

ユーザキーワードの動的定義による情報検索精度に関する検証

陳 幸生[†] 松本 佑也[†] 内海 麻智子[†]

[†] 神奈川工科大学工学部 〒243-0292 神奈川県厚木市下荻野 1030

E-mail: chen@ic.kanagawa-it.ac.jp [†]

あらまし 同じ検索結果を求めるために使用される問い合わせは、ユーザにより異なる。その相違は、検索の精度に影響を与える。本稿では、Latent Semantic Indexing というベクトル検索方式を対象として、ユーザの個人差により生じる問い合わせの相違が情報検索の精度に与える影響について考査した。Latent Semantic Indexing では、問い合わせと検索対象はベクトル化される。検索は、問い合わせに対応するベクトルと検索対象に対応するベクトルとの間の類似度を計算し、順位付けを行うことで達せられる。Latent Semantic Indexing を用いると、キーワードの意味的曖昧による検索精度の低下を抑えられる。そのため、Latent Semantic Indexing は文書検索や多国言語横断検索への応用に注目されている。本稿では、ユーザの個人差を模擬する動的キーワードの定義手法を提案する。また、提案手法を用い、Latent Semantic Indexing 情報検索方式に対し、ユーザの個人差により生じる問い合わせの相違が検索精度に与える影響について検証実験を行う。検証実験により、問い合わせの相違が検索精度に与える変化パターンを明らかにした。

キーワード ベクトル検索方式, LSI, 情報検索, データベース

The Verification on the Information Retrieval Accuracy by a Dynamic User's Keyword Definition Method

Yukio CHEN[†] Yuuya MATUMOTO[†] and Machiko UTUMI[†]

[†] Department of Information & Computer Sciences, Kanagawa Institute of Technology,
1030 Shimoogino, Atsugi-shi, Kanagawa, 243-0292, Japan

E-mail: chen@ic.kanagawa-it.ac.jp [†]

Abstract Queries used to obtain the same retrieval result may be different if they are created by different users. This difference of queries affects the information retrieval accuracy. In this paper, it is examined that how the difference of queries affects the retrieval accuracy of Latent Semantic Indexing (LSI), a vector information retrieval method. In LSI, queries and retrieval candidates are vectorized. Information retrieval is performed by calculating the resemblance between the vectors of queries and retrieval candidates, and ranking the retrieval candidates based on the calculated result. By using LSI, the deterioration of the retrieval accuracy caused by the semantic ambiguity of keywords is held. In this reason, LSI is noticed for the application to the document retrieval and cross-language information retrieval. In this paper, a dynamic keyword definition method is proposed, which simulates the difference of users. Experiments are carried out by using the proposed method to verify the effect of the difference of queries on the retrieval accuracy of LSI. Based on the experimental results, the change patterns of the retrieval accuracy, which are affected by the difference of queries, are clarified.

Keyword vector retrieval method, LSI, information retrieval, database

1. はじめに

近年、WWW の急速な普及、発展にともない、大量の情報へのアクセスが可能となった。一般的に、ユーザが WWW 上で必要な情報を検索する為にはキーワードが必要である。キーワード検索において、現在はパターンマッチングと非パターンマッチングのどちらかを用いるのが主流である。本稿では、多国言語横断検索やマルチメディア情報検索に有効であり、非パターンマッチングの代表的な検索方式であるベクトル検索方式に注目する[1,2,3,4]。ベクトル検索方式の一つ、Latent Semantic Indexing (LSI) は多数の研究者に注目されている[5,6,7,8]。LSI 方式は、文書への検索、特に、言語横断検索（例えば、日本語問い合わせによる英語文書への検索）に利用されている。

検索キーワードとして利用される単語は、ユーザの認識によって意味の違いが生じる。また、同じ検索対象に対して検索の際、利用するキーワードの個数も異なる。LSI ベクトル検索方式は、検索キーワードと検索対象を意味的にベクトル化し、意味的検索空間上に投影し、検索を行う。LSI 検索方式を利用すると、意味的曖昧性による検索精度の低下を抑えることができる。つまり、検索キーワードが同じ意味であれば、異なる単語であっても、同じ検索結果を得ることが可能である[3]。しかし、検索精度は、意味的曖昧性だけではなく、検索の際利用しているキーワードの個数にも影響を受ける。本論文では、キーワードの個数が検索精度へ与える影響について検証する。

本論文は以下の構成となっている。まず2章で、ベクトル検索方式 LSI について概説する。次に3章で、ユーザの個人差を模擬する動的キーワードの定義手法について述べる。4章では、提案手法を用い LSI ベクトル検索方式を対象として、ユーザキーワードの個人差が検索精度へ

与える影響について検証実験を行う。最後に、5章で本稿のまとめを行う。

2. LSI ベクトル検索方式の概要

LSI(Latent Semantic Indexing)では、まず、単語-文書の単語出現頻度行列 $\mathbf{M}$ を作成する。行列 $\mathbf{M}$ の要素 (i, j) は文書 j における単語 i の頻度である。文書から単語の意味を発見するために、 $\mathbf{M}$ を線形代数の特異値分解 (SVD) によって、

$$\mathbf{M} = \mathbf{T}_0 \mathbf{S}_0 \mathbf{D}_0^T$$

と分解する。これにより、似通った文書と単語を近くに配置するように、 $\mathbf{S}_0$ の要素の値が大きい方から k 個の主成分を用い、

$$\hat{\mathbf{M}} = \mathbf{TSD}^T$$

と $\mathbf{M}$ を再構成し、 k -次元の縮退した特徴量空間を形成する。これを LSI 空間と呼ぶ。この空間においては、意味が近い単語と文書は、類似したベクトルとして表現される。

問い合わせは、それを構成するキーワードに重みを付け、問い合わせ行列を作成する。問い合わせ行列の線形変化により、問い合わせは LSI 空間においてベクトルとして表現される。検索は、問い合わせに対応するベクトルと文書に対応するベクトルとの間の類似度を cosine 相関度などにより計算し、順位付けを行うことで達せられる。

3. ユーザキーワードの動的定義手法

問い合わせにおいて、検索キーワード

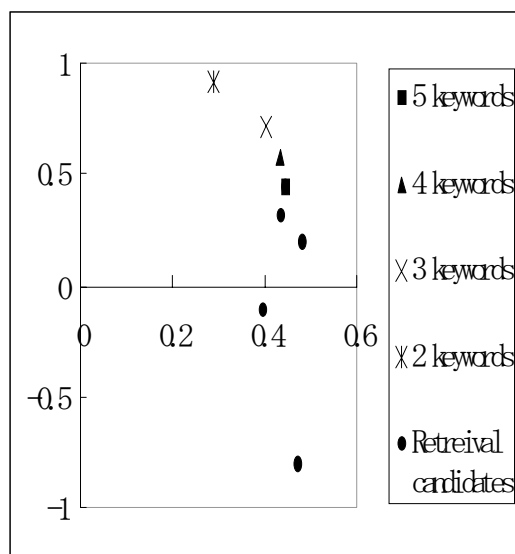


図 1 キーワードの個数の変化による問い合わせの投影変化

Fig.1 The change of the mapping of a query based on the change of the number of the keywords in it

の個数が変化すると、LSI 空間への投影も、それによって変化する。図 1 で、それを示す。図 1 において、キーワードの個数を 5 個から 2 個まで減らすと、その変化に従い問い合わせの LSI 空間への投影も変化することが明らかである。そうすると、問い合わせの投影の変化に従い、検索結果が変わる可能性がある。それが検索精度に影響を与えている。

検索精度を検証するために、問い合わせのキーワードの個数を 1 個から増加していき、個数の増加に従い検索結果を調べていく。単に検索結果を見るだけでは意味がないので、検索結果の正解を予め決めておく必要がある。また、キーワードの個数の最大値を決めなければならない。そこで、検索結果の正解とキーワードの個数の最大値を検討する。LSI 空間において、似通った文書は近くに配置される。その特性に従い、もし、ある問い合わせ Q_i がある文書 F_j の単語数と個々の単語の出現頻度と全く同じであれば、問い

行列 M		文章 ₁	文章 ₂	文章 ₃	文章 ₄	文章 ₅
単語 ₁		0.928009	0.99826	0.474976	0.601196	0.963715
単語 ₂		0.275635	0.36385	0.924011	0.10649	0.149872
単語 ₃		0.700043	0.870178	0.936798	0.928528	0.627502
単語 ₄		0.211029	0.226257	0.534454	0.661072	0.351624
単語 ₅		0.715302	0.666138	0.653931	0.386536	0.820038
Query		0.474976	0.924011	0.936798	0.534454	0.653931

図 2 キーワードの動的定義手法の一例

Fig. 2 An example of the dynamic keyword definition method

合わせ Q_i と文書 F_j は LSI 空間上において同じ点に射影される。すなわち、問い合わせ Q_i である単語-文書行列 M の一列が、文書 F_j の列と同じであれば、問い合わせ Q_i と文書 F_j は LSI 空間上において、同じベクトルとして表現される。ここで、問い合わせ Q_i の検索結果として出現する文書 F_j とその類似文書を検索の正解とする。問い合わせ Q_i に含まれる単語の数はキーワードの個数の最大値とする。以上をふまえ、検索精度を検証検証するために、ユーザキーワードの動的な定義手法を次のようにまとめる。

- (1) 単語-文書行列 M からある 1 列を検証用問い合わせとして切り出す。
- (2) 検証用問い合わせの検索結果を正解とする。
- (3) 検証用問い合わせの単語の数はキーワードの個数の最大値とする。

具体的な例を図 2 に表示する。図 2 では、行列 M の任意の 1 列を、問い合わせとして取り出す。問い合わせの単語数は要素に 0 が無いので 5 つである。

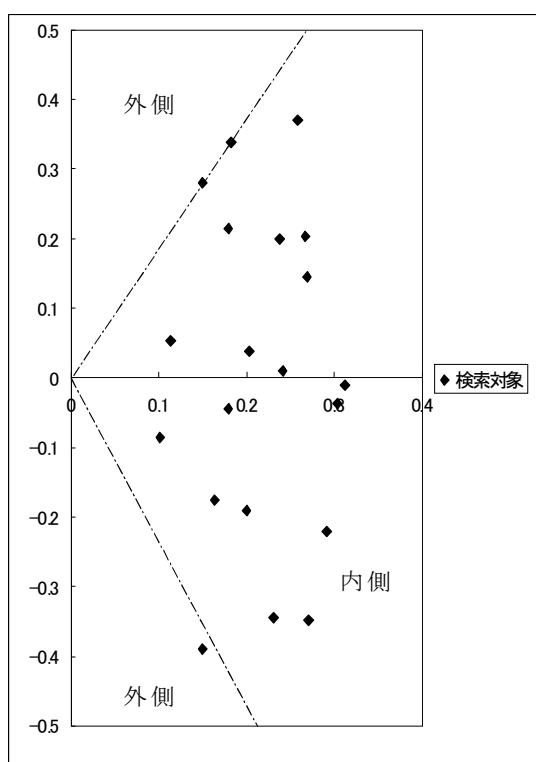


図 3 検索対象の LSI 空間への射影

Fig. 3 The mapping of the retrieval candidates on the LSI space

4. 検証実験

検証実験のために、我々は、特異値分解を求めるプログラムを開発した。このプログラムは、行列 $\mathbf{M}$ の分解値、行列 $\mathbf{T}_0$, $\mathbf{S}_0$, および、 $\mathbf{D}_0$ を求めることが可能である。検証用の行列 $\mathbf{M}$ は、単語-検索対象行列として生成し、20 単語、20 検索対象の行列と設定した。また、ノイズ要素を取り除くために、検証用問い合わせを次のように設定した。

- (1) キーワードの重みは 0 と 1 の間 (≤ 1 , ≥ 0) の値である。
- (2) 重みが 1 になるキーワードは 1 つである。

従って、3 章のキーワードの動的定義

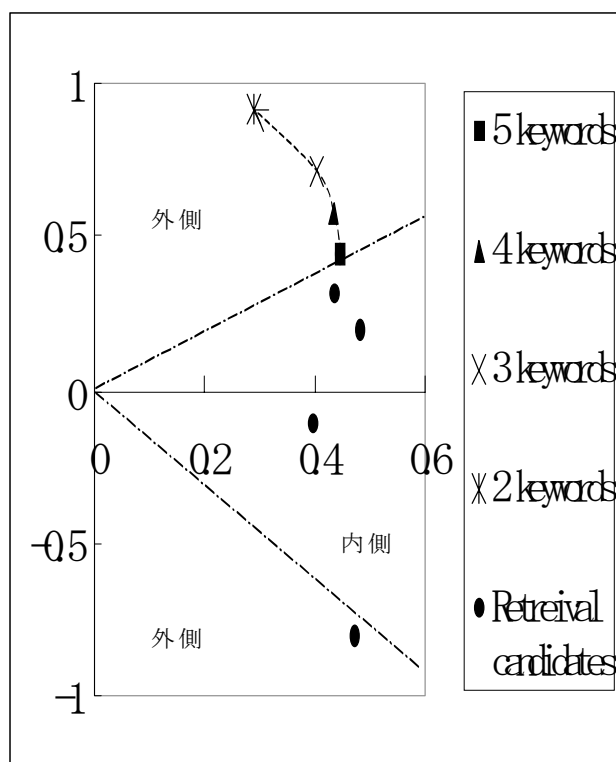


図 4 問い合わせの投影変化パターン (pattern-1)

Fig. 4 A change pattern (pattern-1) of the mapping of a query

手法を用い問い合わせを定義すると、行列 $\mathbf{M}$ の各列もその設定を満たす必要がある。本稿では、行列 $\mathbf{M}$ の各列の値を次のように乱数で生成した。

- (1) $\mathbf{M}$ の各列の値を 0 と 1 の間 (≤ 1 , ≥ 0) の乱数で生成する。
- (2) $\mathbf{M}$ の各列の 1 になる値が 1 つだけである。
- (3) $\mathbf{M}$ の各列の 1 になる値の所在行がそれぞれ異なる。

上記の生成方法に従い、検証用問い合わせを動的に生成できる。また、検索対象も異なる場所に射影され、検索対象の類似性は乱数の生成により定まる。図 3 は検索対象の LSI 空間への射影を示す。

問い合わせキーワードの個数の減少に従い, LSI 空間上の射影位置も変化する. 図 4 はその変化のパターンの一例である. 変化のパターンを説明するために, 我々は, LSI 空間に内側と外側を定義する. 図 4 において, 2つの点線の中を内側, 外を外側と定義する. 図 3 においても同様に定義する. 図 4 の点線矢印で示すように, 問い合わせキーワードの個数の減少に従い, 射影点が内側から外側に移動する. 実験により, キーワードの個数の減少に従い, 問い合わせの LSI 空間上の射影位置は次の 3 種類の変化がある.

(Pattern-1) 内側から外側への移動パターン (図 4)

このパターンは, 問い合わせが外側の検索対象と同じ位置に射影される場合発生する. 類似度を cosine 相関度により計算する場合, 検索精度への影響はない.

(Pattern-2) ほぼ直線で, 0 へ向かって移動するパターン (図 5)

このパターンは, 問い合わせが内側に射影される場合に発生する. 類似度を cosine 相関度により計算する場合, 検索精度への影響はない.

(Pattern-3) 内側から外側へ移動し, 再び内側に移動するパターン (図 6)

このパターンは, 問い合わせが内側に射影される場合発生する. 類似度を cosine 相関度により計算する場合, 検索精度への影響がある.

外側にある検索対象の数が少ないので, Pattern-1, 2 にあてはまる問い合わせの数が少なく, Pattern-3 にあてはまる問い合わせの数が多し. すなわち, 問い合わせキーワードの個数が検索精度に影響を与

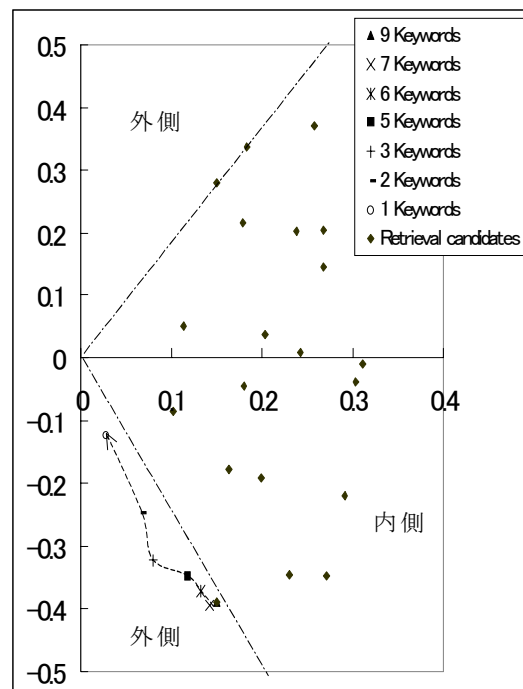


図 5 問い合わせの投影変化パターン (pattern-2)

Fig. 5 The change pattern (pattern-2) of the mapping of a query

えるという結論が出る.

図 6 を観察すると, 次の現象がわかる.

- (1) キーワードの個数の減少に従い, 検索精度が落ちるが, ある程度の減少の後に, 検索精度が再び戻る.
- (2) 精度変化の空間内に検索対象の数が検索精度への影響も大きく, 検索対象の数が多しほど, 検索精度の変化も大きい.

5. おわりに

本論文は LSI ベクトル検索方式を対象として, 実験結果により, キーワードの個数の変化が検索精度に影響を与えることを示した. ユーザキーワードの動的定

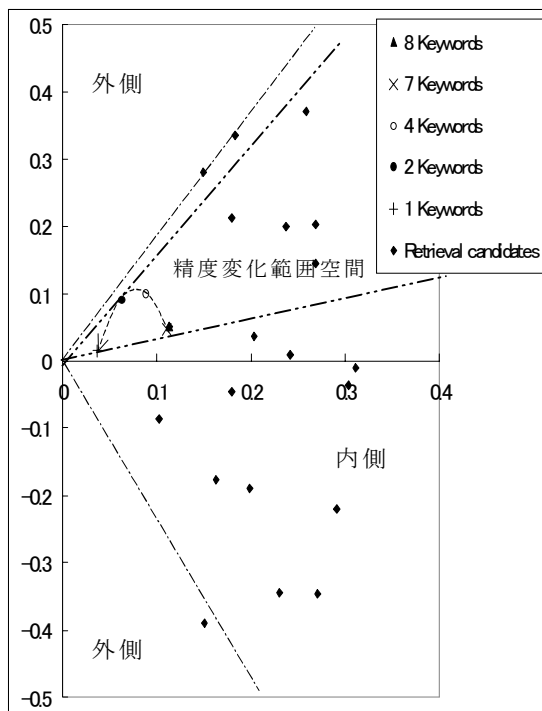


図 6 問い合わせの投影変化パターン (pattern-3)

Fig. 6 The change pattern (pattern-3) of the mapping of a query

義手法により問い合わせを生成し、これを用いて、キーワードの個数の変化が検索精度に与える影響を検証した。今回の実験により、キーワードの個数の変化により 3 種類の問い合わせ投影パターンに分類することができた。また、キーワードの個数の変化により問い合わせ投影の変化はある範囲に限定されたことがわかった。その範囲内に、検索対象の数が多いほど、検索精度の変化も大きいといえる。

謝辞 本研究にあたってのプログラム開発に協力して下さった神奈川工科大学工学部情報工学科 4 年の松野義毅氏、竹田茂馬氏、そして論文作成に協力して下さった中山知映氏、太田祥雅氏に謝意を表す。

文 献

[1]Chen, X., Kiyoki, Y. and Kitagawa, T.,

"A multi-language oriented intelligent information retrieval system utilizing asemantic associative search method," Proceedings of the 17th IASTED International Conference on Applied Informatics, pp.135-140, Innsbruck, Austria, February, 1999.

[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, Dec. 1994.

[3]Scott Deerwester, Susan T. Dumais, Richard Harshman, "Indexing by Latent Semantic Analysis, "Journal of the American Society for Information Science, Vol.41, No.6, pp.391-407, Sept. 1990.

[4]Teruhiko Kanazawa, " (R^2D^2) at NTCIR: Using the Relevance-based Superimposition Model, "Proc. NTCIR Workshop 1, pp83-88, Tokyo, Japan, Aug. 1999.

[5]Tatsunori Mori, Tomoharu Kokubu, Takashi Tanaka, "Cross-Lingual Information Retrieval based on LSI with Multiple Word Spaces, "Proc. NTCIR Workshop 2 Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization, pp.5-67-5-74, Tokyo, Japan, March. 2001.

[6]Susan T. Dumais, Michael L. Littman, Thomas K. Landauer, "Automatic Cross-Language Retrieval Using Latent Semantic Indexing, "AAAI Spring Symposium Technical Report, pp.15-21, March. 1997.

[7]Susan T. Dumais, Bellcore, "Using Latent Semantic Indexing(LSI) for Information Retrieval, Information Filtering and Other Things, "Cognitive Technology Workshop, April 1997.

[8]森辰則, 國分智晴, 田中崇, "空間分割型 CL-LSI による大規模言語横断情報検索, "情報処理学会論文誌: データベース, Vol.43, No.SIG02-05, 2001.