

隣接情報を用いた類似文書検索とリランキング

古川 修平[†] 青野 雅樹[‡]

豊橋技術科学大学情報工学系 〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1

E-mail: [†] frsh@kde.ics.tut.ac.jp, [‡] aono@ics.tut.ac.jp

あらまし 特許の先願調査や先行研究の検索のように、関連した文献を出来得る限り取りこぼしなく発見したいという検索要求が存在する。このような場合、クエリ文書と語彙的に類似していなくとも、クエリ文書と意味的に類似する文書はユーザによって有用である可能性がある。

本研究では、文書の隣接情報を用いて特徴空間の局所的な関係性をとらえる次元削減手法である LPI(Locality Preserving Indexing)、及び、検索結果のランキングが固有の多様体構造を持つことに着目した多様体ランキングを用いて、類似文書検索の性能を向上させる手法を提案する。提案手法を実装し、特許文献のテストセットを用いて行った評価実験の結果から、提案手法を用いた類似文書検索と従来の検索手法との比較、検討を行う。

キーワード 類似文書検索, 次元削減, LPI, 多様体ランキング

Document Similarity Search and Reranking using Adjacency Information

Shuhei FURUKAWA[†] and Masaki AONO[‡]

Department of Information and Computer Sciences, Toyohashi University of Technology

1-1 Hibarigaoka, Tempaku-cho, Toyohashi-shi, Aichi-ken, 441-8580 Japan

E-mail: [†] frsh@kde.ics.tut.ac.jp, [‡] aono@ics.tut.ac.jp

Abstract There is a need to retrieve relevant documents as much as possible, in situations such as invalidation search of patents and exhaustive search of prior research articles. In these situations, it is highly desirable to find “semantically” similar documents, rather than lexically similar ones.

In this paper, we propose methods for enhancing the performance of document similarity search by two different approaches. The first approach employs LPI(Locality Preserving Indexing), a dimensionality reduction taking into account of local structure of feature space based on adjacency information between documents. The second approach employs “manifold ranking”, a semi-supervised learning taking into account of inherent manifold structure iteratively.

We have implemented the proposed method, experimented using patent data set, and compared the results with previous methods.

Keyword Document similarity search, dimensionality reduction, LPI, manifold ranking

1. はじめに

情報技術の発展により、文書データの蓄積が加速しており、それらの文書データを利用するために、大規模データの中から文書を検索する必要性が高まっている。我々は特に、特許の先願調査や先行研究の検索のように、関連した文献を出来得る限り取りこぼしなく発見したいという類似文書検索に注目している。このような場合、クエリ文書自体と語彙的な類似度は低くとも、クエリ文書と意味的な類似度の高い文書はユーザによって有用である可能性がある。

本研究では、文書の隣接情報を用いて特徴空間の局所的な関係性をとらえる次元削減手法である

LPI(Locality Preserving Indexing)[3]、及び、検索結果のランキングが固有の多様体構造を持つことに着目した多様体ランキング[5]を用いて、類似文書検索の性能を向上させる手法を提案する。

提案手法の有効性を確認するために、NTCIR-3[7]の特許検索タスクで用いられたテストセットを用いて評価実験を行った。実験では、提案手法を用いた類似文書検索と従来の検索手法との比較、検討を行った。

以下、2 節で関連研究、3 節で局所的な構造を保存する次元削減を用いた類似文書検索について、4 節で多様体ランキングを用いた検索結果のリランキングについて述べ、5 節で評価と考察を行う。

2. 関連研究

本節では、次元削減と隣接情報を考慮したランキングに関する関連研究について述べる。

2.1. 次元削減

ベクトル空間モデルでは、文書データが増加するに従いベクトルの次元数が増加するため、リソース使用量の増加、不必要な索引語による検索性能の低下が起こる。このような問題に対処するため、文書ベクトルの次元削減が行われる。

LSI(Latent Semantic Indexing)[1]は文書ベクトルの次元削減手法としてよく用いられる。LSIでは、ベクトル空間における単語・文書行列に対して特異値分解(Singular Value Decomposition;SVD)を行い、元の列ベクトルの次元よりも低い次元の列ベクトルに写像する手法である。LSIで次元削減された空間では多くの場合において、潜在的な意味が近いベクトル間の距離が小さくなり、検索性能が向上することが知られている。

HeらはLSIとは異なる次元削減手法としてLPI(Locality Preserving Indexing)を提案した[3]。LPIは、単語・文書行列に対して局所性保存射影(Locality Preserving Projection;LPP)[2]を行い、次元削減を行う手法である。LPIでは文書間の類似行列を定義し、この行列を利用して次元削減を行うため、文書ベクトル間の局所的な関係を保持したまま次元を落とすことが可能である。

本研究では、まず単語・文書行列にLSIで次元削減を行い、その行列に対してLPPを計算するという手法で局所的な構造を保存した次元削減を実現する。

2.2. 隣接情報を考慮したランキング

データの特徴量空間の構造を考慮してランキングを行う手法の一つとして、多様体ランキングが存在する。多様体ランキングは、データの特徴量空間が多様体構造を成すと仮定し、クエリのみをラベル付きデータ、その他のデータをラベルなしデータであるとする半教師付き学習によってランキングを行う[5]。

本研究では、一度単純な類似文書検索を行った後、検索結果下位のデータを負のラベル付きデータとして見なし、多様体ランキングを用いることで検索結果のリランキングを実現する。

3. LSIとLPPを用いた類似文書検索

本研究では、ベクトル空間モデルの単語・文書行列に対してまずLSIを用いて次元削減を行い、その結果得られる行列に対してLPPを用いてさらに次元削減する。本節では、このアルゴリズムについて述べる。

3.1. 提案次元削減手法のアルゴリズム

以下のような m 次元の n 個の文書ベクトルのセットが与えられたとする。

$$\{x_1, \dots, x_i, \dots, x_n\} \subset \mathbb{R}^m$$

提案手法によって、この文書ベクトルを p ($p \ll m$)次元に削減するアルゴリズムは以下の手順になる

1. $\{x_i\}$ に対しLSIを用いて、 l ($p < l \ll m$)次元、 n 個のデータ $\{y_i\}$ を得る。
2. y_i の k 近傍に y_j が存在する、または y_j の k 近傍に y_i が存在する場合、 $s_{ij} = s_{ji} = \text{sim}(y_i, y_j)$ 、そうでない場合 $s_{ij} = s_{ji} = 0$ となるような類似行列 W を作成する。ここで、 sim はベクトル間の類似度を計算する関数で、今回はコサイン類似度を用いる。
3. 以下の一般化固有値問題を解き、固有値 λ 、および固有ベクトル v を求める。

$$YLY^T v = \lambda YDY^T v$$

ここで、 D は $d_{ii} = \sum_j s_{ij}$ である対角行列、

$L = D - W$ である。また Y は i 列が y_i である行列である。

4. 固有値 λ を昇順にソートし、対応する固有ベクトルを p 個選択し、 V_{lpp} とする。つまり

$$V_{lpp} = [v_0, v_1, v_2, \dots, v_{p-1}]$$

となる。ただし、 v_i に対応する固有値は λ_i 、かつ

$$\lambda_0 < \lambda_1 < \dots < \lambda_{p-1}$$

5. 以下の式で、 p 次元の列ベクトル z_i を持つ行列 Z を得る。

$$Z = V^T Y$$

$$V = V_{LSI} V_{LPP}$$

ただし、 V_{LSI} は $\{x_i\}$ にLSIを用いた際の変換行列である。

類似検索を行う際には、クエリ文書の特徴ベクトル q に対し、 p 次元のベクトル $q_{lpp} = V^T q$ を作成し、 q_{lpp} と z_i との比較によって、類似度を得る。

4. 多様体ランキングを用いた検索結果のリランキング

本研究では、類似文書検索の結果上位1000件に対して多様体ランキングを用いることで新たなランキング結果を得る。

多様体ランキングは、特徴量空間が多様体構造を成すと仮定し、特徴量間の類似度をもとに特徴量空間を解析する。本節では、多様体ランキングのアルゴリズムと、検索結果のリランキングへの適用について述べる。

4.1. 多様体ランキングのアルゴリズム

以下のような m 次元の n 個のデータセットが与えられたとする.

$$\{\mathbf{x}_1, \dots, \mathbf{x}_q, \mathbf{x}_{q+1}, \dots, \mathbf{x}_n\} \subset \mathbb{R}^m$$

ここで, $\mathbf{x}_1, \dots, \mathbf{x}_q$ は検索クエリのデータ, その他は検索対象データベースに含まれるデータである. また, $\mathbf{x}_i$ に対してランキングスコア f_i を与えるランキング関数を f とする. この f は, $\mathbf{f} = [f_1, \dots, f_n]^T$ というベクトルと見なす事ができる. 最後に, $\mathbf{x}_i$ が検索クエリである場合に $y_i = 1$, それ以外では $y_i = 0$ となるベクトル $\mathbf{y} = [y_1, \dots, y_n]^T$ を定義する. これらを用いて, 多様体ランキングのアルゴリズムは以下の手順になる.

1. データ同士の距離を昇順に並べ, 連結グラフが作られるまで距離の近いデータ間を接続していく.
2. データ $\mathbf{x}_i$ と $\mathbf{x}_j$ が接続している場合に, 要素が $W_{ij} = \text{sim}(\mathbf{x}_i, \mathbf{x}_j)$ で定義される類似行列 $\mathbf{W}$ を計算する. sim は 2 つのデータの類似度を計算する関数, 対角要素は $W_{ii} = 0$ とする.
3. (i, i) の要素が $\mathbf{W}$ の i 行の総和である対角行列 $\mathbf{D}$ を用いて, $\mathbf{W}$ を $\mathbf{S} = \mathbf{D}^{-1/2} \mathbf{W} \mathbf{D}^{-1/2}$ と対称正規化する.
4. $\mathbf{f}(t+1) = \alpha \mathbf{S} \mathbf{f}(t) + (1-\alpha) \mathbf{y}$ を収束するまで繰り返す. α は 0 以上 1 未満 の値をとるパラメータである.
5. $\mathbf{f}(t)$ の収束値を $\mathbf{f}^*$ とする. $\mathbf{f}^*$ の値 f_i^* を降順にソートし, 対応する文書 i の新たなランキングスコアとする.

$\mathbf{f}(t)$ は, $\mathbf{f}^* = \beta(\mathbf{I} - \alpha \mathbf{S})^{-1} \mathbf{y}$ に収束することが関数 $\{f_i(t)\}$ に関する定理によって知られている. ただし, $\beta = 1 - \alpha$ である.

4.2. リランキングへの適用

本研究では, クエリ文書の文書ベクトルを $\mathbf{x}_l$ とし, 類似文書検索において上位 i 番目の文書ベクトルを $\mathbf{x}_{i+1}$ とすることで, 多様体ランキングを類似文書検索へ適用する. ただし, i は 1000 以下である.

データセットが大きいと, 連結グラフのリンクが少ないと, 関連のあるデータ間が接続されない恐れがある. よって本研究では, 連結グラフに k 近傍グラフ ($k=5$) を用いて, 各データの 5 近傍のデータ群に対してリンクを持つグラフを作成する. グラフを作成するためのデータ間距離には, 文書ベクトル間のコサイン類似度を 1 から引いた値を用いる.

類似行列 W_{ij} の要素の値となる $\text{sim}(\mathbf{x}_i, \mathbf{x}_j)$ にも, 2 つのベクトル $\mathbf{x}_i, \mathbf{x}_j$ のコサイン類似度を使用する.

また, 検索結果の下位の文書はクエリとは似ていないと予想されるため, 下位 r 件の文書ベクトルを負のラベル (-1) 付きデータとして用いる. つまり, 前項におけるベクトル $\mathbf{y}$ を以下のように変更する.

$$\mathbf{y} = [1, 0, \dots, 0, \overbrace{-1, \dots, -1}^r]^T \quad \dots \quad (1)$$

5. 評価実験

5.1. 実験概要

提案手法の類似文書検索に対する効果を確認するために, 3 節の手法 (LSI+LPP), 4 節の手法 (MR) と, 従来手法であるベクトル空間モデル (VSM), LSI を比較した. それぞれの手法の上位 100 件の出力に対し, NTCIR-3 で用いられた特許検索タスクの正解セットによる適合性判定を行った. この正解セットには, 適合, 部分適合, 不適合の 3 値が含まれている.

検索課題として NTCIR-3 の特許検索タスクで用いられた Topic0001 から Topic0020 を用いた. また各手法のパラメータの調整用データとして Topic0021 から Topic0031 を用いた.

検索対象集合として NTCIR-3 で使用されたデータセット中, 1998-1999 年の公開特許公報全文 (約 70 万件) から Topic ごとの正解文献を含む 2000 件を抽出した. (正解文献以外はランダムに抽出)

5.2. 実験条件

文書ベクトルは, 公開特許公報全文を形態素解析器 MeCab[6] にかけた結果が「名詞」, 「未知語」, または連続する「接頭辞, 名詞, 未知語, 句読点以外の記号」で構成されるワードの TF 値を要素とした. また, DF 値が 3 未満, または 800 以上のワードは, 類似文書検索用のベクトルの要素として不適格であると判断し, 削除した. このベクトル作成によって, 2000 件の特許から 8943 次元のベクトルが得られた.

5.3. 評価尺度

ランキングの評価尺度として, 多値の適合度に対応する DCG (Discounted Cumulative Gain)[4] を用いた. DCG はランキング上位に正解が存在するほどスコアが高くなる指標である.

DCG は, 次式で表される.

$$DCG_i = \begin{cases} G_i & \text{if } i=1 \\ DCG_{i-1} + \frac{G_i}{\log_b i} & \text{otherwise} \end{cases} \quad \dots \quad (2)$$

ここで, G_i は i 番目の文書の適合度の度合いを示す値である. 本研究では, 適合を 2, 部分適合を 1, 不適合を 0 とした. また, b は検索結果に対するユーザの寛容さを示すパラメータである. 本研究では $b=2$ とした.

5.4. 実験結果

5.4.1. 提案次元削減手法の結果

提案した次元削減手法 (LSI+LPP) の効果を確認するために, VSM, LSI との比較を行った. 類似行列 $\mathbf{W}$ を計算する際の k 近傍グラフの k は 5, パラメータ l は 500 とした. 比較に用いる LSI では, 80 次元に次元削減を行った.

LSI+LPP と VSM, LSI の検索結果上位 20 件までの DCG の平均値を示すグラフを図 1 に示す.

LSI+LPP を用いて 50 次元に削減されたベクトルでは, VSM, LSI を用いた類似文書検索よりも高い検索精度が得られた. しかし, 10 次元, 20 次元, 100 次元への削減では逆に検索精度が低下した.

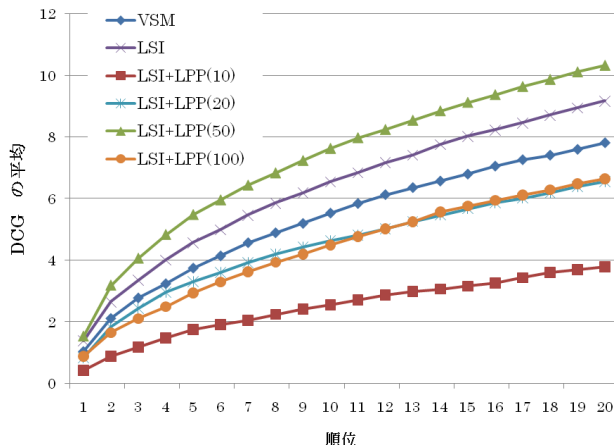


図 1 LSI+LPP と LSI,VSM の DCG

5.4.2. MR の結果

MR の効果を確認するために VSM の検索結果, LSI の検索結果, LSI+LPP の検索結果に MR を適用し, 適用前の結果との比較を行った. 類似行列 W を計算する際の k 近傍グラフの k は 5, MR のパラメータ α は 0.96 とした. また, MR のパラメータの値には VSM に対して MR を適用する実験で調整した値をすべての実験で用いた.

VSM の検索結果に MR を適用した結果 (VSM+MR) と VSM の, それぞれ上位 20 件までの DCG の平均値を示すグラフを図 2 に示す.

式(1)のパラメータ r の値を 0 または 5 にした時, VSM よりも検索性能が向上した. しかし, r の値を 10 以上に設定すると検索性能が低下した.

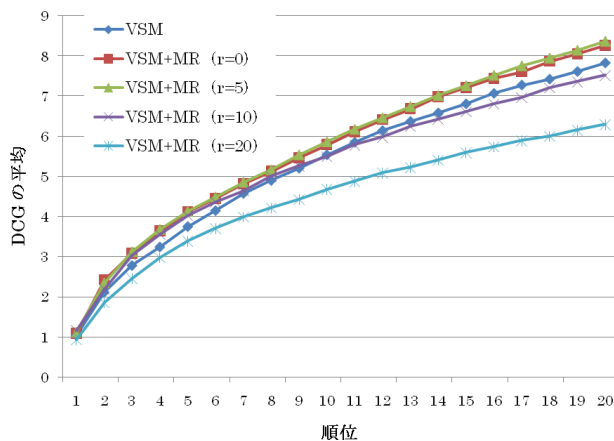


図 2 VSM+MR と VSM の DCG

次に, LSI の結果に MR を適用した結果 (LSI+MR) と LSI の, それぞれ上位 20 件までの DCG の平均値を示すグラフを図 3 に示す. 図 3 中の, LSI の括弧の中の数字は 80 次元へ次元削減を行った結果であることを示す.

図 3 に示すとおり, $r=5$ のときに MR を適用した結

果の中では最も良い結果となったが, MR を適用しない結果のほうが高い DCG を得た.

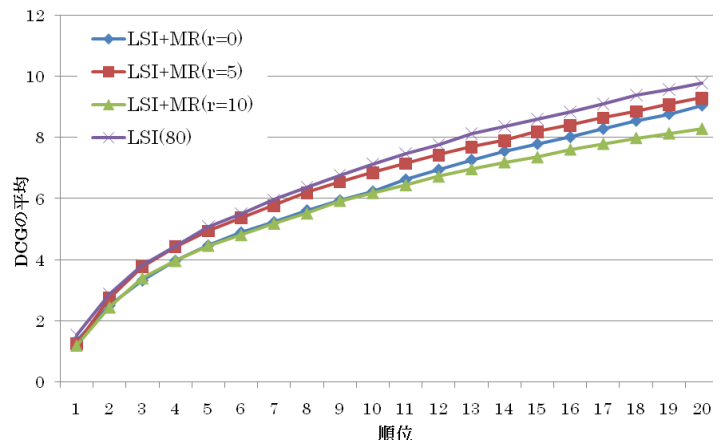


図 3 LSI+MR と LSI の DCG

次に, LSI+LPP の結果に MR を適用した結果 (LSI+LPP+MR) と LSI+LPP の, それぞれ上位 20 件までの DCG の平均値を示すグラフを図 4 に示す. 図 4 中の, LSI+LPP の括弧の中の数字は 50 次元へ次元削減を行った結果であることを示す.

図 4 に示すとおり, $r=0$ のときに MR を適用した結果の中では最も良い結果となったが, LSI と同様, MR を適用しない結果のほうが高い DCG を得た.

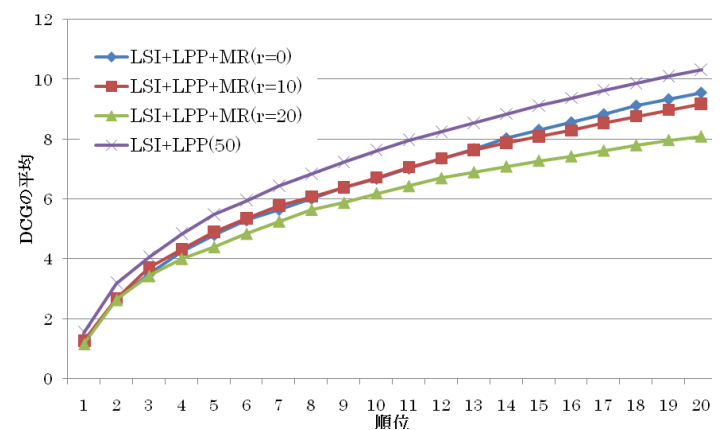


図 4 LSI+MR と LSI の DCG

5.5. 考察

5.5.1. 提案次元削減手法の効果

図 1 に示す通り, 提案次元削減手法 (LSI+LPP) で 50 次元に次元削減した時, 従来手法よりも精度が向上し, 10 次元, 20 次元, 100 次元への次元削減では精度が低下した. 今回用いた検索課題は 20 種類であるため, データ中に含まれるトピックは最低 20 種類存在する. トピック毎の適合文書以外の文献をランダムで抽出し

た結果、トピックの種類が 50 に近くなったことが、50 次元で結果が向上した理由ではないかと予想される。10 次元、20 次元に削減した場合、トピックを分類するために適切な次元が不足し、精度が低下したと考えられる。逆に、100 次元に削減した場合、データを分類するのに適切でない次元が残りすぎたため、精度が落ちたものと考えられる。LSI+LPP での次元削減においては、扱うデータ中に含まれているトピックの種類を考慮して、何次元に削減するかを決定する必要があると考えられる。

また、検索課題文書に関連のある文献が検索対象の文書集合に多く存在する場合、結果が良くなる傾向が見られた。このように、文書集合中に関連文書が多く含まれているトピックでは、文書間の類似度が高くなくとも局所的なクラスタ関係が捉えられやすく、提案手法が有効に働くと考えられる。

5.5.2. MR の効果

図 2 に示す通り、式(1)のパラメータ r の値が 5 以下の時、従来手法よりも精度が向上し、10 以上の時、精度が低下した。

VSM のランキングで、990 位付近に適合文書が存在した場合、式(1)における負のラベルが結果に悪影響を及ぼしていることがあった。MR を行う前のランキングの信頼度、およびデータ中の適合文書の総数などが予測可能な場合、それらを考慮することで、パラメータ r により適切な値を設定することが可能であると考えられる。

また、図 3、図 4 に示したとおり、LSI や LSI+LPP で次元削減を行ったデータでの類似文書検索の結果に対して MR を適用しても結果が向上しなかった。今回の実験では比較を行いやすいように、VSM+MR で調整したパラメータを LSI+MR や LSI+LPP+MR でも固定して実験を行った。しかし、各パラメータの最適値はクエリ文書や、検索対象の文書集合が含むトピック数、他のパラメータの値などに依存すると思われる。特に、式(1)のパラメータ r は、多様体ランキングの反復式 $f(t)$ の内容を変更するパラメータであるため、反復式内のパラメータ α との関連が強いと予想される。これらパラメータの最適値を自動的に推定することが今後の課題である。

また、負のラベルとして今回は -1 を用いた。しかし、正のラベルは“クエリ文書”という絶対的な正解であるのに対して、今回追加した負のラベルは“クエリ文書と関連がない文書である可能性が高い文書”という予想に基づいており、正のラベルに対して確信性が低い。よって、負のラベルとして用いる値は -1 よりも絶対値の小さい値の方が適している可能性がある。

6. おわりに

本研究では、特許の先願調査や先行研究の検索のように、入力した文書と関連した文献を、出来得る限り取りこぼしなく発見したいという検索要求に対して、高い精度の検索結果を返すことを目的として、文書間の隣接情報を用いて次元削減を行う類似文書検索手法、およびランキング手法を提案した。

また、文書間の隣接情報を用いて次元削減を行う提案文書検索手法と従来の文書検索手法を実験によって比較することで、特定のパラメータにおいて従来手法よりも高い精度での類似文書検索が可能であることを示した。また、それらの検索結果に対して適切なパラメータで提案ランキング手法を用いることにより、ランキングの結果が向上することを示した。

今後の課題として、各種パラメータの自動推定や多様体ランキングを行う際の連結グラフの作成方法の検討、負のラベルとして用いる値の検討が挙げられる。

謝辞

NTCIR テストコレクションは国立情報学研究所の許諾を得て使用させていただきました。また本研究は科研費基盤 (C) 20500090 の支援を受けて遂行しました。

文 献

- [1] S.C.Deerwester, S.T.Dumais, T.K.Landauer, G.W.Furnas, and R.A.Harshman, “Indexing by latent semantic analysis”, Journal of the American Society for Information Science, Vol.41, no.6, pp.391-407, October 1990.
- [2] X.He, and P.Niyogi, “Locality Preserving Projections”, in Advances in Neural Information Processing Systems 16, Vancouver, Canada, 2003.
- [3] X.He, D.Cai, H.Liu, and W.Y.Ma, “Locality Preserving Indexing for Document Representation”, Proc. of the 27rd Annual International ACM SIGIR Conference (SIGIR 2004), pp.96-103, July 2004.
- [4] K. Järvelin and J. Kekäläinen, “IR Evaluation Methods for Retrieving Highly Relevant Documents,” Proc. of the 23rd Annual International ACM SIGIR Conference (SIGIR 2000), pp.41-48, July 2000.
- [5] D.Zhou, J.Weston, A.Gretton, O.Bousquet, and B.Schölkopf, “Ranking on Data Manifolds”, Advances in Neural Information Processing Systems 16, pp169-176, MIT Press, Cambridge, MA, 2004.
- [6] MeCab URL: <http://mecab.sourceforge.net/>
- [7] NTCIR(NII Test Collection for IR Systems) URL:<http://research.nii.ac.jp/ntcir/index-ja.html>