ベクトル空間圧縮モデルによるWWW検索処理の効率化

3T－6

原田 晃史　川越 恭二
立命館大学　理工学部

1　はじめに

インターネットの普及によって急激に増大したWeb集合から必要な情報を効率的に検索することは困難である。このため、効率的な手法のアプローチとして、ベクトル空間モデルによる手法が用いられているが、ベクトルの次元が大きくなるという問題がある。この問題に対して、特異値分解を適用してベクトルの次元を減少させる方法があるが、この方法を用いても数十から数百までしか減少しない。

本研究では、ベクトル空間モデルを圧縮する手法を提案し、このモデルを用いて検索効率の向上を図る。

2　ベクトル空間モデル

2.1　概要

ベクトル空間モデルでは、検索条件とドキュメントの両方を、キーワードを軸とした多次元空間におけるベクトルとみなす。予め、ドキュメント集合中のキーワードをすべて列挙し、その各々について各ドキュメントとの適合度を算出し、そのドキュメントのベクトルとする。

2.2　問題点とその対策

このモデルにおける問題点は、個々のベクトルの次元がN次元と高くなることである。ここでNとは、キーワードの総数である。このモデルを行列で表現すると一般に疎行列となることから、本研究では、ドキュメントに関連のあるキーワードだけをベクトルの要素とすることによって、ベクトルの次元を圧縮するモデルを提案する。即ち、n_i：ドキュメントのキーワード数、$h^i(p)$：正規直交系をなす単位ベクトル、w_p^i：キーワードの適合度とすると、ドキュメントのベクトルは次のように定義できる。

$$\vec{d}_i = \sum_{p=1}^{n_i} w_p^i h^i(p) \qquad (1)$$

3　ベクトル空間圧縮モデルによる効率的な検索手法

3.1　主なアイデア

検索の効率を向上させるために、格納すべきページを次のように定義する。即ち、類似関数を用いてベクトルの類似度によってクラスタを生成し、更にその類似度によってクラスタ内を分類し、ページを割り当てる。ここで、クラスタの代表元のことを基準ベクトルと呼ぶことにする。

このように分類することによって、検索時は、問い合わせベクトルと基準ベクトルとの比較のみによって目的のページにたどり着ける。しかし、このままでは検索時に問い合わせベクトルと基準ベクトルとの総当りになり、検索効率がよくない。そこで、キーワードと基準ベクトルを格納している表へのポインタとを保持する表（以下、基準ベクトル対応表）と、基準ベクトルとそれとの類似度による割り当てページへのポインタとを保持する表（以下、ページ対応表）を作成し、検索効率の向上を図る。

3.2　諸定義

以下に、本研究で用いる概念を定義する。

定義1: ベクトルの内積

ベクトルの表現を(1)式のように圧縮して表現すると、ベクトル間の内積は次のように表すことができる。

$$\vec{d}_i \circ \vec{d}_j = \sum_{p=1}^{n_i}\sum_{q=1}^{n_j} w_p^i h^i(p)\cdot w_q^j h^j(q) \qquad (2)$$

$$s.t.\begin{cases} h^i(p)\cdot h^j(q)=1 & if\ h^i(p)=h^j(q) \\ h^i(p)\cdot h^j(q)=0 & if\ h^i(p)\neq h^j(q) \end{cases}$$

定義2: ベクトルの類似関数

類似関数として、コサイン関数を用いる。ベクトルのノルムをその要素数の平方根と考え、そしてベクトル間の内積を(2)式で定義すると、ベクトル間の類似関数SIMは次のように表すことができる。

$$SIM(\vec{d}_i,\vec{d}_j)=\frac{\vec{d}_i\circ\vec{d}_j}{\|\vec{d}_i\|\cdot\|\vec{d}_j\|}=\frac{\vec{d}_i\circ\vec{d}_j}{\sqrt{n_i}\cdot\sqrt{n_j}} \qquad (3)$$

定義3: ベクトルの距離関数

2ベクトルを含む平面における単位円と2ベクトルとの交点のユークリッド距離とする。定義2より類似関数としてコサイン関数を用いているので、ベクトル間の距離関数は次のように表すことができる。

$$D(\vec{d}_i,\vec{d}_j)=\sqrt{2\cdot(1-SIM(\vec{d}_i,\vec{d}_j))} \qquad (4)$$

定義4: 基準ベクトル

クラスタの代表元のことを基準ベクトルとする。即ち、あふ

The efficient WWW retrieval process based on
vector space compression model
Koji HARADA, Kyoji KAWAGOE
Ritsumeikan University
1-1-1,Noji-Higashi,Kusatsu,Shiga, 525-8577,Japan

れページ内の任意のベクトルにおいて、その他のベクトルとの距離の総和が最小となるものが、クラスタの重心に最も近いベクトルとなるので、この条件を満たすベクトルを基準ベクトルとして選ぶ。ベクトル間の距離関数を(4)式で定義すると、基準ベクトルを導出する式は次のように表すことができる。

$$\vec{e}_k = \vec{d}_i \ s.t. \ \min_i\left(\sum_{j=1}^{m} D(\vec{d}_i, \vec{d}_j)\right) \qquad (5)$$

定義5: 拡張基準ベクトル

ベクトルの要素集合から1つ選び、それを基準ベクトルの要素として加えたものを拡張基準ベクトルとする。即ち、オーバーフローページのベクトル要素集合をN^f, $\exists w_p^i h^i(p) \in N^f$, $h^k(q) \neq h^i(p)$ for $\forall h^k(q), q = 1 \cdots n_k$, かつ出現回数の少ないものとすると、拡張基準ベクトルは次のように表せる。

$$\tilde{e}_k = \vec{e}_k + w_p^i h^i(p) = \sum_{q=1}^{\tilde{n}_k} w_q^k h^k(q) \qquad (6)$$
$$\left(\because w_{n_k+1}^k h^k(n_k+1) = w_p^i h^i(p), \tilde{n}_k = n_k + 1\right)$$

3.3 検索方法

初めに、問い合わせベクトルの各要素と基準ベクトル対応表を比較して、基準ベクトルを確定する。ここで、基準ベクトルが確定しない場合は、検索対象ページをあふれページとする。

次に、問い合わせベクトルと基準ベクトルとの類似度を算出し、その値を用いてページ対応表を調べ、その類似度以上のページを検索対象ページとし、そのページ内の任意のベクトルについて、問い合わせベクトルを完全に満たすものを出力する。

3.4 格納方法

初めに、格納ベクトルの各要素と基準ベクトル対応表を比較して、基準ベクトルを確定する。ここで、基準ベクトルが確定しない場合は、格納対象ページをあふれページとする。

次に、格納ベクトルと基準ベクトルとの類似度を算出し、その値を用いてページ対応表を調べ、格納対象ページを確定する。

ここで、格納対象ページがページ内最大ベクトル数（ベクトルの許容数）を満たしていない場合は、そのページにベクトルを格納する。格納対象ページがベクトルの許容数を満たしている場合には、次の3つの方法でページを分割して、ベクトルを格納しなおす。

1. 格納対象ページがあふれページのときは、(5)式から基準ベクトルを求め、$\forall \vec{d}_f \in \Omega$ において $SIM(\vec{e}_k, \vec{d}_f) \succ 0$ を満たすものを新しいページに格納する。ここで、Ω: オーバーフローページ内の任意のベクトル集合とする。
2. 基準ベクトルとの類似度が最も高いページのときは、$\forall \vec{d}_f \in \Omega$ において $SIM(\vec{e}_k, \vec{d}_f) \succ \delta$ を満たすものを新しいページに格納する。ここで、Ω: オーバーフローページ内の任意のベクトル集合、δ: 指標となる類似度とする。
3. 上記以外のページのときは、(6)式から拡張基準ベクトルを求め、$\forall \vec{d}_f \in \Phi$ において $SIM(\tilde{e}_k, \vec{d}_f)$ を算出して、ページを割り当てする。ここで、Φ:オーバーフローページを含むクラスタの任意のベクトル集合とする。

4 評価実験

4.1 実験内容

実験に用いたドキュメント数は100、キーワード数は260とし、この中から6つをランダムにベクトルに割り当てた。この割り当て方で3つのcaseを与える。これは、ベクトル全体の関連性の強度を順に高めている。また、ドキュメントとキーワードの適合度を1、ページのベクトル許容数を5とする。なお、評価項目については、ページ空間効率と平均ページアクセス回数とする。

4.2 実験結果

以上の条件において、シミュレーションを行った結果、3つのcaseの平均ページ空間効率は約76%となった。また、図1に示すような平均ページアクセス回数を得た。

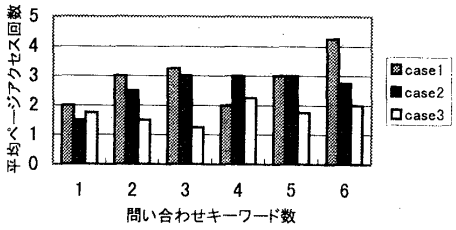

図1：平均ページアクセス回数

一般に、問い合わせのキーワード数が増加するとページアクセス回数は線形に増加するものと考えられる。しかし、提案モデルを用いた手法でのページアクセス回数は、問い合わせキーワード数が増加しても線形には増加せず、一定範囲内に収まる。また、ベクトル全体の関連性の強度が高い情報集合ほど、検索に要するページアクセス回数が減少する。従って、提案モデルを用いた手法は、問い合わせに含まれるキーワード数が多い場合の検索や、ベクトル全体の関連性の強度が高い情報集合に対する検索に適しているものと思われる。

5 おわりに

今後の課題としては、ドキュメントとキーワードの適合度を重み付けすることや適合性フィードバックの概念を取り入れること等が挙げられる。

参考文献

[1] Christos Faloutsos: " SERCHING MULTIMEDIA DATABASES BY CONTENT ", Kluwer Academic Publishers, 1998