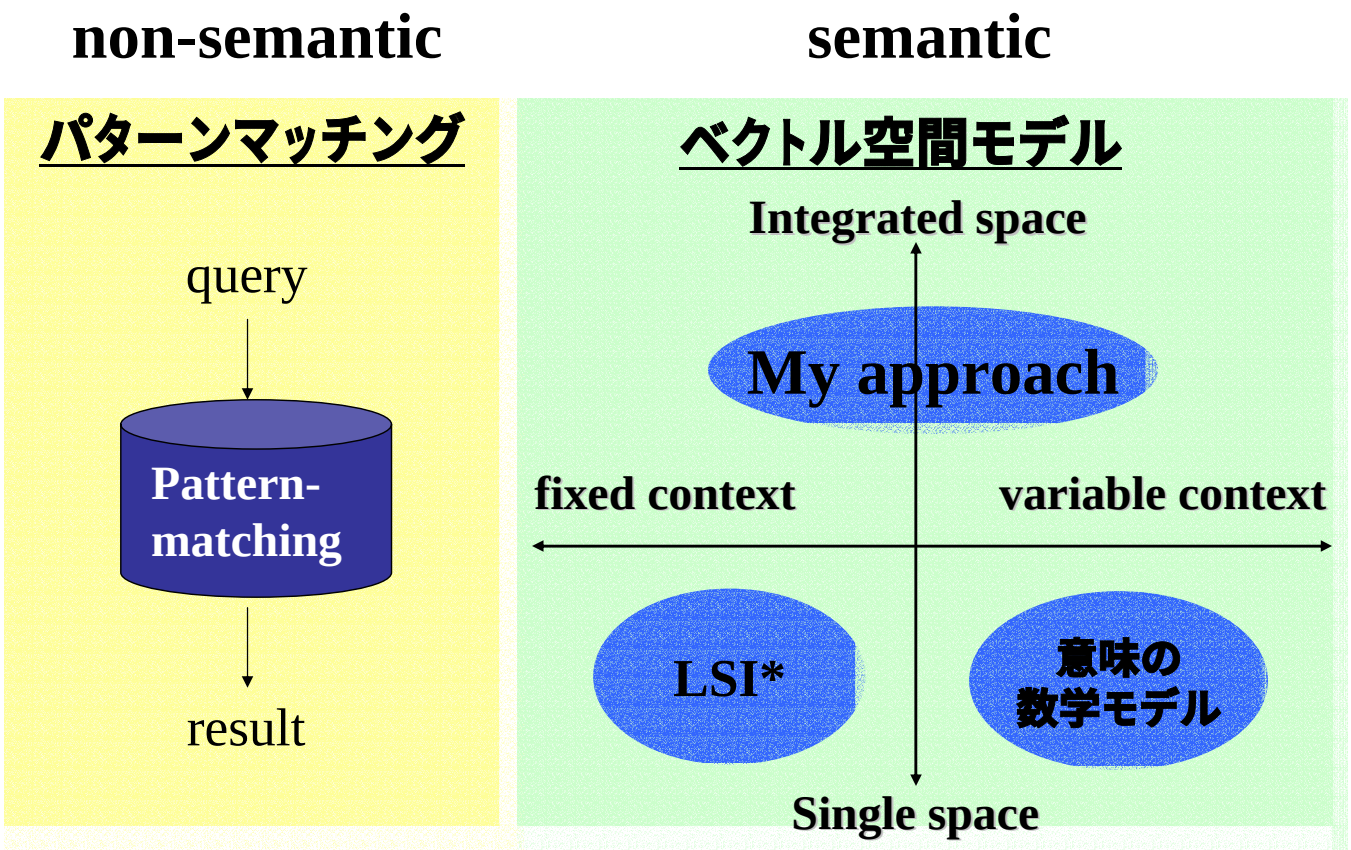
目的：意味の数学モデルでは、検索対象メディアデータのベクトルは2ノルムで正規化されるため、過剰な数のメタデータはドキュメントデータの持っている意味の特徴を弱める傾向がある。

研究の背景

ネットワーク上の膨大なデータ群から必要な情報を獲得するため，様々な検索手法が提案されている．

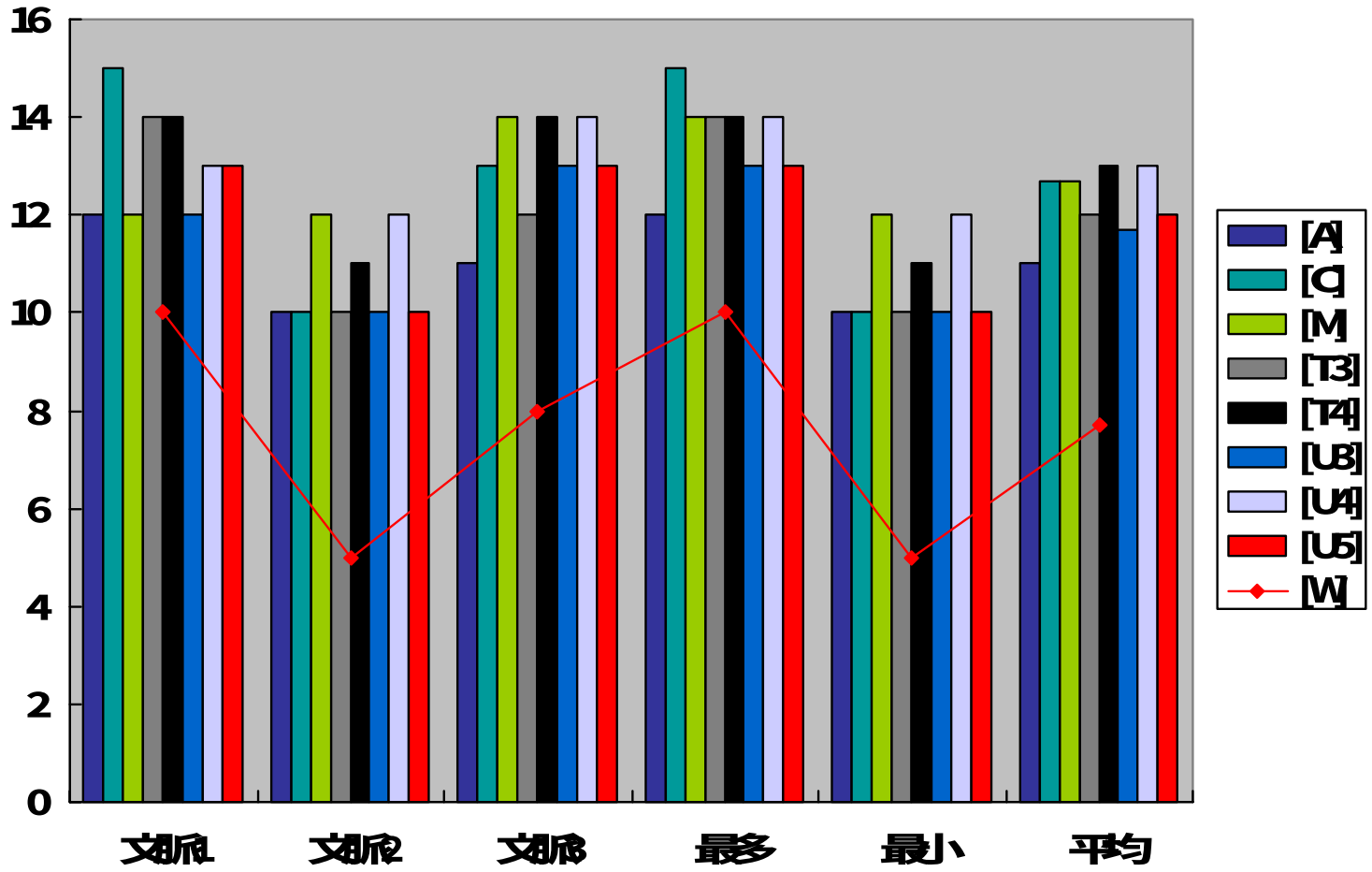
－パターンマッチング検索の限界

－semanticを扱うことが可能なベクトル空間モデル



*LSI：
Latent Semantic Indexing

実験の結果(Excel)



提案方式の概要

- **従来方式**

胃がん02 痛み 出血 胃 粘膜 ポリープ 上皮 胃 壁 筋肉 細胞

胃がん10 胃 粘膜 腹 腫瘍 内臓 白血球 吐き気 脱毛 エイズ

- **提案方式**

胃がん01:胃 苦痛 痛み ポリープ

胃がん02:ポリープ 粘膜 出血 上皮 胃

	胃	胃かいよう	子宮がん	疾病	虫歯
…	1	1	0	1	0
…	0	0	0	1	0
…	0	0	1	1	0
…	0	0	1	1	0

胃
苦痛
痛み
ポリープ

t

オペレータ $\bigoplus_{i=1} \mathbf{o}_i$ **によるドキュメントのベクトル表現**

	胃	胃かいよう	子宮がん	疾病	虫歯
…	1	1	1	1	0

胃がん01

提案方式の概要

- **従来方式**

胃がん02: 痛み 出血 胃 粘膜 ポリープ 上皮 胃壁 筋肉 細胞

胃がん10: 胃 粘膜 腹 腫瘍 内臓 白血球 吐き気 脱毛 エイズ

胃壁: 胃 胃かいよう 胃がん 胃腸薬

胃: 胃 胃かいよう 胃がん 胃腸薬 消化器 消化器疾患

腫瘍: 悪性しゅよう 胃がん 子宮がん 食道がん

悪性しゅよう 胃かいよう 胃がん 消化器 食道がん

$\bigoplus_{i=1}^{t} \mathbf{o}_i$						
胃がん02:	1	1	1	1	1
胃がん10:	1	1	1	1	1

提案方式 Step1 Step2（後述）

- **提案方式**

胃がん02:ポリープ 粘膜 出血 上皮 胃

胃がん10 胃 粘膜 腫瘍 エイズ 内臓

提案方式Step3

悪性しゅよう 胃かいよう 胃がん 消化器 食道がん

$\bigoplus_{i=1}^{t} \mathbf{o}_i$						
胃がん02:	0	1	1	1	0
胃がん10:	1	1	1	1	1