音声言語情報処理 29－17
(1999. 12. 20)

ニュース音声に対する検索方法の比較

鷹尾 誠一　　緒方 淳　　有木 康雄

龍谷大学 理工学部

〒 520-2194 大津市瀬田大江町横谷1-5
Tel: 077-543-7427
E-mail: {tail,ogata}@arikilab.elec.ryukoku.ac.jp, ariki@rins.st.ryukoku.ac.jp

あらまし 近年、放送の多チャンネル化により、多くのニュース番組が放映されるようになった。これを受けて、視聴者側には知りたいニュースだけを見たいという要求が生じている。この要求に対応するには、トピックセグメンテーションや検索などの機能を持つ、ニュースデータベースを構築する必要がある。本研究では、その中からニュース音声に対する記事検索について検討を行った。ニュース音声を対象とする場合、音声認識における単語の湧き出しや欠落が問題となり、従来の単語重要度決定方法やベクトル空間法では対処することができない。この問題点を解決するために本研究では、単語重要度決定方法ではTF-IDFを考慮した相互情報量、ベクトル空間法では単語空間に基づく方法を提案し、良好な結果を得たので報告する。

キーワード：大語彙連続音声認識、記事検索、TF-IDFを考慮した相互情報量、単語空間

Comparison of Retrieval Methods to News Speech

Seiichi Takao　　Jun Ogata　　Yasuo Ariki

Faculty of Science and Technology, Ryukoku University

1-5 Yokotani, Oe-cho, Seta, Otsu-shi, 520-2194 Japan
Tel: +81-77-543-7427
E-mail: {tail,ogata}@arikilab.elec.ryukoku.ac.jp, ariki@rins.st.ryukoku.ac.jp

Abstract Recently, TV news programs are broadcast from all over the world owing to the broadcast digitization. In this situation, TV viewers want to select and watch the most interesting news. In order to satisfy this requirement, news database has to be constructed which has automatic topic segmentation and retrieval function. In this paper, we focus on topic retrieval among them. Conventional term weighting methods and vector space models have no applicability in spoken document retrieval because of error words caused by speech recognition. In order to solve this problem, in this paper, we propose mutual information considering TF-IDF as a new term weighting method, and word space model as a new vector space model.

Key words : LVCSR, article retrieval, mutual information considering TF-IDF, word space model

1 はじめに

近年、放送の多チャンネル化により、多くのニュース番組が放映されるようになった。これを受けて、視聴者側には知りたいニュースだけを見たいという要求が生じている。この要求に対応するには、トピックセグメンテーションや検索などの機能を持つニュースデータベースを構築する必要がある。これを背景に近年、ニュース音声に対するトピックセグメンテーションや検索の研究が数多く行われている[1]-[7]。そこで本研究では、特にニュース音声に対する記事検索に対して検討を行った。

ニュース音声に対する記事検索には3種類の問題点がある。1つ目は音声認識において生じる単語の湧き出しや欠落の問題、2つ目はニュース音声記事からどのように重要語を抽出するかという問題、3つ目は記事間の類似度をどう定義するかという問題である。

1つ目の単語の湧き出しの問題を解決するためには、ディクテーションして得られた単語が湧き出し単語であるか、そうでないかを決定する必要がある。湧き出し単語は主として、記事の内容と関係ない単語である可能性が高い。従って、単語が記事の内容に関係しているか否かを決定することによって、湧き出し単語であるか、そうでないかを決定することができる。記事の内容に関係が深い単語かどうかを決定する技術としては、単語重要度決定法が知られており、よく用いられるものとして相互情報量やTF-IDFが挙げられる。次に単語の欠落の問題を解決する方法を述べる。単語が欠落して検索精度が下がる原因としては、欠落した単語が記事の内容と関係が深い場合が考えられる。具体例として、記事t_1とt_2が関連記事であり、記事t_1とt_2を表す記事ベクトル(3次元)が表1のように表されていたとする。記事ベクトルt_1とt_2の単純類似度は、$\frac{2+25+2}{\sqrt{27}*\sqrt{33}} = \frac{29}{5.19*5.74} = 0.97$である。ディクテーションにより記事$t_1$から単語$w_2$が欠落した場合、両者の単純類似度は$\frac{2+0+2}{\sqrt{2}*\sqrt{33}} = \frac{4}{1.41*5.74} = 0.49$となり、記事$t_1$の内容に関係が深い単語である$w_2$が欠落すると、関連記事の類似度が下がってしまう。

表 1: 記事ベクトル

	w_1	w_2	w_3
t_1	1	5	1
t_2	2	5	2

この問題点を解決するためには、欠落した単語が関連記事に対して持っていた類似度を、残っている他の単語を使って補う方法が効果的であると考えられる。つまり、記事間の単語のオーバーラップ率だけでなく、記事間に存在する単語間の関連度も使って、記事t_1とt_2の類似度を表すことができれば、欠落した単語が関連記事に対して持っていた類似度を残っている他の単語を使って補うことができる。

2つ目の重要語の抽出は、記事ベクトルを作成する場合に、記事の特徴を表さない単語、つまり一般的にどの記事にもよく出現する単語を削除することで実現できる。このため、本研究では単語重要度決定方法を使い、単語が記事の特徴を表す単語であるかどうかを決定した。

3つ目の問題点である記事間の類似度について述べる。人間が任意の2記事を見て関連記事と判断する場合、その判断基準は大きく分けて次の2つが考えられる。

- 記事間で同じ単語が多く出現している。
- 記事間で類似度の高い単語が多く出現している。

1の判断基準を数式化したものは、単純類似度法である。2の判断基準を数式化するためには、単語同士の類似度を考慮する必要がある。この単語同士の類似度を考慮した単純類似度法を、ここでは概念類似度と呼ぶことにする。単語同士の類似度を計算するために単語空間を定義し、この単語空間に基づくベクトル空間法により概念類似度を定義した。

本稿では2でディクテーション、3でベクトル空間法、4でLSI法、5で単語空間に基づくベクトル空間法、6で実験結果について述べる。

2 ディクテーション

2.1 実験条件

用いた言語モデルは、毎日新聞CD-ROM版の45ヶ月分(91年1月～94年9月)の記事から学習したものである。語彙数20Kのback-off bigramで、back-off smoothingにはwitten-bellの推定を用いている。bigramに対するcut-offは1とした。

音響モデルは、男性不特定話者HMMで、単語間の音素文脈依存も考慮したcross-word triphoneモデルである。学習には、日本音響学会新聞記事読み上げ音声コーパスのうち、男性話者137人分の21782発話を用いた。音響特徴量には39次元の特徴パラメータ(12次元のMFCCとパワー、およびそれぞれのΔ、$\Delta\Delta$係数)を用いた。

2.2 連続音声認識実験

評価用音声データには、98年のNHKニュース58記事分(総計1.7時間、総発話数572で4つのニュース番組に含まれている)を用いた。表2に結果を示す。Corrは単語正解率を、Accは単語正解精度をあらわしている。それぞれの値は、式(1),(2)で計算される。

表2の結果で、ディクテーション精度がやや低下している原因について以下に述べる。言語モデルは91年1月～94年9月のデータで学習したものである。これに対して、評価用音声データは98年のニュースであり、時期差が原因となって精度がやや低下していると考えられる。また、言語モデルを構築する際に用いたデータは、トピックの分布が政治、経済に偏っていた。これに対して、評価用音声データでは、それぞれのトピックが均等に存在しており、トピックの分布差が原因となって精度がやや低下していると考えられる。

$$\text{正解率} = \frac{N-D-S}{N} \times 100\% \quad (1)$$

$$\text{正解精度} = \frac{N-D-S-I}{N} \times 100\% \quad (2)$$

N : 全単語数
S : 置換誤り単語数
D : 脱落誤り単語数
I : 挿入誤り単語数

表 2: ディクテーション結果(%)

	Corr	Acc
19980820-12:00NHK	77.83	75.57
19980820-23:00NHK	77.42	75.43
19980824-12:00NHK	76.46	73.74
19980825-12:00NHK	77.81	73.94
Total	76.00	73.27

3 ベクトル空間法

ここではベクトル空間法[8]-[10] について述べる。まず、特に、ベクトル空間法において重要な役割を担う単語重要度決定方法について述べる。次に、ベクトル空間法において類似度を決定する方法を述べる。

3.1 記事検索の手順

ベクトル空間法による記事検索の手順を以下に示す。

1. ニュース音声記事に対してディクテーションを行う。
2. ニュース音声記事におけるそれぞれの単語の重要度を決定する。
3. 記事毎に単語の重要度を閾値処理して、重要語を抽出し記事ベクトルを作成する。
4. 単純類似度法によって、記事ベクトルの比較を行い、クエリー(記事)に対してデータベースから関連記事を抽出する。

3.2 単語重要度決定方法

単語重要度決定方法は先に述べたように、ディクテーションにおいて生じる湧き出し単語の除去、記事の特徴を表す単語の抽出などの役割を担っており、ベクトル空間法において非常に重要な役割を担っている。以下には、単語w_iの記事t_kに対する重要度を決定する方法として、相互情報量、TF-IDF、単語の頻度分布に基づくTF-IDF、TF-IDFを考慮した相互情報量について述べる。

3.2.1 相互情報量

単語w_iと記事t_kとの相互情報量$i(t_k;w_i)$とは、単語w_iを知ることによって、記事t_kに関して得られる情報量の事であり、式(3)で表される。

$$\begin{aligned} i(t_k;w_i) &= i(t_k) - i(t_k|w_i) \\ &= -\log P(t_k) - \{-\log P(t_k|w_i)\} \\ &= \log \frac{P(t_k,w_i)}{P(t_k)P(w_i)} \end{aligned} \quad (3)$$

この情報量は、記事t_kが持っていた情報量$i(t_k)$と、単語w_iを知った後でも、まだ記事t_kが有している情報量$i(t_k|w_i)$の差として定義される。式(3)により相互情報量は、記事t_kの生起確率$P(t_k)$と単語w_iの生起確率$P(w_i)$が独立していれば小さくなり、依存していれば大きな値となる。すなわち、相互情報量が大きければ、記事の特徴をよく表す単語と見なせる。

3.2.2 TF-IDF

TF-IDFは式(4)で表され、単語w_iが記事t_kに現れる回数が高ければ高いほど、TF(Term Frequency)が高くなり、単語w_iが現れる記事数が少なければ少ないほど、IDF(Inverse Document Frequency)が高くなる。したがって、TFは頻度の高い単語という性質を表し、IDFはその記事に偏って現れる単語という性質を表しているので、TF-IDFの値が大きければ、記事の特徴をよく表す単語と見なせる。

$$\begin{aligned} TF \cdot IDF &= TF(w_i,t_k) \cdot IDF(w_i) \quad (4) \\ TF(w_i,t_k) &= \text{単語}w_i\text{が記事}t_k\text{に現れる回数} \\ IDF(w_i) &= \log \frac{\text{索引対象の全記事数}}{\text{単語}w_i\text{が現れる記事数}} \end{aligned}$$

3.2.3 単語の頻度分布を用いたTF-IDF

式(4)のTF-IDFでは、単語w_iが現れる記事数を求めているが、記事の中に現れる単語w_iの頻度に関わらず、

記事数がカウントされている。このため、単語w_iの頻度が低い記事も含めてしまい、IDFが小さく見積もられる可能性がある。この問題を避けるため、式(5)に示すように、単語w_iの頻度を用いる方法が提案されている[11]。

$$I_i = g_i \cdot \log \frac{G_A}{G_i} \tag{5}$$

g_i : (注目している特定の)記事中の単語w_iの頻度
G_i : 全記事中の単語w_iの頻度
$G_A = \sum_i G_i$: 全記事中の総単語数

3.2.4 TF-IDFを考慮した相互情報量

3.2.1節から3.2.3節までの重要度決定方法の基本形は、相互情報量とTF-IDFである。従って、相互情報量とTF-IDFに関して考察する。相互情報量$i(t_k;w_i)$は単語w_iと記事t_kの共起確率しか考慮していないので、単語w_iの発生頻度の大小は無視される。これが相互情報量の弱点である。一方、TF-IDFはTFが単語w_iの発生頻度の大小を考慮しており、相互情報量の弱点を補っている。更に、IDFは単語w_iの出現する記事数によって、単語w_iの記事t_kへの依存度を表しており、共起度を表していると解釈できる。相互情報量とIDFは共に共起度を表しているが、それは異なる観点から見た場合である。従って、相互情報量とTF-IDFを式(6)のように組み合わせれば、より精度良く記事の特徴を表す単語を選択することができる。

$$\begin{aligned} & i(t_k;w_i) \times TF-IDF \\ &= (i(t_k) - i(t_k|w_i)) \cdot TF(w_i,t_k) \cdot IDF(w_i) \\ &= \left(\log \frac{P(t_k,w_i)}{P(t_k)P(w_i)}\right) \cdot TF(w_i,t_k) \cdot IDF(w_i) \end{aligned} \tag{6}$$

3.3 単純類似度法

それぞれのニュース音声記事は、単語重要度決定方法によって抽出された単語を成分とする記事ベクトルで表される。従って、記事ベクトル同士の内積は、ニュース音声記事同士の単語のオーバーラップ率を表し、記事間の類似度を表している。
今、記事t_kの正規化された記事ベクトルを次のように表す。

$$\begin{aligned} X_k &= (x_{1k}, x_{2k}, \cdots, x_{nk})^T \\ &= (x_{1ko}, x_{2ko}, \cdots, x_{nko}, x_{1kc}, x_{2kc}, \cdots, x_{nkc})^T \end{aligned}$$

x_{nko} : 記事t_kのみに出現した単語の頻度
x_{nkc} : 記事t_kと記事t_lに出現した単語の頻度

記事t_lの正規化された記事ベクトルを次のように表す。

$$\begin{aligned} X_l &= (x_{1l}, x_{2l}, \cdots, x_{nl})^T \\ &= (x_{1lo}, x_{2lo}, \cdots, x_{nlo}, x_{1lc}, x_{2lc}, \cdots, x_{nlc})^T \end{aligned}$$

x_{nlo} : 記事t_lのみに出現した単語の頻度
x_{nkc} : 記事t_kと記事t_lに出現した単語の頻度

記事ベクトルX_kとX_lの類似度は次のように表される。

$$\begin{aligned} \cos\theta &= (X_k, X_l) \\ &= (x_{1k}, x_{2k}, \cdots, x_{nk})(x_{1l}, x_{2l}, \cdots, x_{nl})^T \\ &= (x_{1ko}, x_{2ko}, \cdots, x_{nko}, x_{1kc}, x_{2kc}, \cdots, x_{nkc}) \\ &\quad (x_{1lo}, x_{2lo}, \cdots, x_{nlo}, x_{1lc}, x_{2lc}, \cdots, x_{nlc})^T \\ &= (x_{1kc}, x_{2kc}, \cdots, x_{nkc})(x_{1lc}, x_{2lc}, \cdots, x_{nlc})^T \\ &= \sum_i x_{ikc} \cdot x_{ilc} \end{aligned} \tag{7}$$

で定義される。$\cos\theta$が1に近ければ近いほど、記事間の類似度が高い。

4 Latent Semantic Indexing

文献[8]-[10]のベクトル空間法とともに、情報検索で広く使われているLatent Semantic Indexingが文献[12]で提案されている。この方法について、以下に述べる。データベースからXという$w \times d$次元の行列(行と列はそれぞれ単語と記事に対応している。)を作って、特異値分解(式(8))を行う。特異値の高い単語だけを抽出して次元削減を行う。行列X, W_0, S_0, D_0の次元削減を行った行列をそれぞれ$\hat{X}, W, S, D$とすると、式(9)が成り立つ。従って、記事iと記事jの類似度は、式(9)より$\sum_k DS(i_k) \cdot DS(j_k)$に表される。ここで、$DS(i_k)$は記事$i$の$k$番目の要素であり、行列DSの第$i$行$k$列の要素を表している。

$$X = W_0 S_0 D_0^T \tag{8}$$

$$\begin{aligned} \hat{X}^T\hat{X} &= DS^T W^T W S D^T \\ &= DS^2 D^T \end{aligned} \tag{9}$$

式(9)において、W_0とD_0はそれぞれ左特異ベクトル行列、右特異ベクトル行列と呼ばれる直交行列であり、Sは特異値を対角成分に持つ行列である。行列Wは行列Xと行(単語)の次元は同じであり、列(記事)の次元は行列Xに比べて削減されている。Wで得られた基底は，単語の線形結合で表されており，重要語の基底で新しい概念を表している。これは，潜在的に隠れた(Latent)意味的な(Semantic)概念(Indexing)であることからLSI法と呼ばれている。以下にLSI法によるニュース音声記事の検索の手順について述べる。

1. ニュース音声記事に対してディクテーションを行う。

2. 単語×記事行列Xを作成する。

3. 単語×記事行列XにSVDを適用する。

4. $\sum_k DS(i_k) \cdot DS(j_k)$ を使って、データベースから関連記事を抽出する。

5 単語空間法

我々はベクトル空間法[8]-[10]において、相互情報量とTF-IDFを式(6)のように組み合わせると、検索精度が向上することを確かめた(表3)。これは、3.2.4節で考察したように、IDFと相互情報量とは、互いに違った観点からトピックへの依存度を表しており、TFはトピックにおける単語の発生頻度を表している。これらを統合することによって精度が向上することから、これらは互いに補完し合っていると考えられる。この考えをおし進めると、相互情報量とTF、IDFの3つを基底と考えた空間上に、単語を配置することができる。この空間をMutual-TF-IDF 空間と呼ぶ。

それぞれの単語は、記事毎に異なるTFや相互情報量の値を持つことから、この空間上に点集合として配置され(図1)、単語間の距離は式(10)のように表すことができる。式(10)で表される距離は、単語x_iとx_jが同じ記事に出現したとき、その記事t_mに対するそれぞれの単語の重要性の違いを距離として計算し、この距離を全記事で平均している。我々はこの距離を単語間の概念距離と考えている。

一方、記事X_kとX_l間の類似度は、単語間距離$WD(x_i, x_j)$を考慮した内積として、式(11)で表される。通常のベクトル空間法やLSI法では、記事間の単語のオーバーラップ率でしか、その類似度を表現していない。しかし、式(11)を用いると、記事間の単語のオーバーラップ率だけでなく、記事間の単語類似度を用いて、記事間の類似度を表現できるので、検索精度が向上すると考えられる。式(11)に基づくベクトル空間法を単語空間法と呼ぶ。

$$
\begin{aligned}
WD(x_i, x_j) &= \frac{1}{m}\sum_m ((TF(x_i, t_m) - TF(x_j, t_m))^2 \\
&\quad +(IDF(x_i) - IDF(x_j))^2 \\
&\quad +(i(t_m; x_i) - i(t_m; x_j))^2)^{\frac{1}{2}} \qquad (10) \\
(X_k, X_l) &= \sum_i \sum_j x_{ik} \times x_{jl} \times \frac{1}{WD(x_i, x_j)} \qquad (11)
\end{aligned}
$$

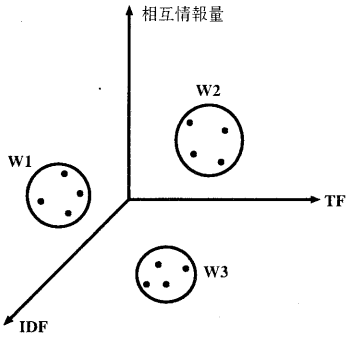

図 1: Mutual-TF-IDF空間における単語空間

6 実験結果

用いたコーパスは1998年8月のニュース記事であり、58記事ある。検索実験においてはジャックナイフ法を用いた。58記事から1つの記事を抽出し、それをクエリーとして残りの記事を検索するという実験を58回繰り返した。クエリーに対して、検索される記事の類似度が閾値よりも高ければ、それを検索記事として出力し、低ければリジェクトする。用いた評価方法は再現率と適合率であり、これらはトレードオフの関係にある。従って、本稿における手法の比較は、再現率と適合率がほぼ等しくなるポイントで評価した。

2で実験したディクテーション結果に対して、記事検索の実験を行った。実験の結果を表3に示す。表における Mutual, TF-IDF, Frq-TF-IDF, Mutual-TF-IDF はそれぞれ、ベクトル空間法に単語重要度決定方法として相互情報量、TF-IDF、単語の頻度分布を用いたTF-IDF、TF-IDFを考慮した相互情報量を適用した結果である。また、Word-Space は単語空間法であり、単語重要度決定方法としてはTF-IDFを考慮した相互情報量を使っている。

表3に示されるように、単語重要度決定方法では、TF-IDF を考慮した相互情報量が一番いい結果を示している。相互情報量、TF-IDF、頻度分布に基づくTF-IDFは、テキストに比べてディクテーションでは精度が悪くなっているのに対し、TF-IDF を考慮した相互情報量は、精度が良くなっている。このことより、TF-IDF を考慮した相互情報量は、ディクテーションにおける湧き出し誤りを解決する手法として有効であることがわかる。

また、従来のベクトル空間法, LSI法に比べて、本研究の提案手法である単語空間法が有効であることがわかる。これは、5で考察したように、ベクトル空間法, LSI法が記事間の単語のオーバーラップ率でしか、その類似度を表現していないのに対して、単語空間法は単語類似

度を用いて、記事間の類似度を表現できたためであると考えられる。ディクテーションと比べ精度の差異がないものの、テキストに対しては4%(54.71%→58.49%)向上した。ディクテーションと比べ精度の差異がない原因は、ディクテーションして得られた結果において、記事の内容に関係のある単語の欠落がほとんどなかったためである。テキストに対して4%向上した原因は、概念類似度を表す尺度として単語空間法が有効だったためである。

表 3: 実験結果

	テキスト	ディクテーション
Mutual	**49.05**	**49.05**
TF-IDF	**56.00**	**52.83**
Frq-TF-IDF	**54.71**	**52.83**
Mutual-TF-IDF	**54.71**	**58.49**
LSI	**54.71**	**49.05**
Word-Space	**58.49**	**58.49**

7 おわりに

本稿では従来の単語重要度決定方法の問題点を解決する手法として、TF-IDFを考慮した相互情報量を提案した。また、従来のベクトル空間法、LSI法の問題点を解決する手法として、単語空間法の提案を行った。TF-IDFを考慮した相互情報量は、ディクテーションにおける湧き出し誤りの問題を解決する手法として、有効であることがわかった。単語空間法は、記事間の単語のオーバーラップ率と単語間の類似度を統合して、記事間の比較を行うことができることから、検索精度を向上できることがわかった。

しかし、まだ、検索精度は再現率 58%、適合率 58% であり、改善の余地があり、今後の課題としてあげられる。

参考文献

[1] Larry Gillick, Yoshiko Ito, Linda Manganaro, Michael Newman, Francesco Scattone, Steven Wegmann, Jon Yamron, Puming Zhan: "Dragon System's Automatic Transcription of New TDT Corpus", Proceedings of Broadcast News Transcription and Understanding Workshop, Lansdowne, Virginia, Feb. 1998.

[2] J.P.Yamron, I.Carp, L.Gillick, S.Lowe, and P.van Mulbregt: "A Hidden Markov Model Approach to Text Segmentation and Event Tracking", ICASSP98, Volume I, pp.333-336, 1998.

[3] P.van Mulbregt, I.Carp, L.Gillick, S.Lowe and J.Yamron: "Text Segmentation and Topic Tracking on Broadcast News Via A Hidden Markov Model Approach", ICSLP98, Volume VI, pp.2519-2522, 1998.

[4] S.Dharanipragada, M.Franz, J.S.McCarley, S.Roukos, T.Ward: "Story Segmentation and Topic Detection for Recognized Speech", Eurospeech99, Volume VI, pp.2435-2438.

[5] Hubert Jin, Rich Schwartz, Sreenivasa Sista, Frederick Walls: "Topic Tracking for Radio, TV Broadcast, and Newswire", Eurospeech99, Volume VI, pp.2439-2442.

[6] Masayuki Nakazawa, Jianxin Zhang and Ryuichi Oka: "Topic Spotting and Its Description of Summary from Spontaneous Speech", Eurospeech99, Volume VI, pp.2447-2450.

[7] Frederick Walls, Hubert Jin, Sreenivasa Sista and Richard Schwartz: "Topic Detection in Broadcast News", Eurospeech99, Volume VI, pp.2451-2454.

[8] G.Salton, A.Wong and C.S.Yang, "A Vector Space Model for Information Retrieval", Journal of the ASIS, pp.613-620, November 1975.

[9] G.Salton and M.J.McGill: "Introduction to Modern Information Retrieval", McGraw-Hill, 1983.

[10] G.Salton: "Automatic Text Processing", Addison-Wesley, Reading, Massachusetts, 1989.

[11] S.Furui, K.Takagi, A.Iwasaki: "Japanese Broadcast News Transcription and Topic Detection", DARPA98, 1998.

[12] Scott Deerwester, Susan T.Dumais, George W.Furnas, Thomas K.Landauser, and Richard Harshman: "Indexing by Latent Semantic Analysis", in Journal of the Amercan Society for Information Science, Vol.41-6, pp.391-407, 1990.