김일성종합대학학보
(자연과학)
주체103(2014)년 제60권 제8호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 60 No. 8 JUCHE103(2014).

GMM언어모형구축을 위한 대규모성긴2－그람빈도 행렬의 특이값분해실현의 한가지 방법

리현순, 리혁철

지난 시기 제기된 고전적인 행렬의 특이값분해(SVD)[1]는 입력행렬에 직교변환을 직접 적용하여 실현하였다.

이러한 특이값분해방법들은 차수가 수만～수십만의 대규모행렬인 경우에는 계산량이 많은것으로 하여 적용불가능하다.[2]

GMM언어모형구축을 위한 2－그람빈도행렬은 차수가 매우 큰 대규모성긴행렬이다. 따라서 음성인식을 위한 GMM언어모형구축에서는 수만차수의 2－그람빈도행렬을 특이값분해하는데 적용할수 없다.

GMM언어모형구축을 위한 수만차수의 2－그람빈도행렬을 특이값분해하기 위해서는 이 행렬을 효과적으로 표현하는 문제가 제기된다.

이로부터 론문에서는 대규모 2－그람빈도행렬의 성김특성을 리용하여 대규모행렬을 효과적으로 표현할수 있는 행렬압축표현방법을 제기하고 2－그람빈도행렬의 특이값분해를 진행하는 방법을 제안한다.

1. 성김특성을 가진 대규모2－그람빈도행렬의 압축표현 및 특이값분해방법

먼저 성김특성을 가진 대규모 2－그람빈도행렬의 압축표현방법에 대하여 보자.

GMM언어모형구축을 위한 특이값분해행렬은 2－그람빈도행렬로서 어휘의 수를 V(수만～수십만개)라고 할 때 차수가 $V \times V$ 인 대규모성긴행렬이다.

대규모 2－그람빈도행렬에서 령빈도들을 포함하여 행렬의 모든 원소들을 보관하는것은 기억랑비이며 비효률적이다.

따라서 령아닌 원소만을 압축된 렬보관형태로 화일에 보관하는 방법으로 대규모 2－그람빈도행렬을 표현한다.

여기로부터 대규모 2－그람빈도행렬표현을 위한 자료구조를 3개의 자료배렬 즉 령아닌 원소들만을 렬별로 보관하는 실지의 자료값배렬(value), 자료값배렬의 매 원소들에 대한 행번호를 보관하는 행첨수배렬(rowind), 압축된 형식의 렬지시자를 보관하는 렬지시자배렬(pointr)로서 표현한다.

대규모 2－그람빈도행렬 $A(i,\ j)$ 를 표현하는 3개의 자료배렬사이에는 다음과 같은 관계가 있다.

$\text{value}(k) = a_{ij}$ 라고 하면 $\text{rowind}(k) = i$ 이고 pointr는 $\text{pointr}(j) \le k < \text{pointr}(j+1)$ 을 만족시키는 값이다.

여기서 a_{ij}는 대규모 2－그람빈도행렬의 i번째 행, j번째 렬의 원소이다.

우와 같은 관계로부터 pointr배렬의 첫 요소값은 0과 같으며 마지막 요소값은 대규모 2－그람빈도행렬의 령아닌 원소개수와 같다는것을 알수 있다.

한편 pointr 배렬의 매 요소는 대규모 2－그람빈도행렬의 령아닌 원소들의 렬별시작번호를 나타낸다.

즉 $\text{pointr}(j) \le k < \text{pointr}(j+1)$을 만족시키는 $\text{pointr}(j+1) - \text{pointr}(j)$개의 실지 자료값배렬 요소들 $\text{value}(k)$는 대규모 2－그람빈도행렬의 j번째 렬의 령아닌 원소들을 표현한다.

실지 자료값배렬(value)과 행첨수배렬(rowind)의 크기는 대규모 2－그람빈도행렬의 령아닌 원소개수와 같으며 렬지시자배렬은 대규모 2－그람빈도행렬의 렬개수보다 하나 큰 값을 가진다.

론문에서는 대규모 2－그람빈도행렬의 령아닌 원소들에 대한 압축된 렬보관형식의 표현을 2－그람언어모형구축모듈을 리용하여 진행한다.

이 모듈은 본문코퍼스로부터 2－그람들의 빈도수를 계산하여 2－그람완충기에 보관하고 2－그람빈도화일을 생성하며 이때 얻어진 2－그람완충기를 대규모 2－그람빈도행렬을 표현하는데 리용한다.

2－그람완충기를 리용한 대규모 2－그람빈도행렬의 압축표현알고리듬은 다음과 같다.

① 본문코퍼스로부터 한 단어씩 읽어서 다음의 처리를 반복 수행한다.

우선 읽어들인 단어가 이미 등록되여있는가를 검사하고 등록되여있으면 그에 대한 식별번호(id)를 돌려주고 없으면 그 단어의 식별번호를 새로 창조하여 단어사전에 등록하고 돌려준다.

다음 밀기등록기를 리용하여 왼쪽밀기방법으로 2－그람을 완성하고 그것을 거꿀순서로 배치하여 2－그람완충기에 보관하는데 그것은 대규모 2－그람빈도행렬의 렬별압축을 진행하는데 편리하게 하기 위해서이다.

② 2－그람완충기를 단어사전에 등록된 순서대로 정렬시키는데 이때 얻어지는 2－그람완충기는 대규모 2－그람빈도행렬의 령아닌 원소들에 대한 렬별정렬형태를 표현한다.

③ 얻어진 2－그람완충기를 리용하여 행렬의 정보(행수, 렬수, 령아닌 원소개수) 등을 보관한다. 여기서 행수와 렬수는 등록된 단어개수와 같으며 령아닌 원소개수는 2－그람완충기에 들어있는 2－그람개수와 같다.

④ 렬별로 정렬된 2－그람완충기의 매 요소값들을 보관하여 실지 자료값배렬(value)을 만든다.

⑤ 2－그람완충기의 매 요소값들의 행별첨수값을 보관하여 자료값배렬(value)의 매 요소에 대한 행번호를 표현하는 행첨수배렬(rowind)을 만든다.

⑥ 2－그람완충기의 정렬된 렬별첨수값을 리용하여 해당 렬번호로 시작되는 2－그람개수들을 루계하여 렬별지시자배렬 pointr 를 만든다.

다음으로 압축표현된 대규모 2－그람빈도행렬의 특이값분해를 보자.

우와 같은 방법으로 압축된 렬보관형태로 표현된 3개의 자료배렬(pointr, rowind, value)을 읽어들여 Lanczos알고리듬에 따라 지정된 개수의 안정된 특이값과 특이벡토르를 계산한다. 즉 대규모성긴 2－그람빈도행렬에 대한 특이값분해에 실대칭행렬에 대한 Lanczos 재귀방법을 적용하여 본래의 대규모성긴행렬과 가장 류사한 위수가 R인 근사특이벡토르행렬을 얻는 방법으로 특이분해를 실현한다.

2. 실 험 결 과

론문에서는 GMM언어모형구축을 위한 학습자료로서 형태부단위로 분할된 품사표식이 붙은 120만개의 문장으로 이루어진 본문자료를 리용하였다.

우의 학습자료에 대하여 구축된 GMM언어모형을 위한 2－그람빈도행렬은 차수가 80 000×80 000인 대규모성긴행렬로서 이 행렬의 밀도는 대략 0.25%이다.

대규모성긴2－그람빈도행렬에 대하여 령아닌 원소들만을 압축된 렬보관방법으로 10MB로 압축표현하고 특이값분해를 진행하여 120개의 특이벡토르를 생성하였다.

우의 행렬을 특이값분해하는데 Pentium 4(2.6GHz CPU, 512M RAM)콤퓨터에서 2h정도 걸렸다.

맺 는 말

압축된 렬보관형식의 행렬표현방법으로 대규모성긴2－그람빈도행렬을 압축표현하는 방법을 제안하고 2－그람언어모형구축모듈을 리용하여 령아닌 원소들만을 렬별압축형태로 보관함으로써 대규모 2－그람빈도행렬을 손쉽게 표현하였다. 또한 실대칭행렬에 대한 Lanczos알고리듬을 리용하여 GMM언어모형구축을 위한 대규모성긴2－그람빈도행렬의 특이값분해를 실현하였다.

참 고 문 헌

[1] M. W. Berry; Int. J. Supercomp. Appl., 6, 12, 1992.

[2] N. Bassiou et al.; Computer Speech and Language, 25, 31, 2011.

주체103(2014)년 4월 5일 원고접수

A Method on the Singular Value Devision Completation of a Large-Scale Sparse 2-Gram Count Matrix for Constructing a GMM Language Model

Ri Hyon Sun, Ri Hyok Chol

We propose the method to represent efficiently a large-scale 2-gram count matrix by using one's sparseness and implement its singular value division for constructing a GMM language model with Lanczos algorithm on a real symmetry matrix.

Key words: GMM language model, singular value division