



دانشگاه صنعتی امیرکبیر  
(پلی تکنیک تهران)  
دانشکده ریاضی و علوم کامپیوتر

پایان نامه کارشناسی ارشد  
گرایش سیستم‌های کامپیوتری

کاهش بعد داده‌های بزرگ مقیاس با استفاده از نگاشت  
تصادفی

نگارش  
سیامک دهد

استاد راهنما  
دکتر عادل محمدپور

استاد مشاور  
دکتر هادی زارع

دی ۱۳۹۷

## صفحه فرم ارزیابی و تصویب پایان نامه - فرم تأیید اعضاء کمیته دفاع

در این صفحه فرم دفاع یا تأیید و تصویب پایان نامه موسوم به فرم کمیته دفاع- موجود در پرونده آموزشی- را قرار دهید.

### نکات مهم:

- نگارش پایان نامه/رساله باید به **زبان فارسی** و بر اساس آخرین نسخه دستورالعمل و راهنمای تدوین پایان نامه های دانشگاه صنعتی امیرکبیر باشد.(دستورالعمل و راهنمای حاضر)
- رنگ جلد پایان نامه/رساله چاپی کارشناسی، کارشناسی ارشد و دکترا باید به ترتیب مشکی، طوسی و سفید رنگ باشد.
- چاپ و صحافی پایان نامه/رساله بصورت **پشت و رو(دورو)** بلامانع است و انجام آن توصیه می شود.



دانشگاه صنعتی امیرکبیر  
(پلی تکنیک تهران)

## تعهدنامه اصالت اثر

تاریخ: دی ۱۳۹۷

اینجانب **سیامک دهبید** متعهد می‌شوم که مطالب مندرج در این پایان‌نامه حاصل کار پژوهشی اینجانب تحت نظارت و راهنمایی اساتید دانشگاه صنعتی امیرکبیر بوده و به دستاوردهای دیگران که در این پژوهش از آنها استفاده شده است مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذکر گردیده است. این پایان‌نامه قبلاً برای احراز هیچ مدرک هم‌سطح یا بالاتر ارائه نگردیده است.

در صورت اثبات تخلف در هر زمان، مدرک تحصیلی صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پیگیری قانونی خواهد داشت.

کلیه نتایج و حقوق حاصل از این پایان‌نامه متعلق به دانشگاه صنعتی امیرکبیر می‌باشد. هرگونه استفاده از نتایج علمی و عملی، واگذاری اطلاعات به دیگران یا چاپ و تکثیر، نسخه‌برداری، ترجمه و اقتباس از این پایان‌نامه بدون موافقت کتبی دانشگاه صنعتی امیرکبیر ممنوع است. نقل مطالب با ذکر مآخذ بلامانع است.

**سیامک دهبید**

امضا

# سپاسگزاری

با تشکر از استاد گرامی دکتر محمدپور بابت همراهی و صبر ایشان

سیامک دهب  
دی ۱۳۹۷

## چکیده

روش تصویر تصادفی برای کاهش بعد داده‌های بزرگ مقیاس مزایای متعددی نسبت به روش‌های دیگر کاهش بعد دارد. در این پایان‌نامه این روش برای داده‌های بزرگ مقیاس با دیگر روش‌های کاهش بعد مقایسه شده است. همچنین توانایی این روش برای داده‌های با توزیع پایدار غیر نرمال با دیگر روش‌های کاهش بعد مقایسه شده است.

## واژه‌های کلیدی:

کاهش بعد، تصویر تصادفی، توزیع پایدار، داده‌های بزرگ مقیاس

## فهرست مطالب

صفحه

عنوان

۱	مقدمه	۱
۲	۱-۱ مقدمه	۲
۲	۱-۱-۱ مقدمه	۲
۳	۲ مرور ادبیات	۳
۴	۱-۲ مقدمه	۴
۵	۲-۲ داده‌های حجیم	۵
۵	۱-۲-۲ داده‌های حجیم وب	۵
۷	۲-۲-۲ جریان‌های داده‌ی حجیم	۷
۷	۳-۲ چالش‌های نمونه‌گیری از داده‌های حجیم	۷
۸	۱-۳-۲ مزایای نمونه‌گیری تصادفی مختصات	۸
۸	۲-۳-۲ معایت نمونه‌گیری تصادفی مختصات	۸
۹	۴-۲ تصویر تصادفی پایدار	۹
۹	۵-۲ کاهش بعد	۹
۹	۶-۲ کاهش بعد	۹
۹	۱-۶-۲ بارگیری مراجع	۹
۱۰	منابع و مراجع	۱۰
۱۲	پیوست	۱۲
۱۳	واژه‌نامه‌ی فارسی به انگلیسی	۱۳
۱۵	واژه‌نامه‌ی انگلیسی به فارسی	۱۵

صفحه

## فهرست اشکال

شکل

## فهرست جداول

صفحه

جدول

- ۱-۲ تعداد بازدید صفحات برای کلمات با بازخورد بالا و کلمات با بازخورد نادر . . . . . ۵
- ۲-۲ با افزایش تعداد عبارات در درخواست، باید فرکانس‌های جفت شده کاهش پیدا کنند. ولی
- ۶ تخمین‌های بیان شده توسط موتورهای جستجو گاهی این موضوع تثبیت شده را نقض می‌کنند.



## فهرست نمادها

نماد	مفهوم
$\mathbb{R}^n$	فضای اقلیدسی با بعد $n$
$\mathbb{S}^n$	کره $n$ یکه بعدی
$M^m$	خمینه $m$ -بعدی $M$
$\mathfrak{X}(M)$	جبر میدان‌های برداری هموار روی $M$
$\mathfrak{X}^1(M)$	مجموعه میدان‌های برداری هموار $1$ یکه روی $(M, g)$
$\Omega^p(M)$	مجموعه $p$ -فرمی‌های روی خمینه $M$
$Q$	اپراتور ریچی
$\mathcal{R}$	تانسور انحنای ریمان
$ric$	تانسور ریچی
$L$	مشتق لی
$\Phi$	۲-فرم اساسی خمینه تماسی
$\nabla$	التصاق لوی-چویتای
$\Delta$	لاپلاسین ناهموار
$\nabla^*$	عملگر خودالحاق صوری القا شده از التصاق لوی-چویتای
$g_s$	متر ساساکی
$\nabla$	التصاق لوی-چویتای وابسته به متر ساساکی
$\Delta$	عملگر لاپلاس-بلترامی روی $p$ -فرم‌ها

# فصل اول

## مقدمه

۱-۱ مقدمه

۱-۱-۱ مقدمه

## فصل دوم

## مرور ادبیات

## ۱-۲ مقدمه

عمومیت پیدا کردن داده‌های حجیم مانند داده‌های حجیم تحت وب و جریان‌های داده بزرگ در کاربردهای جدید، موجب به وجود آمدن فرصت‌های و چالش‌هایی برای مهندسين و دانشمندان شده است. [۱۵] برای مثال، زمانی که ماتریس داده  $A \in \mathbb{R}^{n \times D}$  ابعادی در حد وب داشته باشد، عملیات ساده‌ای مانند محاسبه  $AA^T$  سخت می‌شود. برای ارائه و نگهداری داده‌های حجیم در حافظه‌ای کوچک و برای استخراج اطلاعات آماری اصلی از مجموعه‌ای از بیانی محدود، روش‌های گوناگونی نمونه‌برداری توسعه یافته است. به طور کلی روش تصویر تصادفی پایدار<sup>۱</sup> برای داده‌های با دم سنگین خیلی خوب کار می‌کند.

روش تصویر تصادفی پایدار، ماتریس داده‌های اولیه  $A \in \mathbb{R}^{n \times D}$  را در ماتریس تصادفی  $R \in \mathbb{R}^{D \times k}$  ضرب می‌کند و نتیجه ماتریس  $B = AR \in \mathbb{R}^{n \times k}$  است. معمولاً درایه‌های ماتریس تصادفی  $R$  به صورت i.i.d از یک توزیع  $\alpha$ -پایدار متقارن انتخاب می‌شوند ( $0 < \alpha \leq 2$ ). ما می‌توانیم مشخصه‌های  $l_\alpha$  را در  $A$  بر اساس  $B$  تخمین بزنیم. در مورد حالت  $l_2$  مزیت توزیع تصادفی پایدار توسط لم JL<sup>۲</sup> برجسته شده است. لم JL بیان می‌دارد که کافی است  $k = O(\frac{\log n}{\epsilon^2})$  باشد تا هم فاصله دو به دویی با نرم  $l_\alpha$  در  $A$  را بتوان با ضریب  $1 \pm \epsilon$  از روی ماتریس  $B$  تخمین زد. در تز [۱۵] Ping Li<sup>۳</sup> لمی مشابه لم JL برای  $0 < \alpha < 2$  اثبات شده است. روش تصویر تصادفی پایدار به یک مسئله تخمین آماری کاهش می‌یابد برای تخمین پارامتر مقیاس برای یک توزیع پایدار  $\alpha$  متقارن. این مسئله از این جهت مورد توجه قرار می‌گیرد زیرا ما به دنبال برآوردی می‌گردیم که هم از نظر آماری درست باشند و هم از نظر محاسباتی مقرون به صرفه. برآوردگرهای مختلفی را مطالعه و مقایسه کردیم. شامل میانگین حسابی، میانگین هندسی، میانگین هارمونیک، تقسیم توانی<sup>۴</sup> و برآوردگر حداکثر بزرگنمایی.

در این پایان‌نامه ما به بررسی موارد خاصی از تصویر تصادفی پایدار می‌پردازیم. برای نرم  $l_2$  ارتقای را با استفاده از اطلاعات حاشیه‌ای پیشنهاد می‌کنیم. همچنین برای حالت  $l_2$  می‌توان ماتریس تصویرگر را از یک توزیع زیرگوسی<sup>۵</sup> بسیار کوچکتر به جای توزیع نرمال انتخاب کرد. با در نظر گرفتن محدودیت‌های معقولی می‌توان، از یک توزیع خاص زیرگوسی استفاده کرد. این توزیع شامل  $[-1, 0, 1]$  با احتمالات  $\{\frac{1}{s}, 1 - \frac{1}{s}, \frac{1}{s}\}$  با مقادیر بسیار بزرگی برای  $s$  (به عبارتی، تصویر تصادف خیلی گسسته<sup>۶</sup>) می‌تواند به خوبی تصویر تصادفی نرمال عمل کند. برای حالت نرم  $l_1$  به عبارتی دیگر تصویر تصادفی کوچی<sup>۷</sup> انجام تخمین کاری نسبتاً جذاب است. برای مثال، محاسبه برآوردگر بیشینه درست‌نمایی MLE در این حالت از لحاظ محاسباتی ممکن است. و یک توزیع معکوس گاوسی<sup>۸</sup> برای مدل‌سازی دقیق توزیع MLE بیان شده است.

روش تصویر تصادفی از پراکندگی داده‌ها استفاده نمی‌کند. در حالی که داده‌های بزرگ مقیاس معمولاً بسیار پراکنده هستند. از روش تصویر تصادفی می‌توان برای حل مسائل بزرگ مقیاس در علوم و مهندسی در موتورهای جستجو و سیستم‌های اخذ داده، پایگاه‌های داده، سیستم‌های جریان داده جدید، جبر خطی عددی و بسیاری از کارهای یادگیری ماشین و داده کاوی که شامل محاسبه حجیم فاصله‌ها است، استفاده کرد.

<sup>1</sup> Stable Random Projection<sup>2</sup> Independent and identically distributed random variables<sup>3</sup> Johnson-Lindenstrauss<sup>4</sup> fractional power<sup>5</sup> sub-Gaussian<sup>6</sup> Very sparse random projections<sup>7</sup> Cauchy random projections<sup>8</sup> inverse Gaussian

## ۲-۲ داده‌های حجیم

عبارات زیر از سایت *Information Week* نقل قول شده‌اند<sup>۹</sup>:

- مقدار داده‌ای که توسط کسب و کارها ذخیره می‌شود تقریباً هر ۱۲ تا ۱۸ ماه دو برابر می‌شود.
- پایگاه داده‌ها بیشتر هم زمان شده‌اند. فروشگاه‌های زنجیره‌ای Wall-Marat داده‌های فروش را هر ساعت به روز می‌کند.
- اضافه شدن یک میلیون خط داده اجازه جستجوهای پیچیده‌تری را می‌دهد. شرکت EBay به کارمندان اجازه می‌دهد برای بدست آوردن درکی عمیق‌تر در خصوص رفتار مشتریان در میان داده‌های حراج در بازه‌های زمانی کوتاه جستجو کنند.
- بزرگترین پایگاه داده‌ها توسط، مرکز شتاب‌دهنده خطی استاندارد، مرکز تحقیقات ناسا، آژانس امنیت ملی و ... در ابعادی در محدوده‌ی پتابایت (هزار ترابایت ۱۰<sup>۱۵</sup> بایت)، اداره می‌شوند.

پدیده نو ظهور مجموعه داده‌ای حجیم، چالش‌های محاسباتی در بسیاری کاربردهای علمی و تجاری به وجود آورده است. شامل اختریف‌یک، بیوتکنولوژی، جمعیت‌شناسی<sup>۱۰</sup>، مالی، سیستم‌های اطلاعات جغرافیایی، دولت، دارو، ارتباطات از راه دور، محیط زیست و اینترنت.

## ۱-۲-۲ داده‌های حجیم وب

وب چقدر بزرگ است؟ **جدول ۱-۲** نشان‌دهنده تعداد بازدید صفحات در موتورهای جستجوی امروزی است. به طور تخمینی حدود  $D = ۱۰^{۱۰}$  صفحه‌ی وب را می‌توان بر اساس بازدید دو واژه‌ی بسیار پر کاربرد «A» و «THE» تخمین زد. **جدول ۱-۲** همچنین نشان می‌دهد که حتی کلماتی که به ندرت کاربرد دارند هم تعداد زیادی بازدید دارند.

Query	Google	Bing
A	25,270,000,000	175,000,000
The	25,270,000,000	101,000,000
Kalevala	7,440,000	939,000
Griseofulvin	1,163,000	332,000
Saccade	1,030,000	388,000

جدول ۱-۲: تعداد بازدید صفحات برای کلمات با بازخورد بالا و کلمات با بازخورد نادر

کلماتی با بازخورد معمولی چه میزان بازدید دارند؟ برای جواب این سوال ما به طور تصادفی ۱۵ صفحه از لغتنامه‌ی آموزشی انتخاب می‌کنیم. [۱۱] (لغتنامه‌ای با ۱۰۰,۵۷ کلمه) و اولین کلمه در هر صفحه را مد نظر قرار می‌دهیم. میانه‌ی آماری بر اساس جستجوگر گوگل ۱۰ میلیون صفحه برای کلمه است.

زبان انگلیسی چند کلمه دارد؟ در اینجا عبارتی را از AskOxford.com نقل قول می‌کنیم:

«این بیان میدارد که حداقل یک چهارم میلیون واژه‌ی انگلیسی مستقل وجود دارد. به جز افعال صرفی و کلمات فنی و ناحیه‌ای که توسط OED<sup>۱۱</sup> تحت پوشش قرار نمی‌گیرند یا کلماتی که هنوز به لغتنامه‌های منتشر

<sup>۹</sup><http://www.informationweek.com/news/showArticle.jhtml?articleID=175801775>

<sup>۱۰</sup>demographics

<sup>۱۱</sup>Oxford English Dictionary

شده اضافه نشده‌اند. در صورتی که این موارد هم در نظر گرفته شوند تعداد لغات در حدود سه چهارم میلیون لغت خواهد بود»

بنابراین اگر یک ماتریس «عبارت به سند»  $A \in \mathbb{R}^{n \times D}$  در نظر بگیریم. در ابعاد وب این ماتریس در ابعاد  $n \approx 10^6$  و  $D \approx 10^5$  بزرگ خواهد شد. در اینجا عدد  $(i, j)$  در  $A$  تعداد ظهور واژه  $i$  در سند  $j$  را نشان می‌دهد. کارکردن با ماتریس در این ابعاد بزرگ چالش برانگیز است. برای مثال، شاخص LSI<sup>۱۲</sup> [۷] و یک مدل موضوعی فراگیر، از SVD<sup>۱۳</sup> بر روی ماتریس عبارت به سند استفاده می‌کند. که انجام این عملیات در ابعاد وب قطعاً غیرممکن است.

یک مشکل اصلی در قبال مجموعه داده‌های سنگین، حافظه کامپیوتر است. به این دلیل که ابعاد و سرعت حافظه فیزیکی بسیار رشد کمتری در مقایسه با پردازنده‌ها (CPU) دارد. این پدیده به عنوان دیوار حافظه شناخته می‌شود [۱۶، ۲۰]. برای مثال، هر چند ممکن است تمامی رخدادهای همزمان دوتایی از پیش محاسبه شوند، ولی نگهداری این حجم از داده در حافظه غیر ممکن است. علاوه بر این، گاهی اوقات تخصیص‌هایی با بیش از دو عامل هم اهمیت پیدا می‌کنند زیرا درخواست‌ها ممکن است شامل بیش از دو واژه هم باشند. یک راه حل ممکن این است که یک «نمونه» از  $A$  نگهداری شود و همزمانی‌ها بر اساس این نمونه در حین کار تخمین زده شوند. ما حدس می‌زنیم که این روش توسط موتورهای جستجوی امروزی مورد استفاده قرار می‌گیرد، هر چند که روش واقعی قطعاً جزو اسرار تجاری آن‌ها است.

هر چند که انتظار می‌رود تخمین‌ها سازگار باشند و فرکانس‌های جفت شده باید با افزایش عبارت به درخواست، کاهش پیدا کنند. جدول ۲-۲ نشان می‌دهد که تخمین‌های بیان شده با موتورهای جستجوی فعلی، همیشه سازگار نیستند.

Query	Hits(Bing)	Hits(Google)
America	150,731,182	393,000,000
America & China	15,240,116	66,000,000
America & China & Britain	235,111	6,090,000
America & CHina & Britain & Japan	154,444	23,300,000

جدول ۲-۲: با افزایش تعداد عبارات در درخواست، باید فرکانس‌های جفت شده کاهش پیدا کنند. ولی تخمین‌های بیان شده توسط موتورهای جستجو گاهی این موضوع تثبیت شده را نقض می‌کنند.

با اینکه، تعداد کل واژه‌های انگلیسی (که به‌طور صحیح نوشته شده‌اند) هم اکنون شگفت‌آور است، در بسیاری کاربردهای متن کاوی، ما باید با ابعاد بسیار بزرگتری سر و کار داشته باشیم. در حالی که یک سند ممکن است بیانگر برداری از تک واژه‌ها باشد (به عبارت دیگر، مدل کیسه لغات<sup>۱۴</sup>). معمولاً بهتر است سند به عنوان یک بردار از لغات به صورت  $I$  پیوسته<sup>۱۵</sup> [۴] بیان شود. برای مثال، با استفاده از مدل ۳ پیوسته، جمله‌ی "It is a nice day" به مجموعه‌ی زیر تجزیه می‌شود.  $\{ "it is a", "is a nice", "a nice day" \}$  این مدل به طور چشمگیری ابعاد داده‌ها را افزایش می‌دهد. به خاطر اینکه، اگر مجموعه‌ی  $10^6$  تک لغت انگلیسی موجود داشته باشد. مدل ۳ پیوسته تعداد ابعاد را از  $10^6$  به  $10^{18}$  افزایش می‌دهد.

<sup>12</sup>latent semantic indexing

<sup>13</sup>singular value decomposition

<sup>14</sup>bag-of-words

<sup>15</sup>l-shingles

## ۲-۲-۲ جریان‌های داده‌ی حجیم

در بسیاری کاربردهای جدید پردازش داده، جریان‌های داده‌ی حجیم نقش بنیادی دارند. جریان‌های داده‌ای که از روترهای اینترنت، سوئیچ‌های تلفن، رصد امسفر، شبکه‌های سنسور، شرایط ترافیکی بزرگراهی، داده‌های مالی و غیره [۱، ۱۷، ۶، ۳، ۱۲، ۱۳، ۱۰] حاصل می‌شوند.

برخلاف پایگاه داده‌های سنتی، معمول نیست که جریان‌های داده‌ی حجیم (که با سرعت زیادی منتقل می‌شوند) در جای نگهداری شوند. بنابراین پردازش معمولاً به طور همزمان انجام می‌شوند. برای مثال، گاهی اوقات «رصد تصویری» داده‌ها با رصد تغییرات زمانی برخی آماره‌ها کفایت می‌کند. برای مثال آماره‌های نظیر: مجموع، تعداد آیتم‌های مجزا، برخی نرم‌های  $l_\alpha$ . در برخی کاربردها (برای مثال، طبقه‌بندی صدا/محتوا و جدا سازی) نیاز است یک مدل یادگیری آماری برای کلاسه‌بندی<sup>۱۶</sup> یا خوشه‌بندی<sup>۱۷</sup> جریان داده‌های حجیم توین شود. ولی معمولاً فقط می‌توانیم یک‌بار داده‌ها را مورد بررسی قرار دهیم.

یک خاصیت مهم جریان‌های داده‌ای این است که دینامیک هستند. به عنوان یک مدل محبوب، جریان  $u$  شامل ورودی‌های  $(i, u_i)$  است که  $i \in D$  برای مثال،  $D = 2^64$  زمانی که جریان بیان‌گر IP آدرس‌ها است.<sup>۱۸</sup> ورودی‌ها ممکن است به هر ترتیبی باشند و ممکن است مرتباً به روز شوند. ذات دینامیک جریان داده‌های حجیم فرآیند نمونه‌گیری را بسیار چالش‌برانگیزتر از زمانی می‌کند که با داده‌های ایستا سر و کار داریم.

## ۳-۲ چالش‌های نمونه‌گیری از داده‌های حجیم

در حالی که مسائل جذاب و چالش‌برانگیزی با ورود داده‌های حجیم شکل گرفته‌اند، این پایان‌نامه بر روی توسعه‌ی روش‌های نمونه‌گیری برای محاسبه فاصله در داده‌هایی با ابعاد بسیار بالا با استفاده از حافظه محدود تمرکز دارد. در کاربردهای مدل‌سازی آماری و یادگیری ماشین، در اغلب موارد به جای داده‌های اصلی به فاصله، به خصوص فاصله‌ی جفتی نیاز داریم. برای مثال، محاسبه ماتریس گرام<sup>۱۹</sup>  $AA^T$  در آمار و یادگیری ماشین معمول است.  $AA^T$  بیانگر هم‌بستگی ضرب‌های داخلی دوتایی در ماتریس داده‌ی  $A$  است.

دو داده‌ی  $u_1, u_2 \in \mathbb{R}^D$  داده شده‌اند. ضرب داخلی آن‌ها (که با  $a$  نمایش داده می‌شود) و  $l_\alpha$  (که با  $d_{(\alpha)}$  نمایش داده می‌شوند با عبارات زیر تعریف می‌شوند):<sup>۲۰</sup>

$$a = u_1^T u_2 = \sum_{i=1}^D u_{1,i} u_{2,i} \quad (1-2)$$

$$d_{(\alpha)} = \sum_{i=1}^D |u_1 - u_2|^\alpha \quad (2-2)$$

<sup>16</sup>Classification<sup>17</sup>Clustering

<sup>18</sup> هرچند ما بیشتر اوقات تعداد دقیق ابعاد ( $D$ ) یک جریان داده را نمی‌دانیم ولی در بیشتر کاربردها کافی است حد بالایی محافظه‌کارانه‌ای را در نظر بگیریم. برای مثال  $D = 2^64$  زمانی که جریان بیانگر IP های ورودی است. همچنین این یکی از دلایلی است که داده‌ها بسیار پراکنده هستند. به این نکته توجه داشته باشید که ابعاد بسیار بزرگ تأثیری در محاسبه فاصله‌ها و نمونه‌گیری طی الگوریتم‌های معرفی شده در این پایان‌نامه ندارد.

<sup>19</sup>Gram matrix

<sup>20</sup> ما فاصله  $l_\alpha$  را به صورت  $d_{(\alpha)} = \sum_{i=1}^D |u_1 - u_2|^\alpha$  تعریف کرده‌ایم. به جای اینکه به شکل  $(\sum_{i=1}^D |u_1 - u_2|^\alpha)^{1/\alpha}$  تعریف کنیم. زیرا شکل اول در کاربردهای عملی عمومیت بیشتری دارد. برای مثال،  $l_2$ ، در ادبیات معمولاً به شکل توان دو  $l_2$  بیان می‌شود.  $\sum_{i=1}^D |u_1 - u_2|^2$  به جای  $(\sum_{i=1}^D |u_1 - u_2|^2)^{1/2}$ . در این پایان‌نامه، ما برای سادگی  $\sum_{i=1}^D |u_1 - u_2|^2$  را «فاصله  $l_2$ » بیان می‌کنیم به جای «مربع فاصله  $l_2$ ».



به این نکته توجه داشته باشید که هم ضرب داخلی و هم فاصله به شکل جمع  $D$  جمله تعریف می‌شوند. بنابراین، زمانی که داده‌ها به اندازه‌ای حجیم هستند که نمی‌توان به طور کارا آن‌ها را مدیریت کرد، نمونه‌گیری خیلی عادی به نظر می‌رسد تا بتوان با انتخاب تصادفی  $k$  عضو از  $D$  جمله تخمینی از مجموع به دست آوریم (با ضریب مقیاس  $\frac{D}{k}$ ). در خصوص ماتریس داده‌ای  $A \in \mathbb{R}^{n \times D}$  انتخاب تصادفی مختصات <sup>۲۱</sup>،  $k$  ستون را از ماتریس داده به طور یکنواخت و تصادفی انتخاب می‌کند.

نمونه‌گیری از این جهت سودمند است که هم سایکل‌های کاری CPU را کاهش می‌دهد و هم در حافظه صرفه‌جویی می‌کند. در کاربردهای جدید، در اغلب موارد صرفه‌جویی در حافظه از اهمیت بیشتری برخوردار است. در نیم قرن گذشته گلوگاه محاسباتی حافظه بوده است، نه پردازشگر. سرعت پردازشگرها با نرخ تقریبی ۷۵ درصد در سال رو به افزایش است. در حالی که سرعت حافظه تقریباً سالی ۷ درصد افزایش می‌یابد [۱۶]. این پدیده به عنوان «دیوار حافظه» <sup>۲۲</sup> شناخته می‌شود [۱۶، ۲۰]. بنابراین در کاربردهایی که شامل مجموعه داده‌های حجیم می‌شوند، بحرانی‌ترین کار بیان کردن داده‌ها است. برای مثال، از طریق نمونه‌گیری با فرمی فشرده برای قرارگیری در ابعاد حافظه در دسترس.

## ۲-۳-۱ مزایای نمونه‌گیری تصادفی مختصات

نمونه‌گیری تصادفی مختصات به دو دلیلی معمولاً انتخاب پیش‌فرض است.

- **سادگی** این روش از لحاظ زمانی تنها از مرتبه  $O(nk)$  برای نمونه‌گیری  $k$  ستون از  $A \in \mathbb{R}^{n \times D}$  طول می‌کشد.

- **انعطاف پذیری** یک مجموعه نمونه را می‌توان برای تخمین بسیاری از شاخص‌های آماری استفاده کرد. شامل: ضرب داخلی، فاصله  $l_\alpha$  (برای هر مقداری از  $\alpha$ )

## ۲-۳-۲ معایب نمونه‌گیری تصادفی مختصات

با این حال نمونه‌گیری تصادفی مختصات دو ایراد اساسی دارد.

- معمولاً دقیق نیست زیرا مقادیری با مقدار زیاد محتمل است که گم شوند. مخصوصاً زمانی که داده‌ها دم سنگینی داشته باشند. داده‌های بزرگ مقیاس دنیای واقعی (مخصوصاً داده‌های مربوط به اینترنت) همیشه دم سنگین هستند و از قاعده توانی پیروی می‌کنند. [۱۸، ۹، ۵، ۱۴] زمانی که فاصله  $l_2$  یا ضرب داخلی را تخمین می‌زنیم. واریانس تخمین‌ها بر اساس ممان چهارم داده‌ها تعیین می‌شود. در حالی که در داده‌های دم سنگین، گاهی اوقات حتی ممان اول هم معنی‌دار نیست (محدود نیست) [۱۸].

- این روش داده‌های پراکنده را به خوبی مدیریت نمی‌کند. بسیاری از داده‌های بزرگ مقیاس به شدت پراکنده هستند، به عنوان مثال، داده‌های متنی [۸] و داده‌های بر اساس بازار [۲، ۱۹]. به جز برخی واژه‌های کاربردی مانند "A" و "The" بیشتر لغات با نسبت بسیار کمی در مستندات ظاهر می‌شوند ( $\square <$ ) اگر ما داده‌ها را با در نظر گرفتن تعدادی از ستون‌های ثابت نمونه‌گیری کنیم. خیلی محتمل است که بیشتر داده‌های (مقادیر غیر صفر) را از دست بدهیم. به خصوص موارد جذابی که دو مقدار با هم غیر صفر شده‌اند.

در این پایان‌نامه ما روش تصویر تصادفی را مورد بررسی قرار می‌دهیم و نشان خواهیم داد که این روش به خوبی قابلیت مدیریت داده‌های دم‌سنگین را دارد.

<sup>21</sup>Random coordinate sampling

<sup>22</sup>Memory wall

## ۴-۲ تصویر تصادفی پایدار

تصویر شکل ، ایده تویر تصادفی را نشان می دهد. ایده اصلی تصویر تصادفی ضرب ماتریس داده ی

## ۵-۲ کاهش بعد

## ۶-۲ کاهش بعد

## ۱-۶-۲ بارگیری مراجع

## منابع و مراجع

- [1] Aggarwal, Charu C. *Data streams: models and algorithms*, volume 31. Springer Science & Business Media, 2007.
- [2] Aggarwal, Charu C, Wolf, Joel L, and Yu, Philip S. *A new method for similarity indexing of market basket data*. ACM, 1999.
- [3] Babcock, Brian, Babu, Shivnath, Datar, Mayur, Motwani, Rajeev, and Widom, Jennifer. Models and issues in data stream systems. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 1–16. ACM, 2002.
- [4] Broder, Andrei Z. On the resemblance and containment of documents. In *Compression and complexity of sequences 1997. proceedings*, pages 21–29. IEEE, 1997.
- [5] Crovella, Mark E and Bestavros, Azer. Self-similarity in world wide web traffic: evidence and possible causes. *IEEE/ACM Transactions on networking*, 5(6):835–846, 1997.
- [6] Datar, Mayur and Indyk, Piotr. Comparing data streams using hamming norms. In *Proceedings 2002 VLDB Conference: 28th International Conference on Very Large Databases (VLDB)*, page 335. Elsevier, 2002.
- [7] Deerwester, Scott, Dumais, Susan T, Furnas, George W, Landauer, Thomas K, and Harshman, Richard. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.
- [8] Dhillon, Inderjit S and Modha, Dharmendra S. Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1-2):143–175, 2001.

- [9] Faloutsos, Michalis, Faloutsos, Petros, and Faloutsos, Christos. On power-law relationships of the internet topology. In *ACM SIGCOMM computer communication review*, pages 251–262. ACM, 1999.
- [10] Henzinger, Monika Rauch, Raghavan, Prabhakar, and Rajagopalan, Sridhar. Computing on data streams. *External memory algorithms*, 50:107–118, 1998.
- [11] Hornby, Albert Sydney, editor. *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, Oxford, UK, fourth edition, 1989.
- [12] Indyk, Piotr. Stable distributions, pseudorandom generators, embeddings and data stream computation. In *focs*, page 189. IEEE, 2000.
- [13] Kannan, J Feigenbaum S, Strauss, M, and Viswanathan, M. An approximate 11-difference algorithm for massive data streams. *Unknown*, Unknown.
- [14] Leland, Will E, Willinger, Walter, Taqqu, Murad S, and Wilson, Daniel V. On the self-similar nature of ethernet traffic. *ACM SIGCOMM Computer Communication Review*, 25(1):202–213, 1995.
- [15] Li, Ping. *Stable random projections and conditional random sampling, two sampling techniques for modern massive datasets*. Stanford, 2007.
- [16] McKee, Sally A. Reflections on the memory wall. In *CF'04: Proceedings of the 1st conference on Computing frontiers*, page 162, 2004.
- [17] Muthukrishnan, S. Data streams: Algorithms and applications (foundations and trends in theoretical computer science). *Hanover, MA: Now Publishers Inc*, 2005.
- [18] Newman, Mark EJ. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.
- [19] Strehl, Alexander and Ghosh, Joydeep. A scalable approach to balanced, high-dimensional clustering of market-baskets. In *International Conference on High-Performance Computing*, pages 525–536. Springer, 2000.
- [20] Wulf, Wm A and McKee, Sally A. Hitting the memory wall: implications of the obvious. *ACM SIGARCH computer architecture news*, 23(1):20–24, 1995.

## پیوست

موضوعات مرتبط با متن گزارش پایان نامه که در یکی از گروه‌های زیر قرار می‌گیرد، در بخش پیوست‌ها آورده شوند:

۱. اثبات‌های ریاضی یا عملیات ریاضی طولانی.
۲. داده و اطلاعات نمونه (های) مورد مطالعه (Case Study) چنانچه طولانی باشد.
۳. نتایج کارهای دیگران چنانچه نیاز به تفصیل باشد.
۴. مجموعه تعاریف متغیرها و پارامترها، چنانچه طولانی بوده و در متن به انجام نرسیده باشد.

## کد میپل

```
with(DifferentialGeometry):  
with(Tensor):  
DGsetup([x, y, z], M)  
frame name: M  
a := evalDG(D_x)  
D_x  
b := evalDG(-2 y z D_x+2 x D_y/z^3-D_z/z^2)
```

## واژه‌نامه‌ی فارسی به انگلیسی

Automorphism . . . . . خودریختی	آ
د	
Degree . . . . . درجه	اسکالر . . . . . Scalar
ر	ب
microprocessor . . . . . ریزپردازنده	بالابر . . . . . Lift
ز	پ
Submodule . . . . . زیرمدول	پایا . . . . . Invariant
س	ت
Character . . . . . سرشت	تناظر . . . . . Correspondence
ص	ث
Faithful . . . . . صادقانه	ثابت‌ساز . . . . . Stabilizer
ض	ج
Inner product . . . . . ضرب داخلی	جایگشت . . . . . Permutation
ط	چ
Loop . . . . . طوقه	چند جمله‌ای . . . . . Polynomial
ظ	ح
Valency . . . . . ظرفیت	حاصل ضرب دکارتی . . . . . Cartesian product
ع	خ

Nonadjacency . . . . . عدم مجاورت

ف

Vector space . . . . . فضای برداری

ک

Complete reducibility . . . . کاملاً تحویل پذیر

گ

Graph . . . . . گراف

م

Permutation matrix . . . . ماتریس جایگشتی

ن

Disconnected . . . . . ناهمبند

و

Invertible . . . . . وارون پذیر

ه

Connected . . . . . همبند

ی

Edge . . . . . یال

# واژه‌نامه‌ی انگلیسی به فارسی

<b>A</b>	Lift ..... بالا بر
Automorphism ..... خودریختی	<b>M</b>
<b>B</b>	Module ..... مدول
Bijection ..... دوسویی	<b>N</b>
<b>C</b>	Natural map ..... نگاشت طبیعی
Cycle group ..... گروه دوری	<b>O</b>
<b>D</b>	One to One ..... یک به یک
Degree ..... درجه	<b>P</b>
<b>E</b>	Permutation group ..... گروه جایگشتی
Edge ..... یال	<b>Q</b>
<b>F</b>	Quotient graph ..... گراف خارج‌قسمتی
Function ..... تابع	<b>R</b>
<b>G</b>	Reducible ..... تحویل پذیر
Group ..... گروه	<b>S</b>
<b>H</b>	Sequence ..... دنباله
Homomorphism ..... همریختی	<b>T</b>
<b>I</b>	Trivial character ..... سرشت بدیهی
Invariant ..... پایا	<b>U</b>
<b>L</b>	



Unique ..... منحصر بفرد

Vector space ..... فضای برداری

**V**

## **Abstract**

This page is accurate translation from Persian abstract into English.

### **Key Words:**

Write a 3 to 5 KeyWords is essential. Example: AUT, M.Sc., Ph. D, ..



**Amirkabir University of Technology**  
**(Tehran Polytechnic)**

**Department of ...**

**MSc Thesis**

**Title of Thesis**

**By**

**Name Surname**

**Supervisor**

**Dr.**

**Advisor**

**Dr.**

**Month & Year**