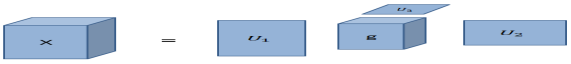
テンソル分解を用いた情報検索の次元縮小効果

横林 亮平† 三浦 孝夫†
† 法政大学理工学部創生科学科

1. 前書き

多次元データは記憶域量が膨大であり，情報検索操作の実行が高価であるため実用化されることは少ない．2 次元データでの記憶量を減少する方法として潜在意味索引付(LSI)が知られている．LSI はモデルが理論的に簡単であり縮小効果が大きく，特異値を対角行列として大きな順に並べ小さな固有値の影響を省くことにより近似復元可能となる．これに対して多次元テンソルでは，Tucker 分解[1]により，コア(g)とファクタ(U)に分解するが，コアは対角ではない．

本実験では，テンソル分解におけるファクタの寄与率を用いたコアの次元縮小手法を提案する．Tucker 分解を 3 次元で考えると以下で表される．



分解で得るコア g は一般に対角ではないため 3 つの問題が存在する．第 1 は，必要とするメモリサイズがファクタの線形和とならないので，記憶量の見積もりが立たない．第 2 は，テンソルの次元縮小と情報検索に手間がかかる．最後に，潜在意味を用いた語の検索が直接的に表現できない．

2. テンソル分解による次元縮小

情報検索精度を保ったまま記憶域量を減少させ計算や検索処理を向上する手法としてテンソル分解を用いる．本研究では寄与率を用いた次元縮小を行う．即ち，主要な寄与率部分を残して次元を削減し，次元縮小前後のコアに質問ベクトルを与え検索方式を定義する。検索に対して，次元縮小後のコアが検索精度を保持したまま削減できれば，元のテンソルより小さなサイズのテンソルで操作が可能となる．本稿では，縮小後の検索精度を実験で確認する．

3. 実験

テンソル分解を行い,コアとファクタを求める．ファクタの寄与率を用いてコアの次元縮小を行う．質問ベクトルを与えることで次元縮小による検索精度を確認する．その結果からコアに有用性があるかを検証する．

テキストデータには Reutes Corpus Volume 1(RCV1)の 19,960,821 件に含まれる 750 記事を使用． TreeTagger を用いて原型に直し，さらに不要語の消去を行う． 処理した記事から全単語を抽出し重複を消去する．得られた単語は 11,169 語となる．低頻度(頻度 38 以上)の語である 424 語を,単語かつ索引語とみなす．カテゴリは，記事中に記されているものを使用する．本記事中に存在するカテゴリ数は 86 個である．

次元縮小前のコアから正解を与える．コアに関して質問ベクトル$\vec{q}$と記事ベクトル$\vec{d} \in X$の余弦(cos)類似度を測定し，この類似度が 0.5 以上のものをすべて正解とする．同様に次元縮小後のコアに対しても類似度を求める．このとき質問ベクトルにも次元合わせを行い$\overrightarrow{q_0} = {U_1}'^T \times {U_3}'^T \times \vec{q}$となる．質問ベクトル$\overrightarrow{q_0}$と記事ベクトル$\overrightarrow{d_0}$の余弦類似度 0.5 以上を抽出する．

次元縮小前のコア(424×750×86)での余弦類似度 0.5 以上は 521 件であり，これを検索の正解とする．それに対して寄与率 9 割，8 割，7 割での余弦類似度と適合率を以下の表に示す．

	単語数	カテゴリ数	cos0.5 以上	正回数	適合率
縮小前	424	86	521		
寄与率9割	258	50	549	521	94.9
寄与率8割	184	35	605	521	86.1
寄与率7割	134	25	706	519	73.5

縮小前後の余弦類似度と適合率

寄与率制限を設けて次元縮小すると，寄与率 9 割，8 割，7割となるにつれて単語数，カテゴリ数が減少する．寄与率7割では単語 31.6%，カテゴリ 29.1%にまで次元縮小される．次元が圧縮されても正解数は保持しており，ノイズのみが増加している．従って，次元縮小しても検索精度は保持されている．ノイズには余弦類似度が 0.51 に満たない低位な記事が多く含まれる．これはプログラムでの計算誤差によるものと考えられる．

4. 結論

本研究では，情報検索精度を保ったまま記憶域量を減少させ，計算や検索の処理を向上させる手法としてテンソル分解を用いた．70%程度の次元縮小を行ってもその特徴は保持したままである． コアは密テンソルであるが， 3 次元の構造を保持しているためモデル化が容易であり，テンソル分解による次元縮小後のコアを利用した検索は有効的である．

参考文献

[1] Tamara G.Kolda and Brett W.Bader, "Tensor Decompositions and Applications", SAND2007-6702 Unlimited Release Printed November 2007