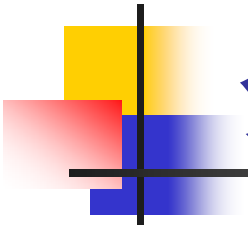
リンク先ページの内容を反映させたWebページの特徴ベクトル改良法

奈良先端科学技術大学院大学
情報科学研究科
杉山一成　波多野賢治　吉川正俊　植村俊亮



発表の手順

- 本研究の目的
- 提案手法
- 評価実験･結果
- まとめ
- 今後の課題

本研究の目的

- Webページの検索
  - 検索語の内容に関係のないページが多数提示される

➡ より内容を考慮した索引付けが望まれる

- Webページの特徴
  - ページ間にハイパーリンクによる内容的な結びつきがある

➡ 注目ページの近傍のページを含めた索引付けをする必要がある

本研究の目的（1）

- **これまでの研究**(DBWeb2001)
  - **リンク先ページの内容を考慮した索引付け手法を提案**（tf-idf**法を改良**）

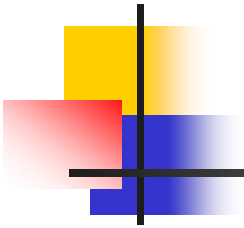
問題点

- **リンク先ページの扱い方について、検討不足**
- **離れたリンク先に類似ページがある場合には、そのページの内容が反映されにくい**
- **何リンク先のページまで扱えばよいか、検討不足**

本研究の目的 (2)

- リンク先ページの扱い方について、3つの特徴ベクトル改良手法を提案

- 各手法について、次の事項を検証
  - リンク先ページを考慮したことによるキーワード抽出精度
  - 改良特徴ベクトルを用いた検索精度の比較と、何リンク先までたどればよいか



3つの提案手法

ある注目しているWebページに対して、

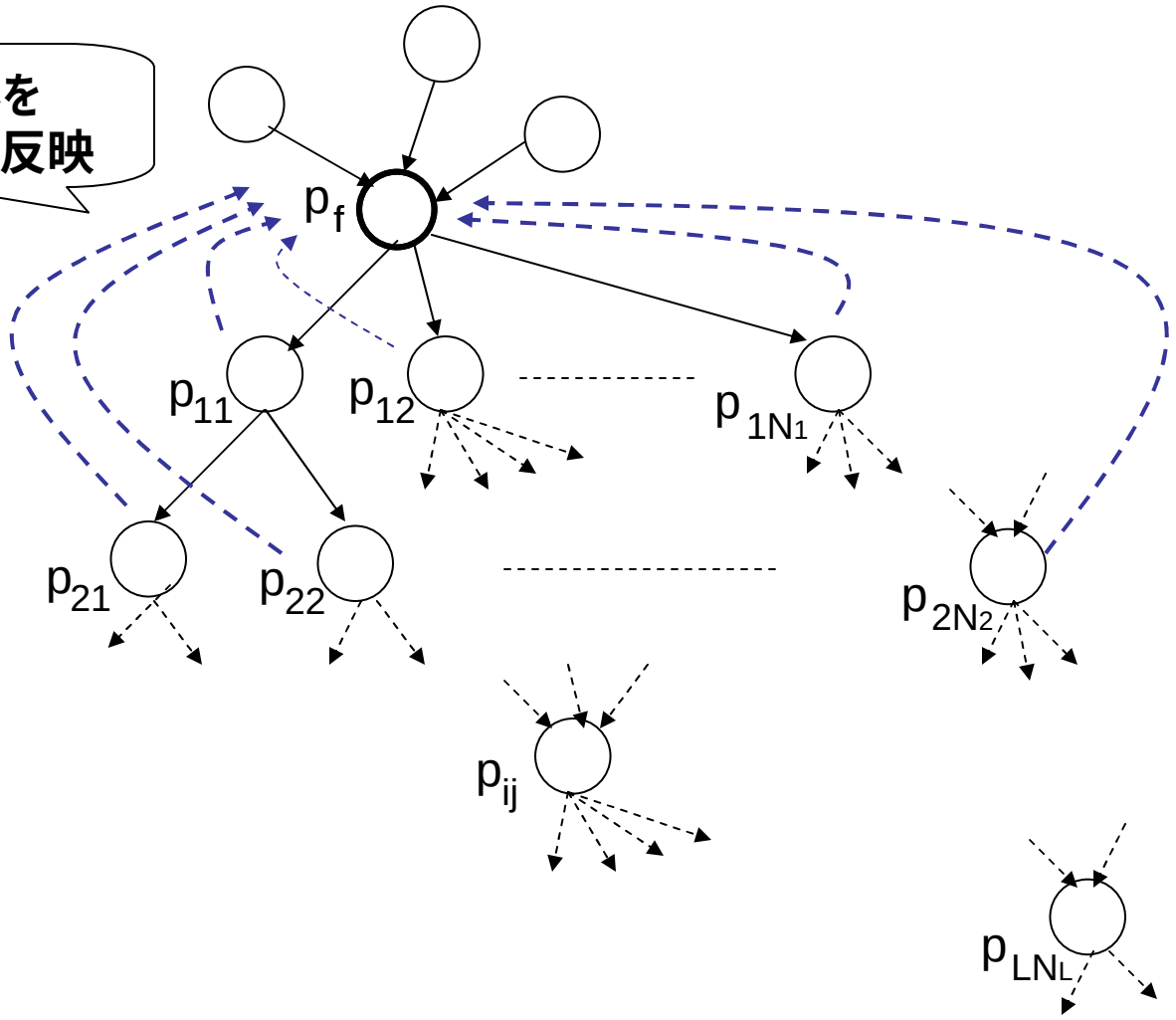
(1) 個々のリンク先ページの内容を反映させる

(2) リンク先ページから構成される
クラスタの重心ベクトルを反映させる

(3) リンク先の各階層のWebページから構成される
クラスタの重心ベクトルを反映させる

提案手法 (1)

各ページの内容を
参照元ページに反映



各単語の重み付け (手法1)

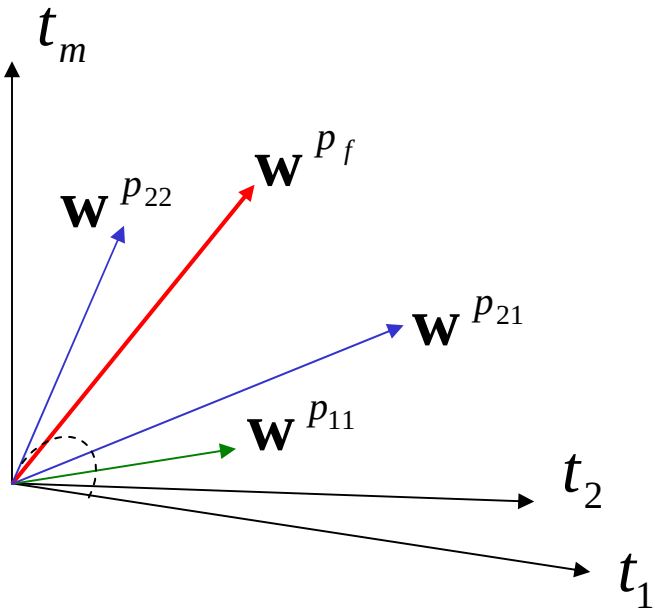
注目ページ p_f の各単語 t_k の重み

$$w'^{p_f}_{t_k} = w^{p_f}_{t_k} + \sum_{i=1}^{L} \sum_{j=1}^{N_i} \frac{1}{dis(\mathbf{w}^{p_f}, \mathbf{w}^{p_{ij}})} w^{p_{ij}}_{t_k}$$

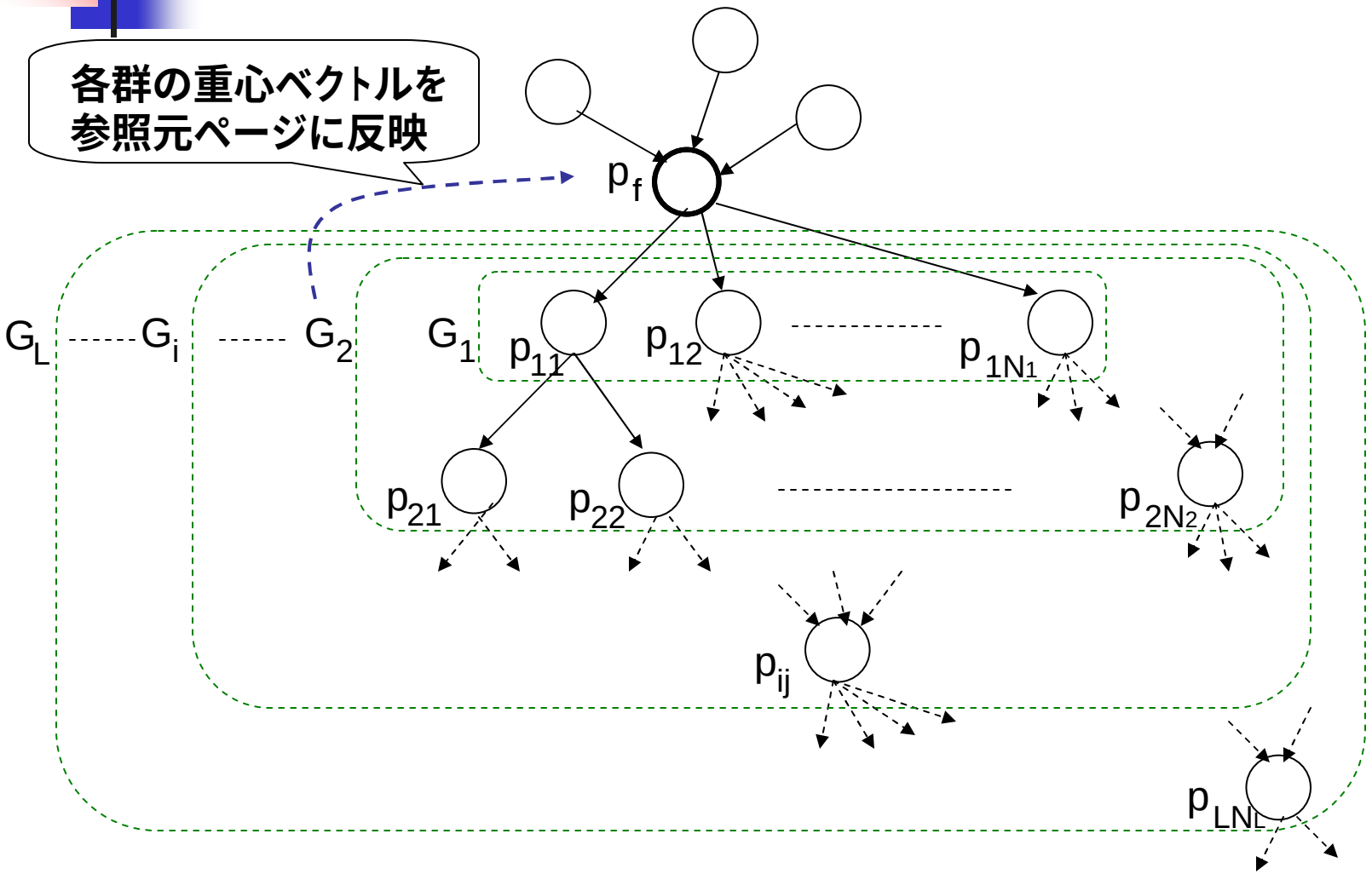
ただし、

$w^{p_f}_{t_k}$ ： tf-idf**によって求められた重み**

$$dis(\mathbf{w}^{p_f}, \mathbf{w}^{p_{ij}}) = \sqrt{\sum_{k=1}^{m} (w^{p_f}_{t_k} - w^{p_{ij}}_{t_k})^2}$$



提案手法 (2)


各群の重心ベクトルを
参照元ページに反映
p_f
G_L
G_i
G_2
G_1
p_{11}
p_{12}
p_{1N_1}
p_{21}
p_{22}
p_{2N_2}
p_{ij}
p_{LN_L}

各単語の重み付け (手法2)

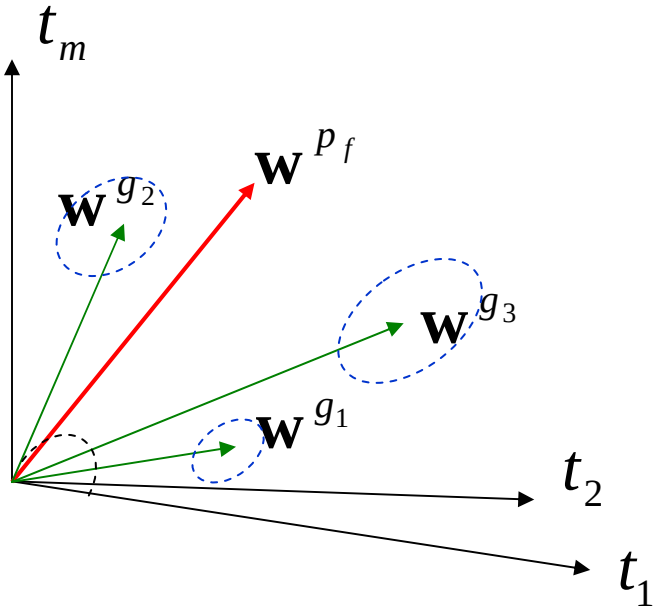
注目ページ p_f の各単語 t_k の重み

$$w'^{p_f}_{t_k} = w^{p_f}_{t_k} + \sum_{c=1}^{K} \frac{1}{dis(\mathbf{w}^{p_f}, \mathbf{w}^{g_c})} w^{g_c}_{t_k}$$

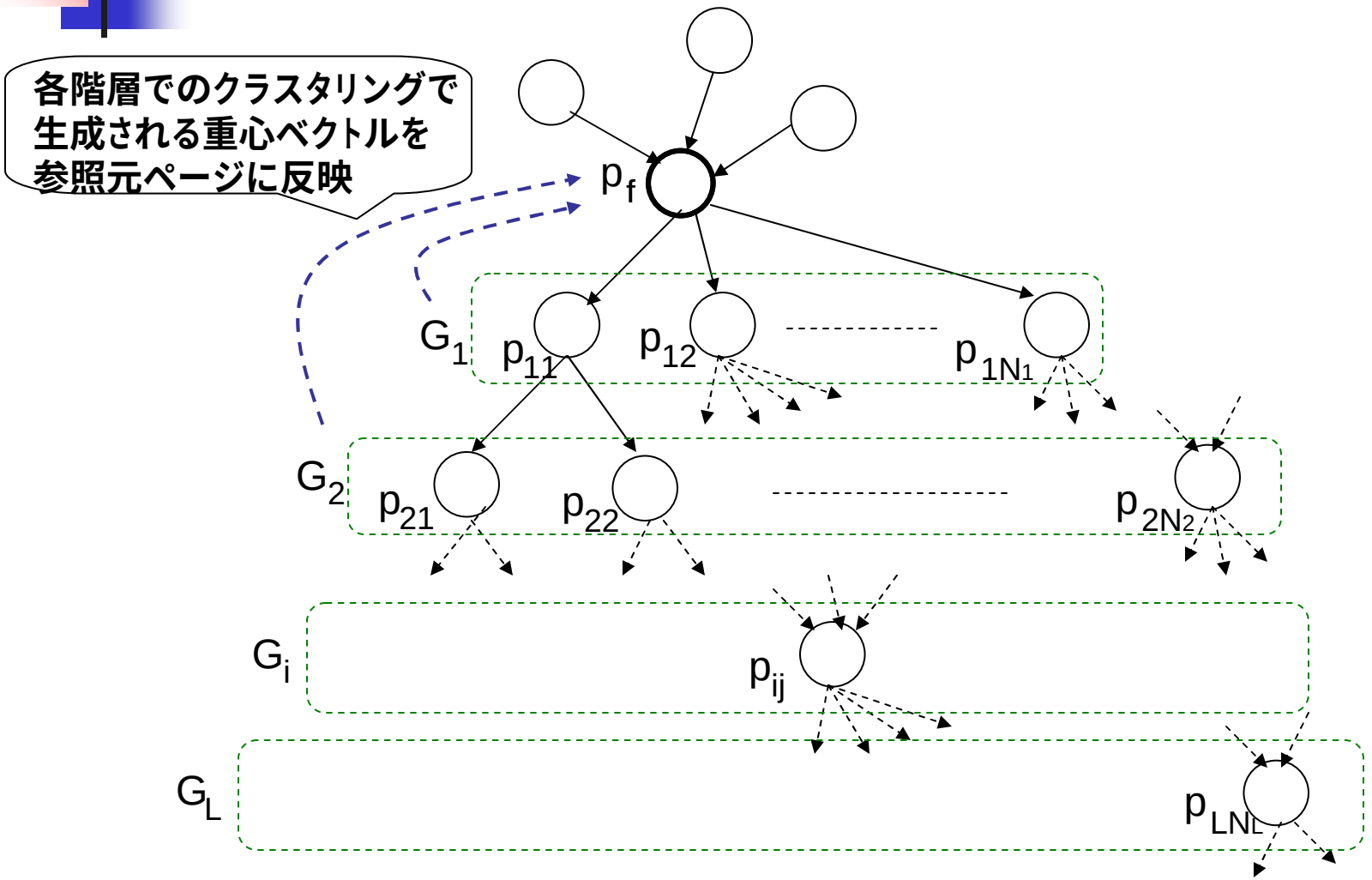
ただし、

$w^{p_f}_{t_k}$ ： tf-idf**によって求められた重み**

$$dis(\mathbf{w}^{p_f}, \mathbf{w}^{g_c}) = \sqrt{\sum_{k=1}^{m} (w^{p_f}_{t_k} - w^{g_c}_{t_k})^2}$$



提案手法 (3)


各階層でのクラスタリングで
生成される重心ベクトルを
参照元ページに反映
p_f
G_1
p_{11}
p_{12}
p_{1N_1}
G_2
p_{21}
p_{22}
p_{2N_2}
G_i
p_{ij}
G_L
p_{LN_L}

各単語の重み付け (手法3)

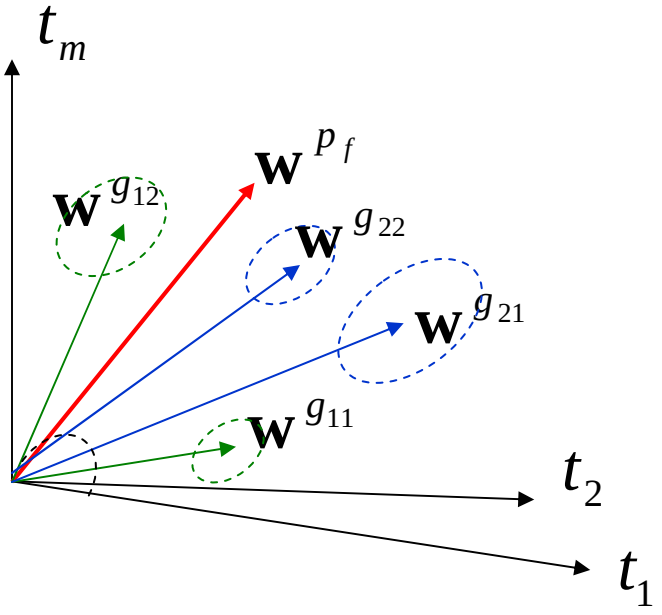
注目ページ p_f の各単語 t_k の重み

$$w'^{p_f}_{t_k} = w^{p_f}_{t_k} + \sum_{i=1}^{L} \sum_{c=1}^{K} \frac{1}{dis(\mathbf{w}^{p_f}, \mathbf{w}^{g_{ic}})} w^{g_{ic}}_{t_k}$$

ただし、

$w^{p_f}_{t_k}$ ： tf-idf**によって求められた重み**

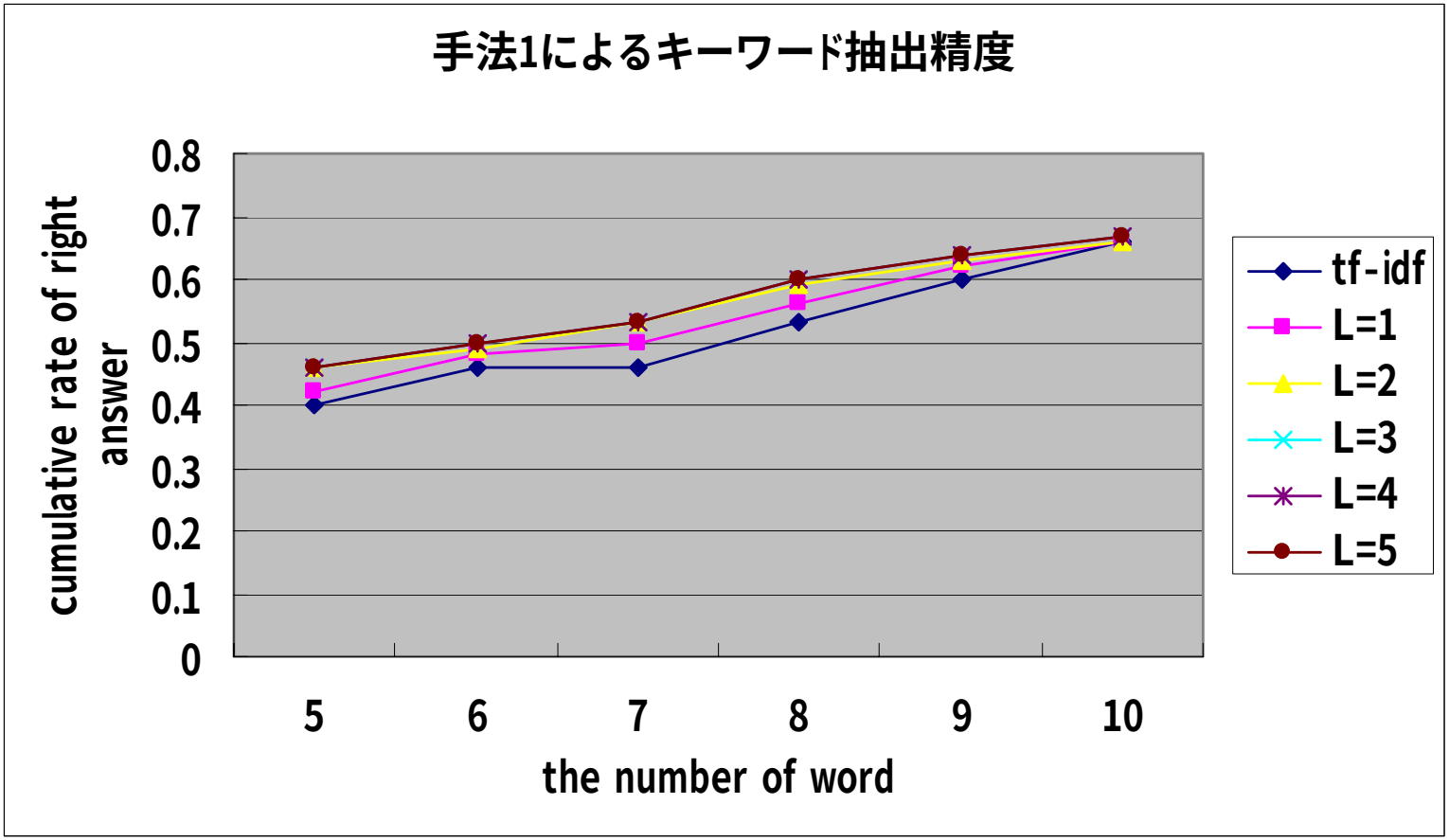
$$dis(\mathbf{w}^{p_f}, \mathbf{w}^{g_{ic}}) = \sqrt{\sum_{k=1}^{m} (w^{p_f}_{t_k} - w^{g_{ic}}_{t_k})^2}$$



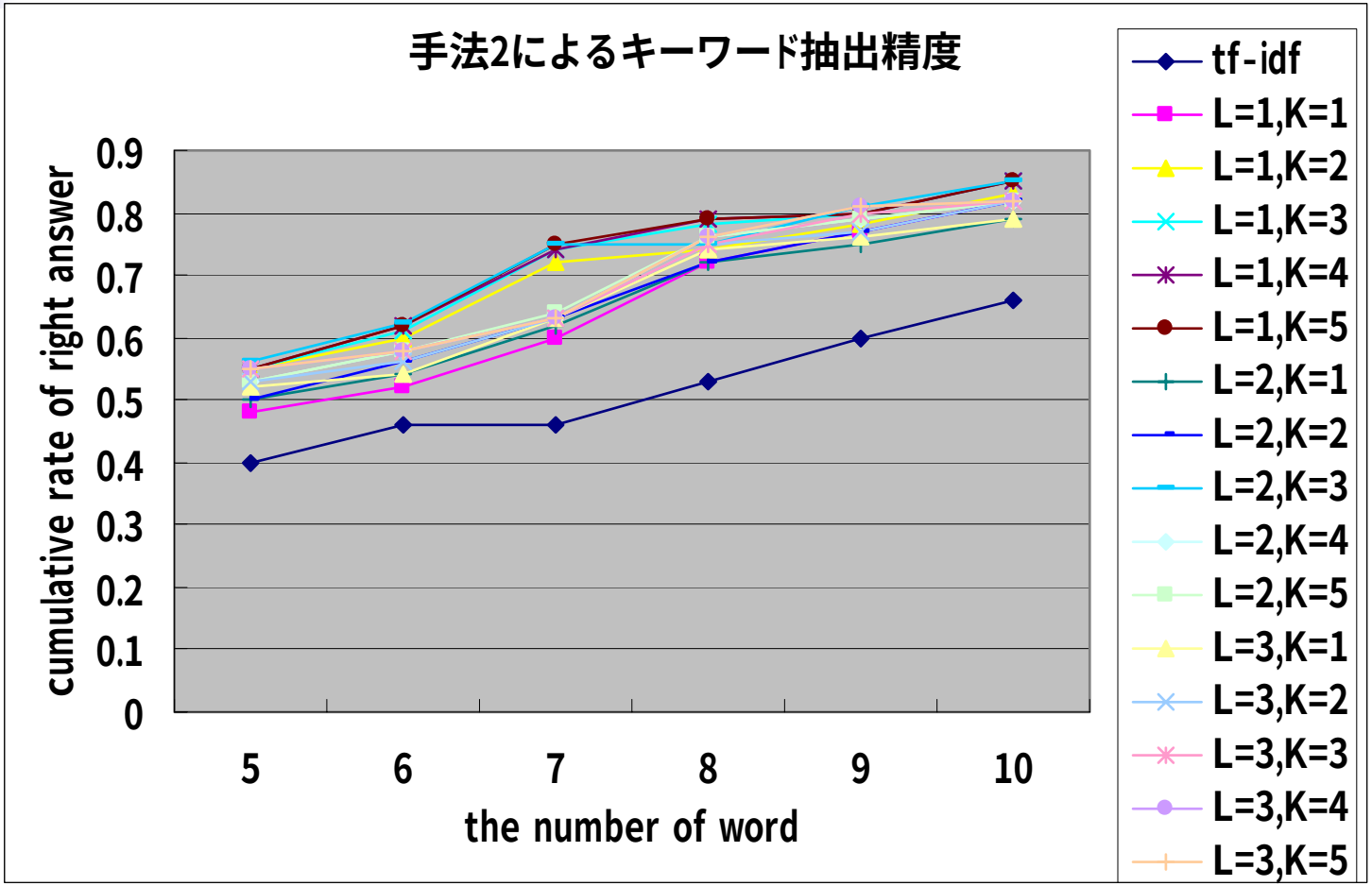
実験

- 実験1
  - リンク先ページを考慮したことによるキーワード抽出精度の比較（スコアの高い上位10語までに、いくつのキーワードが含まれるか）
- 実験2
  - 改良した特徴ベクトルによる検索精度の比較

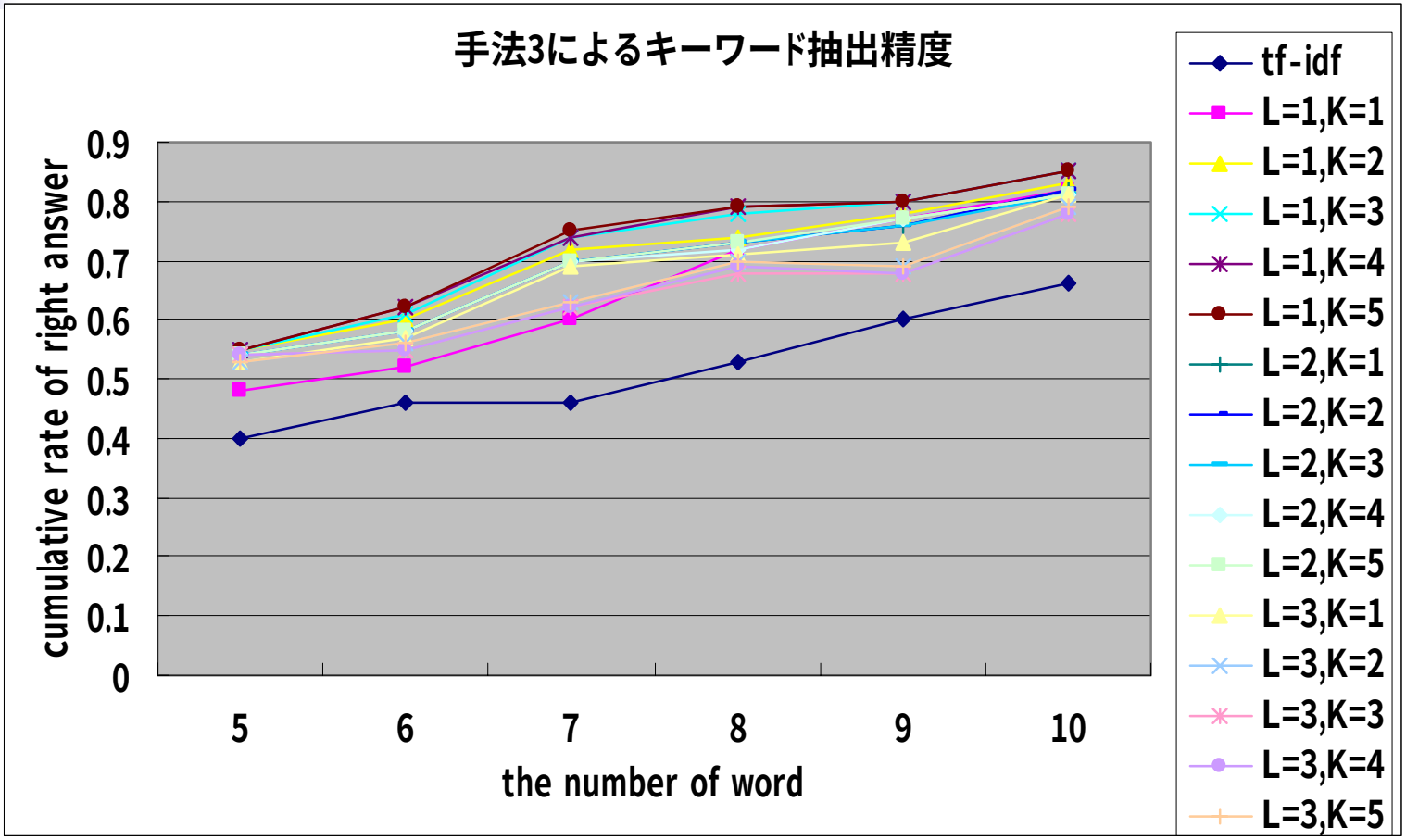
実験1 (手法1)


手法1によるキーワード抽出精度
cumulative rate of right answer
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5
6
7
8
9
10
the number of word
tf-idf
L=1
L=2
L=3
L=4
L=5

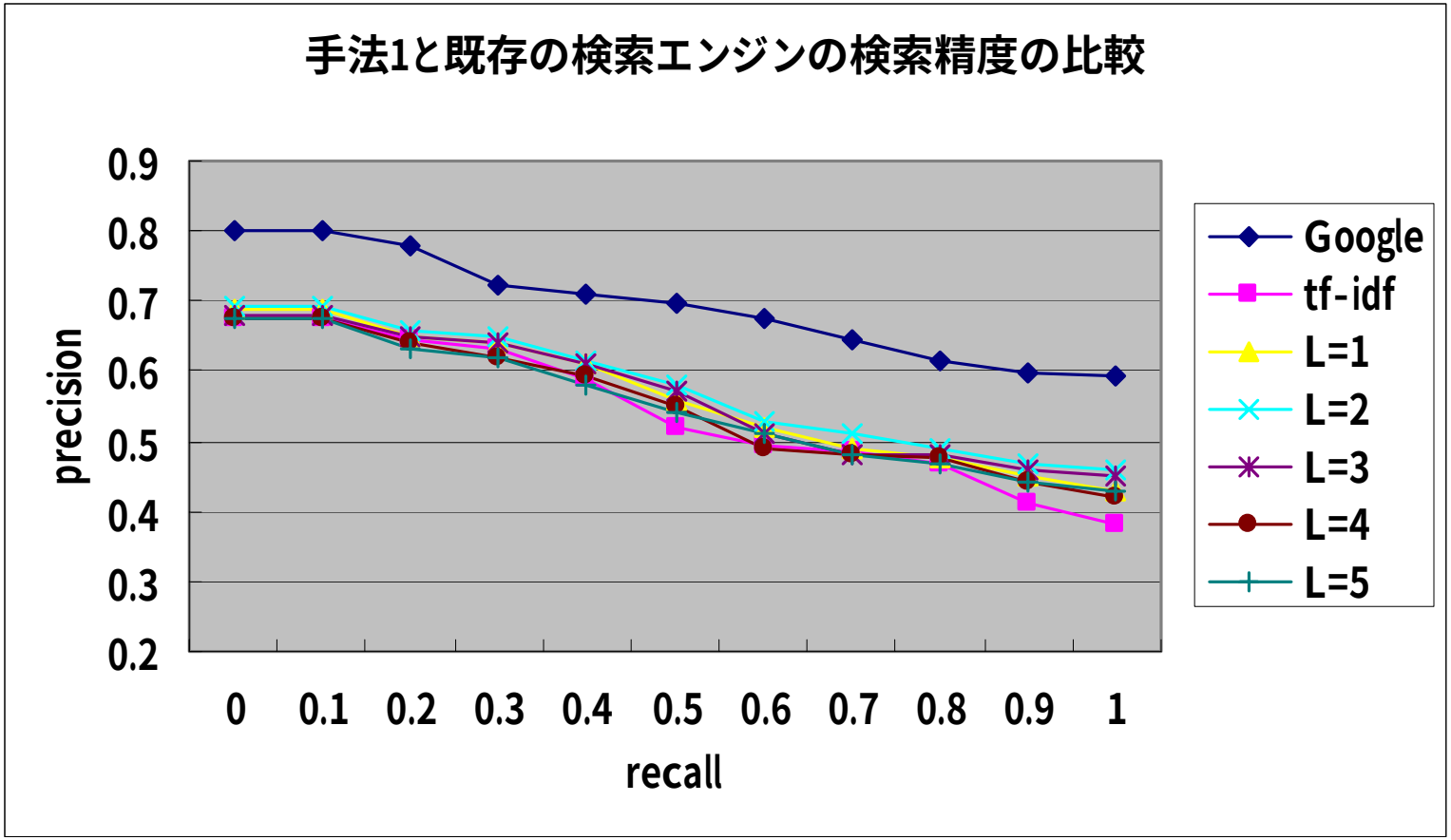
実験1（手法2）


手法2によるキーワード抽出精度
cumulative rate of right answer
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5
6
7
8
9
10
the number of word
tf-idf
L=1,K=1
L=1,K=2
L=1,K=3
L=1,K=4
L=1,K=5
L=2,K=1
L=2,K=2
L=2,K=3
L=2,K=4
L=2,K=5
L=3,K=1
L=3,K=2
L=3,K=3
L=3,K=4
L=3,K=5

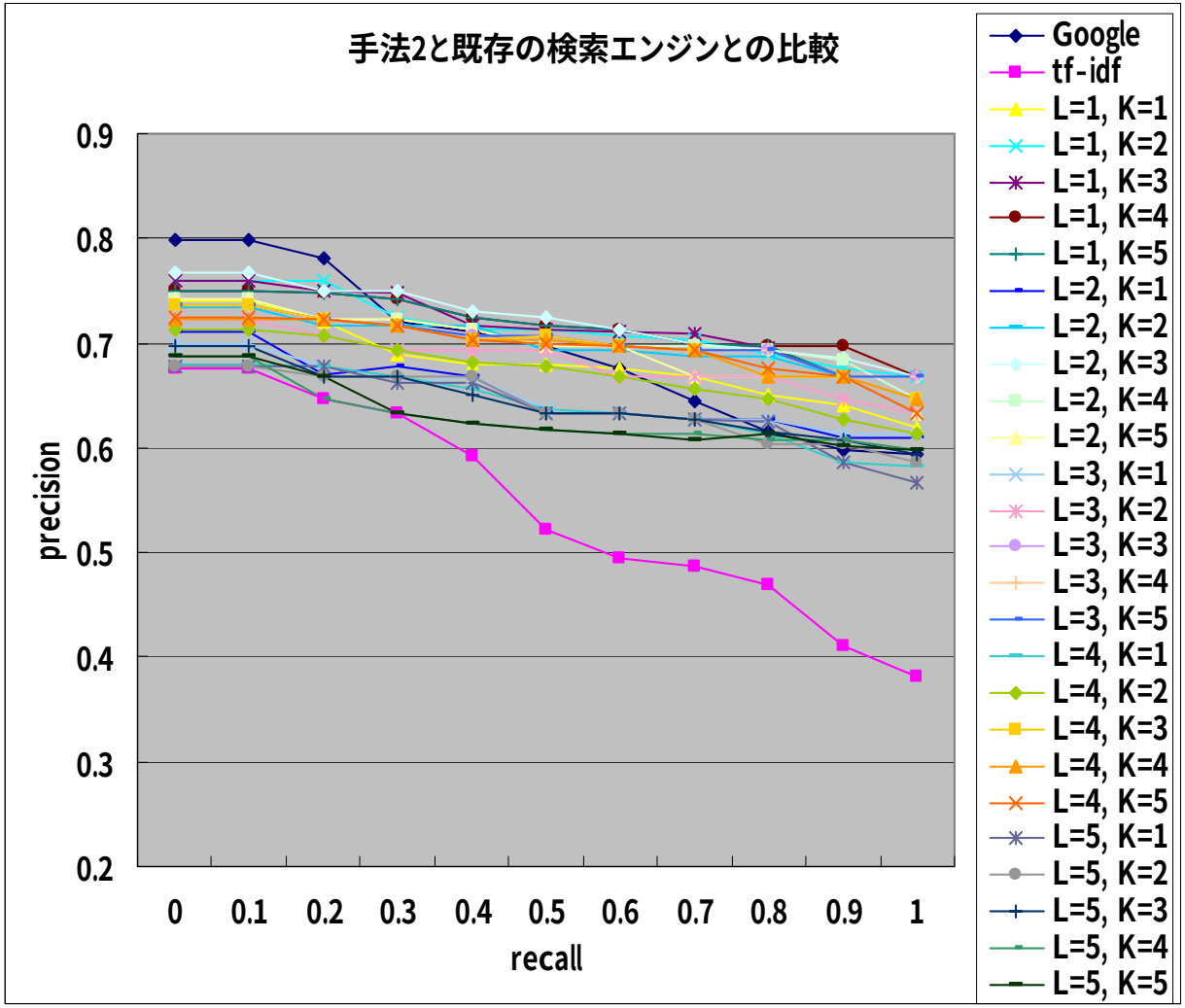
実験1 (手法3)


手法3によるキーワード抽出精度
cumulative rate of right answer
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5
6
7
8
9
10
the number of word
tf-idf
L=1,K=1
L=1,K=2
L=1,K=3
L=1,K=4
L=1,K=5
L=2,K=1
L=2,K=2
L=2,K=3
L=2,K=4
L=2,K=5
L=3,K=1
L=3,K=2
L=3,K=3
L=3,K=4
L=3,K=5

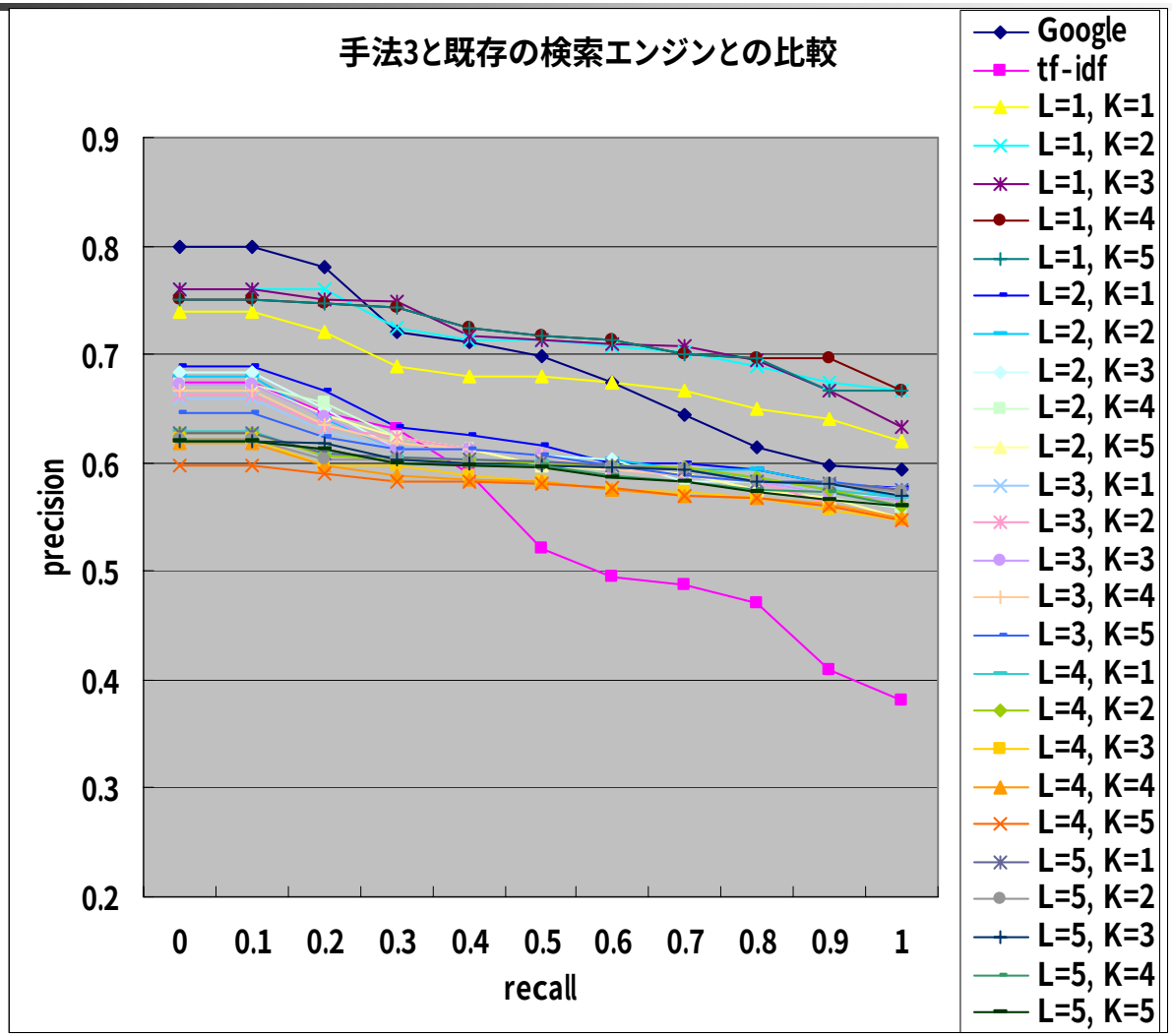
実験2 (手法1)


手法1と既存の検索エンジンの検索精度の比較
precision
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
recall
Google
tf-idf
L=1
L=2
L=3
L=4
L=5

実験2 (手法2)


手法2と既存の検索エンジンとの比較
precision
recall
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Google
tf-idf
L=1, K=1
L=1, K=2
L=1, K=3
L=1, K=4
L=1, K=5
L=2, K=1
L=2, K=2
L=2, K=3
L=2, K=4
L=2, K=5
L=3, K=1
L=3, K=2
L=3, K=3
L=3, K=4
L=3, K=5
L=4, K=1
L=4, K=2
L=4, K=3
L=4, K=4
L=4, K=5
L=5, K=1
L=5, K=2
L=5, K=3
L=5, K=4
L=5, K=5

実験2 (手法3)


手法3と既存の検索エンジンとの比較
precision
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
recall
Google
tf-idf
L=1, K=1
L=1, K=2
L=1, K=3
L=1, K=4
L=1, K=5
L=2, K=1
L=2, K=2
L=2, K=3
L=2, K=4
L=2, K=5
L=3, K=1
L=3, K=2
L=3, K=3
L=3, K=4
L=3, K=5
L=4, K=1
L=4, K=2
L=4, K=3
L=4, K=4
L=4, K=5
L=5, K=1
L=5, K=2
L=5, K=3
L=5, K=4
L=5, K=5

まとめ

- tf-idf法によって作成した特徴ベクトルをリンク先ページを用いて改良する手法を提案
  - リンク先ページを使用することは、キーワード抽出や検索のための索引付けとして有効
  - Webページの内容は、そのページから2～3リンク先のWebページに集約される

  といった知見が得られた。

今後の課題

- 個々のWebページのリンク環境に応じた特徴ベクトルの生成法についての検討、および検索精度の検証