通过对应分析而非潜在语义分析改进信息检索

Qianqian Qi DavidJ.Hessen乌得勒支大学
乌得勒支大学q.qi@uu.nl d.j.hessen@uu.nl
Peter G.M. van der Heijden乌得勒支大学与南
安普敦大学p.g.m.vanderheijden@uu.nl

arXiv:2303.08030v1 [cs.IR] 14 Mar 2023

摘要：
潜在语义分析（LSA）和对应分析（CA）都是利用奇异值分解（SVD）进行信息检索的降维技术。从理论上讲，LSA 的结果既体现了文档与词项之间的关联，也体现了边际效应；相比之下，CA 只关注文档与词项之间的关联。边际效应通常与信息检索无关，因此从理论角度看，CA 更适用于信息检索。

本文对 LSA 和 CA 进行实证比较。我们对原始文档-词项矩阵的元素进行加权，并调整奇异值加权指数以提高 LSA 的性能。我们探究这两种加权方式是否也能提高 CA 的性能。此外，我们比较 LSA 和 CA 的最优奇异值加权指数，以确定 LSA 中的初始维度对应于什么。

四个实证数据集的结果表明，对应分析的表现始终优于潜在语义分析。对原始数据矩阵的元素进行加权可以提升对应分析的表现；然而，这取决于数据集，且提升幅度较小。调整奇异值加权指数通常可以改善对应分析的性能；然而，性能提升的程度取决于数据集和维度数量。总体而言，与潜在语义分析相比，对应分析需要更大的奇异值加权指数才能获得最佳性能。这表明，对应分析比潜在语义分析更重视初始维度，因此，边际在潜在语义分析的初始维度中发挥着重要作用。
关键词：奇异值分解；潜在语义分析；对应分析；奇异值加权指数；初始维度；信息检索。

1 引言

在信息检索中，给定的用户查询会与相关文档（Al- Qahtani、Amira 和 Ramzan，2015年；Guo 等人，2022；Kolda 和 O'leary，1998；Zhang、Yoshida 和 Tang，2011年）相匹配，而文档集合通常表示为文档-词项矩阵。计算给定用户查询与文档-词项矩阵中每篇文档的相似度，并返回相似度最高的文档。不幸的是，返回的文档可能并不相关，因为相似度是基于词语匹配计算的，而不同的词可能具有相同的含义，同一个词也可能具有不同的含义。

潜在语义分析（LSA）是信息检索中向量空间模型的一种变体（Deerwester、Dumais、Furnas、Landauer 和 Harshman，1990年；Dumais，1991；Dumais、Furnas、Landauer、Deerwester 和 Harshman，1988年）。LSA 使用奇异值分解（SVD）将原始的文档-词项矩阵替换为潜在语义向量。它

试图捕捉文档与词项之间的关联，从而帮助发现文档之间隐藏的语义关联。在 LSA 中，即使文档与查询没有任何共同词语，它们也可能彼此相似。

潜在语义分析已被广泛用作信息检索的常见基线方法（Chang、Lee、Wu、Liu 和 Liu，2021年；Duan、Gao、Ni 和 Wang，2021年；Paralı、Zontul 和 Ertu˘grul，2019），也用于文本挖掘（Bellaouar、Bellaouar 和 Ghada，2021年；Dzisevi˘c 和 ˘Se˘sok，2019）以及自然语言处理（Mohan 和 Kumar，2022年；Rezende、Rodrigues、Resendo 和 Komati，2022年）。此外，采用潜在语义分析得到的结果可与神经网络模型的结果相媲美（Altszyler、Sigman、Ribeiro 和 Slezak，2016年；Levy、Goldberg 和 Dagan，2015年；Phillips 等，2021年）。例如，Altszyler 等（2016）比较了 Word2Vec（Skip-Gram 模型）和潜在语义分析从梦境报告中提取相关语义模式的能力，并表明潜在语义分析取得了更好的性能。Phillips 等（2021）表明，在教育数据的文本分类任务中，潜在语义分析的表现优于神经网络方法（例如 ELMo 词嵌入）。此外，研究者还提出了依托潜在语义分析的新方法（Azmi、Al-Jouie 和 Hussain，2019年；Gupta 和 Patel，2021年；Hassani、Iranmanesh 和 Mansouri，2021年；Horasan，2022年；Patil，2022年；Suleman 和 Korkontzelos，2021年）。例如，Gupta 和 Patel（2021）提出了一种文本摘要算法，该算法采用潜在语义分析、词频-逆文档频率关键词提取器和 BERT编码器模型。该算法的表现优于潜在狄利克雷分配。Horasan（2022）提出了一种基于协同过滤的推荐系统，该系统采用潜在语义分析并取得了良好的性能。Patil（2022）开发了一种采用潜在语义分析和词频-逆文档频率的信息检索新程序。该程序能够以 0.943 的准确率，从文档中的句子识别出最显著和最不显著的词。

在计算机科学领域，捕捉文档与词项之间的关联似乎是潜在语义分析（LSA）取得成功的必要条件；然而，LSA 的解既混合了文档与词项之间的关联，也包含由文档长度和词项边际频率所产生的边际效应（Qi, Hessen, Deoskar, & Van der Heijden，2023年即将刊载于 Natural Language Engineering）。Hu 等人(2003）和 Qi 等人(2023年即将刊载于 Natural Language Engineering）指出，边际在 LSA 提取的第一维度中发挥着重要作用。

对应分析（CA）是另一种使用奇异值分解（SVD）的技术（Beh &Lombardo，2021年；Greenacre，1984年，2017年）。在计算机科学领域，对应分析的探索远不如潜在语义分析深入。对应分析通常用于制作二维图形展示（Arenas-M´arquez,Martinez-Torres, & Toral，2021年；Hou & Huang，2020年；Van Dam et al.，2021年）。例如，Arenas-M´arquez 等人(2021年）使用对应分析绘制双标图，展示卷积神经编码器的文档编码能够突出属于不同类别的文档之间的差异性。与潜在语义分析不同，对应分析通过预处理数据，在结果中忽略文档之间以及词项之间的边际频率差异信息，而仅关注文档与词项之间的关系（Qi 等人，2023年即将刊载于 Natural Language Engineering）。因此，对应分析似乎更适用于信息检索。

S´egu´ela 和 Saporta (2011年) 以及 Qi 等人 (2023年即将刊于 Natural Language Engineering)

分别通过实验比较了潜在语义分析和对应分析在文本聚类和文本分类中的表现，发现对应分析优于潜在语义分析。尽管潜在语义分析最初是为信息检索而提出的，但据我们所知，该领域仍缺乏对潜在语义分析和对应分析进行实证比较的研究。本文使用四个数据集（包括三个英语数据集和一个荷兰语数据集），比较潜在语义分析和对应分析在信息检索中的性能。

对原始文档-词项矩阵中的元素进行加权，是提升潜在语义分析（Bacciu 等人，2019年；Dumais，1991；Horasan、Erbay、Varcın 和 Deniz，2019年）性能的一种常见且有效的方法。潜在语义分析通常涉及对原始或预处理后的文档-词项矩阵进行奇异值分解。此外，潜在语义分析之所以广受欢迎，是因为它适用于不同的矩阵。在对应分析中，对原始文档-词项矩阵中的元素进行加权并不常见，而处理原始文档-词项矩阵是对应分析不可或缺的一部分（Beh& Lombardo，2021年；Greenacre，1984年，2017年）。对应分析基于标准化残差矩阵的奇异值分解。本文研究对条目进行加权的文档-词项矩阵的对应分析，以探究这种加权是否会影响对应分析的性能。

除了对文档-词项矩阵中的元素进行加权以提升信息检索性能之外，Caron （2001年）还提出改变潜在语义分析中的奇异值加权指数。他的研究结果表明，调整奇异值加权指数可以提升信息检索性能。自Caron （2001年）以来，奇异值加权指数已在由词-上下文矩阵生成的词嵌入中得到广泛研究和应用（Bullinaria &Levy，2012年；Drozd、Gladkova 和 Matsuoka，¨ ¨2016；Osterlund 、 Odling 和 Sahlgren，2015 年；Yin & Shen，2018）。此外，研究者还探讨了通过 Word2Vec 和 GloVe 创建的词嵌入中其他改变奇异值加权指数的变体（Liu、Ungar 和 Sedoc，2019年；Mu & Viswanath，2018年）。据我们所知，尚无研究探讨对应分析中改变奇异值加权指数的做法。鉴于调整潜在语义分析中奇异值加权指数取得了成功，我们将探究这种做法在对应分析中是否同样有效。

奇异值的加权指数越大，赋予初始维度的权重就越高。标准加权指数为 1。根据 Caron（2001）的实验结果，提高初始维度的权重通常可以提升标准测试数据集上的信息检索性能，但在问答匹配任务中，提高初始维度的权重则可能降低性能。关于词嵌入的论文往往会降低初始维度的贡献，以提升性能（Bullinaria & Levy，2012；¨Drozd 等，2016；Liu 等，2019；Mu & Viswanath，2018；Osterlund 等，2015；Yin & Shen，2018），不过，奇异值加权指数的最优值取决于任务（Osterlund 等，2015）。Bullinaria and Levy（2012）报告称，在基于 ukWaC 语料库（Baroni、Bernardini、Ferraresi 和 Zanchetta，2009）构建的词项—上下文矩阵上，对于托福、距离比较、语义分类和聚类纯度任务，降低初始维度的权重可以提升性能。他们认为，总体而言，初始维度似乎并不能提供最有用的语义信息，且包含大量“噪声”，最好将其去除或降低其影响。不过，初始维度具体对应什么仍不清楚。

如果初始维度对应于文档之间以及词项之间的边际差异，那么对应分析（CA）可能是最优程序，因为它会忽略这些无关信息。

总而言之，本研究作出了三项贡献。第一，比较潜在语义分析和对应分析在信息检索中的表现。结果表明，对应分析的性能优于潜在语义分析。因此，对应分析可作为通用基线，并广泛应用于信息检索、文本挖掘和自然语言处理。第二，探究各种加权方式能否提升对应分析的性能，包括对原始文档-词项矩阵元素的加权，以及调整奇异值加权指数。第三，研究潜在语义分析的初始维度代表什么，以及对应分析能否有效忽略原始或预处理后的文档-词项矩阵中的无用信息；这些信息构成了潜在语义分析提取出的初始维度中的很大一部分。我们对潜在语义分析和对应分析在四个数据集上的表现进行了广泛比较，采用的度量包括欧几里得距离、点积相似度和余弦相似度。

本文其余部分安排如下。第2节简要介绍潜在语义分析和对应分析。第3 节阐述本文所采用的方法论。第4节介绍欧几里得距离的结果，第5节介绍点积相似度和余弦相似度的结果。最后，第6 节总结并讨论研究结果。

2 潜在语义分析与对应分析

本节简要介绍潜在语义分析和对应分析。关于这些方法的详细介绍，请参阅 Qi 等人 (2023年即将发表于 Natural Language Engineering)。

2.1 潜在语义分析

考虑一个原始文档-词项矩阵 $\boldsymbol{F} = [f_{ij}]$，它有 m 行 $(i = 1, \ldots\ldots, m)$ 和 n 列$(j = 1, \ldots\ldots, n)$，其中行表示文档，列表示词项。加权可用于避免文档长度差异对表示产生过大影响，也可用于施加关于哪些词项更重要的某些先验假设（Deerwester 等人，1990年）。词项 a_{ij} 在文档 j 中 i 的加权元素为

$$a_{ij} = L(i, j) \times G(j) \times N(i), \tag{1}$$

其中，局部加权项 $L(i, j)$ 是词项 j 在文档 i中的权重，$G(j)$ 是词项 j 在整个文档集合中的全局权重，而 $N(i)$ 是文档 i的加权部分。选择 $L(i, j) = f_{ij}$、$G(j) = 1$和 $N(i) = 1$，即可得到原始文档-词项矩阵 F。常用的词频-逆文档频率（TF-IDF）可写为$L(i, j) = f_{ij}$、$G(j) = 1 + \log_2(n\text{docs}/df_j)$ $N(i) = 1$，其中 $n$docs 是集合中的文档数量，而 df_j 是词项 j 出现过的文档数量（Dumais,

1991）。 $\boldsymbol{A} = [a_{ij}]$ 是

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{2}$$

其中， $\boldsymbol{U}^T\boldsymbol{U} = \mathrm{I}$、 $\boldsymbol{V}^T\boldsymbol{V} = \mathrm{I}$ 和 $\boldsymbol{\Sigma}$ 是一个对角矩阵，其对角线上的奇异值按降序排列。我们将包含$\boldsymbol{U}$的前 k 列、包含 $\boldsymbol{V}$的前 k 列，以及 $\boldsymbol{\Sigma}$ 最大的 k 个奇异值的矩阵分别记为 $\boldsymbol{U}_k$、 $\boldsymbol{V}_k$和 $\boldsymbol{\Sigma}_k$。那么， $\boldsymbol{U}_k\boldsymbol{\Sigma}_k(\boldsymbol{V}_k)^T$ 在最小二乘意义上给出了 $\boldsymbol{A}$ 的最优秩-k 近似，这表明奇异值分解可用于数据降维。在潜在语义分析（LSA）中， $\boldsymbol{U}_k\boldsymbol{\Sigma}_k$ 和$\boldsymbol{V}_k\boldsymbol{\Sigma}_k$ 的行分别给出了行点和列点的坐标。 $\boldsymbol{U}_k\boldsymbol{\Sigma}_k$ ($\boldsymbol{V}_k\boldsymbol{\Sigma}_k$) 的行间欧几里得距离近似于 $\boldsymbol{A}$的行（列）间欧几里得距离（Deerwester 等，1990年；Parhizkar，2013年）。

在 LSA 的 k 维子空间中表示样本外文档或查询，对许多应用（包括信息检索）都很重要。假设新的加权文档是行向量 $\boldsymbol{d}$。由于 $\boldsymbol{V}^T\boldsymbol{V} = \mathbf{I}$和 $\boldsymbol{U}^T\boldsymbol{U} = \mathbf{I}$，我们有

$$\boldsymbol{A}\boldsymbol{V}_k = \boldsymbol{U}_k\boldsymbol{\Sigma}_k \tag{3}$$

and

$$\boldsymbol{A}^T\boldsymbol{U}_k = \boldsymbol{V}_k\boldsymbol{\Sigma}_k \tag{4}$$

因此，利用公式（3），样本外文档 $\boldsymbol{d}$ 在 k维 LSA 子空间中的坐标为 $\boldsymbol{d}\boldsymbol{V}_k$。类似地，利用公式（4），样本外词项 $\boldsymbol{t}$ （表示为行向量）在 k维 LSA 子空间中的坐标为$\boldsymbol{t}\boldsymbol{U}_k$。

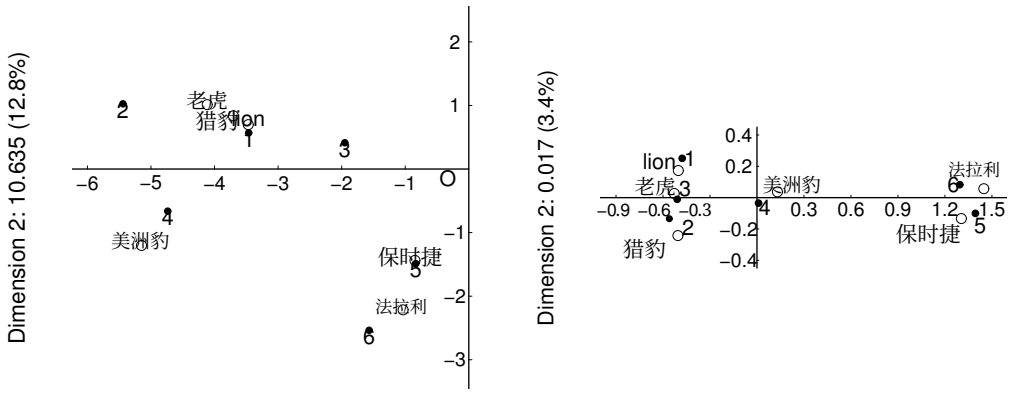
与Qi 等人（2023年即将刊载于 Natural Language Engineering）一样，我们首先用一个小型数据集来说明 LSA。这个小型数据集由Aggarwal （2018）（见表1）引入，包含 6 篇文档。对于每篇文档，我们关注六个词项的出现频率。前三篇文档主要涉及猫，最后两篇主要涉及汽车，第四篇则同时涉及两者。第四个词项“美洲豹”具有多义性，因为它既可以指猫科动物，也可以指汽车。

表1：文档-词项矩阵 $\boldsymbol{F}$：大小 6×6

	lion	老虎	猎豹	美洲豹	保时捷	法拉利
doc1	2	2	1	2	0	0
doc2	2	3	3	3	0	0
doc3	1	1	1	1	0	0
doc4	2	2	2	3	1	1
doc5	0	0	0	1	1	1
doc6	0	0	0	2	1	2

对原始文档-词项矩阵进行潜在语义分析（LSA-RAW）时， $\boldsymbol{F}$的行和列均未加权，因此，我们可以将**方程**（2）中的A $\boldsymbol{F}$替换为 $\boldsymbol{U}_2\boldsymbol{\Sigma}_2$。LSA-RAW 中文档和词项在前两个维度上的坐标为 $\boldsymbol{U}_2\boldsymbol{\Sigma}_2$

和 $\boldsymbol{V}_2\boldsymbol{\Sigma}_2$。图1a 展示了文档和词项的二维图。猫类词项（狮子和老虎）彼此接近；汽车类词项（美洲豹、保时捷和法拉利）彼此接近；汽车类文档（5 和 6）彼此接近。然而，猫类文档（1、2 和 3）彼此并不接近，文档 4 也没有位于猫类文档和汽车类文档之间，美洲豹也没有位于猫类词项和汽车类词项之间。这可以归因于潜在语义分析不仅显示文档与词项之间的关系，也显示文档和词项的大小：就后者而言，例如，美洲豹在文档中出现的频率最高，因此离原点最远。



(a) LSA-RAW

(b) 对应分析

图1：文档和词项的二维图：(a) LSA-RAW，(b) 对应分析。

2.2 对应分析

In CA，残差标准化矩阵经过奇异值分解，所得结果为Greenacre (20 17)

$$\boldsymbol{S} = \boldsymbol{D}_r^{-\frac{1}{2}}(\boldsymbol{P} - \boldsymbol{E})\boldsymbol{D}_c^{-\frac{1}{2}} \tag{5}$$

其中 $\boldsymbol{P} = [p_{ij}]$ 是联合观察比例矩阵，且 $p_{ij} = f_{ij}/\sum_i \sum_j f_{ij}$；$\boldsymbol{D}_r$是对角线上为 $r_i = \sum_j p_{ij} (i = 1, 2, \cdots, m)$ 的对角矩阵；$\boldsymbol{D}_c$ 是对角线上为 $c_j = \sum_i p_{ij} (j = 1, 2, \cdots, n)$ 的对角矩阵；而 $\boldsymbol{E} = [r_i c_j]$ 是在文档与词项统计独立的条件下得到的期望比例矩阵。^{1} ^{1} $^{-}$ $^{-}\boldsymbol{D}_r^{2}(\boldsymbol{P} - \boldsymbol{E})\boldsymbol{D}_c^{2}$中的元素是统计独立性模型下的标准化残差。这些元素的平方和得出总惯量，即 Pearsonχ^2 统计量除以样本量 N。对标准化残差矩阵进行奇异值分解后，我们得到

$$\boldsymbol{D}_r^{-\frac{1}{2}}(\boldsymbol{P} - \boldsymbol{E})\boldsymbol{D}_c^{-\frac{1}{2}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{6}$$

在对应分析中，$\boldsymbol{\Phi}_k\boldsymbol{\Sigma}_k$ 和 $\boldsymbol{\Gamma}_k\boldsymbol{\Sigma}_k$ 的各行分别提供行点和列点的坐标，^{1} ^{1} $^{-}$ $^{-}$其中 $\boldsymbol{\Phi}_k = \boldsymbol{D}_r^{2}\boldsymbol{U}_k$ 和 $\boldsymbol{\Gamma}_k = \boldsymbol{D}_c^{2}\boldsymbol{V}_k$。坐标的加权和

为 0： $\sum_i r_i\phi_{ik} = 0 = \sum_j c_j\gamma_{jk}$。 $\boldsymbol{\Phi}_k\boldsymbol{\Sigma}_k$ 的各行之间的欧几里得距离（$\boldsymbol{\Gamma}_k\boldsymbol{\Sigma}_k$）近似于$\boldsymbol{F}$的各行（列）之间的 χ^2-距离，其中行 k 和 l 之间的平方 χ^2-距离为

$$\delta_{kl}^2 = \sum_j \frac{(p_{kj}/r_k - p_{lj}/r_l)^2}{c_j} \tag{7}$$

在方程（7）中，行被转换为条件比例向量，每一行的比例之和为 1，例如第 k 行： p_{kj}/r_k，$j = 1, 2, \cdots, n$；并且，转换后的行中第 j 列的列元素之间的差异会根据 c_j进行校正，它表示第 j列的大小。

转移公式为

$$\boldsymbol{D}_r^{-1}\boldsymbol{P}\boldsymbol{\Gamma}_k = \boldsymbol{\Phi}_k\boldsymbol{\Sigma}_k \tag{8}$$

and

$$\boldsymbol{D}_c^{-1}\boldsymbol{P}^T\boldsymbol{\Phi}_k = \boldsymbol{\Gamma}_k\boldsymbol{\Sigma}_k \tag{9}$$

方程（8）表明，当使用 $\boldsymbol{D}_r^{-1}\boldsymbol{P}$ 的各行作为权重时，行点是列点的加权平均；当使用 $\boldsymbol{D}_r^{-1}\boldsymbol{P}$ 的各行作为权重时，方程（9）则表明，列点同时也是行点的加权平均。

根据方程（8），新的文档 $\boldsymbol{d}$（以行向量表示）可以通过将其置于列点的加权平均中，投影到 k 维子空间：$(\boldsymbol{d}/\sum_{j=1}^n d_j)\boldsymbol{\Gamma}_k$。新词项 $\boldsymbol{t}$也可以用类似的方法处理。

对于表1的对应分析，对应分析中前两个维度上的文档和词项坐标分别为 $\boldsymbol{\Phi}_2\boldsymbol{\Sigma}_2$ 和 $\boldsymbol{\Gamma}_2\boldsymbol{\Sigma}_2$。图1b展示了文档和词项的二维图。猫类词项（狮子和老虎）彼此接近；汽车类词项（美洲豹，保时捷和法拉利）彼此接近；美洲豹位于猫类词项和汽车类词项之间；汽车类文档（5 和 6）彼此接近，猫类文档（1、2 和 3）彼此接近；文档 4 则位于猫类文档和汽车类文档之间。图1b展示了所有数据特性。对比图1b 和1a 可见，与潜在语义分析相比，对应分析能够更清晰地呈现数据的重要特性。这是因为，由于 $\sum_i r_i\phi_{ik} = 0 = \sum_j c_j\gamma_{jk}$，各维度的坐标与边际正交，而且对应分析只关注文档与词项之间的关系。

3 方法论

本节介绍对条目经过加权的文档-词项矩阵进行对应分析。我们还将探讨如何研究初始维度的影响。随后，我们将说明所采用的研究设计、数据集和评估方法。

3.1 加权频数文档-词项矩阵的对应分析

对原始数据文档-词项矩阵中的条目进行加权，是提升潜在语义分析性能的有效方法，这促使我们研究

对应分析输入矩阵中各元素的加权。

$^{-1}$ $^{-1}$ $^{-}$通过 $\boldsymbol{D}_r^{-\frac{1}{2}}(\boldsymbol{P}-\boldsymbol{E})\boldsymbol{D}_c^{-\frac{1}{2}}$（见方程（5））对原始数据矩阵进行处理，被视为对应分析不可或缺的一部分。这一处理步骤有效地消除了边际，使对应分析能够聚焦于文档与词项之间的关系。方程（1）中原始文档-词项矩阵各项的加权（例如采用词频-逆文档频率），可以为更能体现文档含义的词项赋予更高的值。因此，对原始文档-词项矩阵各项进行加权，也可能是提升对应分析性能的有效方法。

要对加权频数构成的文档-词项矩阵进行对应分析，我们首先使用方程（1）得到加权频数文档-词项矩阵 $\boldsymbol{A}$，然后对该矩阵 $\boldsymbol{A}$ 而非 $\boldsymbol{F}$进行对应分析。

3.2 改变奇异值分解中初始维度的贡献

Caron （2001年提出，可以使用 $\boldsymbol{U}_k\boldsymbol{\Sigma}_k^{\alpha}$ 或 $\boldsymbol{V}_k\boldsymbol{\Sigma}_k^{\alpha}$ 作为坐标，而不是 $\boldsymbol{U}_k\boldsymbol{\Sigma}_k$ 或 $\boldsymbol{V}_k\boldsymbol{\Sigma}_k$，从而调整潜在语义分析中向量分量的相对强度，其中 α 是用于调整各维度重要性的奇异值加权指数。加权指数 α 决定了相对于第 $\alpha = 1$ 2.1节所述的标准 $\alpha = 1$情形，分量的加权方式。与 $\alpha < 1$ 相比， $\alpha > 1$对初始维度的强调较少，而 $\alpha > 1$则更为强调。

Bullinaria and Levy（2012年）同时采用了加权指数 $\alpha < 1$ 和排除初始维度的方法，两者带来的性能提升程度相近。他们认为，总体上，奇异值最高的维度往往并未提供关于语义的最有用信息，而且包含大量“噪声”，最好将其去除或减弱。然而，初始维度究竟对应什么尚不明确。鉴于此，我们改变了潜在语义分析和对应分析所提取的初始维度的贡献，并比较了它们的性能。

我们使用表 1 来说明 α 对奇异值和坐标的影响。我们使用$\alpha = -0.5$， $\alpha = 0$，$\alpha = 0.5$， $\alpha = 1$，以及 $\alpha = 1.5$。在文献中，我们经常会遇到$\alpha = 0.5$，因为它与

$$\boldsymbol{F} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = \left(\boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\right)\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{V}^T\right) \tag{10}$$

随后可使用坐标对 $\boldsymbol{U}_2\boldsymbol{\Sigma}_2^{1/2}$ 和 $\boldsymbol{V}_2\boldsymbol{\Sigma}_2^{1/2}$ 绘制双标图（Gabriel，1971年）。双标图在同一图形表示中呈现矩阵的行和列；其他常用的坐标对还有 $\boldsymbol{U}_2\boldsymbol{\Sigma}_2$ 和 $\boldsymbol{V}_2$，以及 $\boldsymbol{U}_2$ 和$\boldsymbol{V}_2\boldsymbol{\Sigma}_2$。对于坐标对 $\boldsymbol{U}_2\boldsymbol{\Sigma}_2$ 和 $\boldsymbol{V}_2$， $\boldsymbol{U}_2\boldsymbol{\Sigma}_2$ 各行之间的欧几里得距离近似于 $\boldsymbol{F}$各行之间的欧几里得距离；对于坐标对 $\boldsymbol{U}_2$ 和 $\boldsymbol{V}_2\boldsymbol{\Sigma}_2$， $\boldsymbol{V}_2\boldsymbol{\Sigma}_2$ 各行之间的欧几里得距离近似于 $\boldsymbol{F}$各列之间的欧几里得距离。坐标对 $\boldsymbol{U}_2\boldsymbol{\Sigma}_2^{1/2}$ 和 $\boldsymbol{V}_2\boldsymbol{\Sigma}_2^{1/2}$ 在这两种方案之间取得折衷，同时仍以图形方式表示矩阵 $\boldsymbol{F}$。在实际应用中，人们常会使用坐标对 $\boldsymbol{U}_2\boldsymbol{\Sigma}_2$ 和 $\boldsymbol{V}_2\boldsymbol{\Sigma}_2$；然而，这

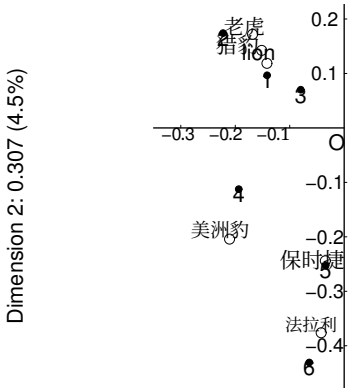
并非双标图表示，因为 $\boldsymbol{\Sigma}_2$ 被使用了两次。在双标图中，如果行点为 $\boldsymbol{U}_2\boldsymbol{\Sigma}_2^a$，那么列点为 $\boldsymbol{V}_2\boldsymbol{\Sigma}_2^{1-a}$，也就是说，矩阵中的任一项都近似等于相应行向量与列向量的内积。下文中，我们不绘制双标图，而绘制对称图，其中文档和词项具有相同的 α，因为实验中通常使用对称坐标（Berry、Dumais 和 O'Brien，1995年；Deerwester 等人，1990年；Dumais 等人，1988年；Levy 等人，2015年）。

表2列出了幂为 α 的奇异值： σ^α，幂为 α 的奇异值平方： $\sigma^{2\alpha}$，以及比例 $\sigma^{2\alpha}/\sum_\sigma \sigma^{2\alpha}$。其中，我们将幂为 α 的奇异值平方总和 $\sum_\sigma \sigma^{2\alpha}$ 称为 α 惯量。这些比例显示了所有分量到原点的欧几里得距离之和在各分量间的分布。 α 越大，对初始分量的强调就越多。当 α 为负时，对初始分量的强调程度低于后续分量。例如，对于 $\sigma^{-0.5}$，第一维度占 α 惯量的 0.017，而第五维度占 0.536。当 α 为 0 时，坐标的所有分量受到的强调程度相同；由于 $\boldsymbol{F}$ 的秩为 5，每个维度占 α 惯量的 0.2。随着 α 增大，对坐标初始分量的强调程度增加，对后续分量的强调程度则降低。当 α 分别为 0.5、1 和 1.5 时，第一维度分别占 α 惯量的 0.623、0.855 和 0.943，而第五维度分别占 0.020、0.001 和 0.000。标准的潜在语义分析解满足 $\alpha = 1$。

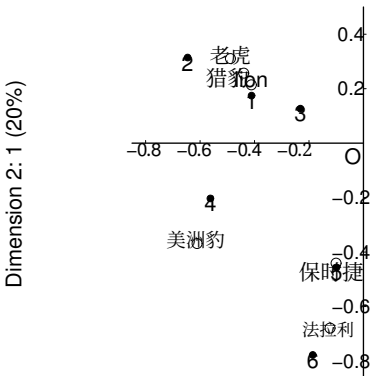
表2：LSA-RAW各维度的 σ^α、 $\sigma^{2\alpha}$以及解释的 α-惯量比例 $\sigma^{2\alpha}/\sum_\sigma \sigma^{2\alpha}$ 。

	dim1	dim2	维度3	维度4	维度5
$\sigma^{-0.5}$	0.345	0.554	1.006	1.320	1.917
σ^{-1}	0.119	0.307	1.012	1.741	3.675
$\sigma^{-1}/\sum_\sigma \sigma^{-1}$	0.017	0.045	0.148	0.254	0.536
σ^0	1	1	1	1	1
σ^0	1	1	1	1	1
$\sigma^0/\sum_\sigma \sigma^0$	0.2	0.2	0.2	0.2	0.2
$\sigma^{0.5}$	2.903	1.806	0.994	0.758	0.522
σ^1	8.425	3.261	0.988	0.574	0.272
$\sigma^1/\sum_\sigma \sigma^1$	0.623	0.241	0.073	0.042	0.020
σ^1	8.425	3.261	0.988	0.574	0.272
σ^2	70.985	10.635	0.976	0.330	0.074
$\sigma^2/\sum_\sigma \sigma^2$	0.855	0.128	0.012	0.004	0.001
$\sigma^{1.5}$	24.455	5.889	0.982	0.435	0.142
σ^3	598.063	34.684	0.964	0.189	0.020
$\sigma^3/\sum_\sigma \sigma^3$	0.943	0.055	0.002	0.000	0.000

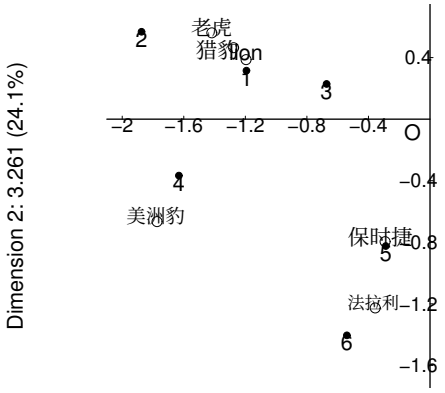
图2 展示了在不同的选择下，LSA-RAW中文档和词项的二维图。图2d 与图1a相同。当 α 增大时，第一维度上的行点（列点）之间的欧几里得距离相对于第二维度会增大。



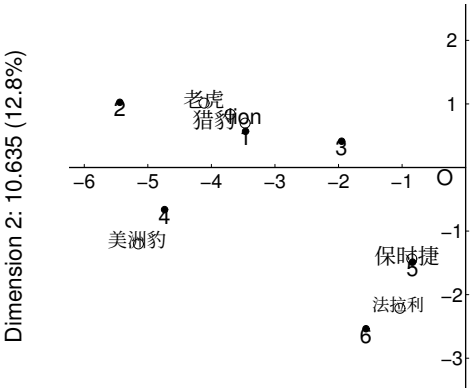
（a） $\alpha = -0.5$



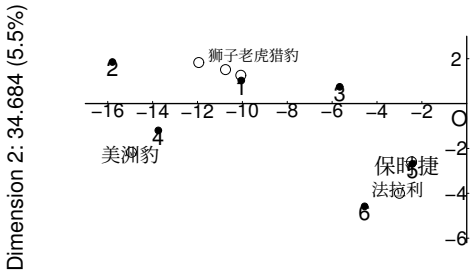
（b） $\alpha = 0$



（c） $\alpha = 0.5$



（d） $\alpha = 1$



（e） $\alpha = 1.5$

图2：LSA-RAW 中文档和词项的二维图，其中 (a) $\alpha = -0.5$、(b) $\alpha = 0$、(c) $\alpha = 0.5$、(d) $\alpha = 1$ 和 (e) $\alpha = 1.5$。

3.3 设计

我们比较潜在语义分析（LSA）和对应分析（CA）在信息检索中的性能。在潜在语义分析中，我们研究两种加权方式：对原始文档-词项矩阵的元素进行加权，并改变加权指数 α 。我们还探究这些加权方式对对应分析的影响。最后，我们研究在潜在语义分析和对应分析的最优维度数下，平均准确率均值（MAP）随 α 变化的情况。我们将维度数k 设为 1、2、$\cdots$、20、22、$\cdots$、50、60、$\cdots$至 100，并将 α 的值设为 -6、-5.5、$\cdots$、-2、-1.8，$\cdots$、4、4.5、$\cdots$至 8；我们考察所有 $40 \times 47 = 1,880$ 种参数值组合。

在研究原始文档-词项矩阵元素的加权时，我们对潜在语义分析和对应分析进行

- 原始数据矩阵$\boldsymbol{F}$，记为 RAW，
- L1行归一化矩阵 $\boldsymbol{F}^{L1}$，其中 $L(i,j) = f_{ij}$、$G(j) = 1$和 $N(i) = 1/\sum_{j=1}^{n} f_{ij}$，记为 NROWL1，
- L2行归一化矩阵 $\boldsymbol{F}^{L2}$，其中 $L(i,j) = f_{ij}$、$G(j) = 1$和 $N(i) = 1/\sqrt{\sum_{j=1}^{n} f_{ij}^2}$，记为 NROWL2，以及
- 词频-逆文档频率矩阵$\boldsymbol{F}_{\text{词频-逆文档频率}}$，如第 2.1 节所述，记为 TFIDF。

我们将对应分析与 TF-IDF 矩阵的组合称为 CA-TFIDF。同理，我们得到 LSA-RAW、LSA-NROWL1、LSA-NROWL2、LSA-TFIDF、CA-RAW、CA-NROWL1 和 CA-NROWL2。为了比较性能，RAW、NROWL1、NROWL2 和 TFIDF 用于不进行降维的词项匹配。

3.4 数据集

我们使用三个英语数据集和一个荷兰语数据集比较潜在语义分析和对应分析。三个英语数据集分别为 BBC体育（Greene &Cunningham，2006）、BBC新闻（Greene& Cunningham，2006）以及 20 Newsgroups 数据集（20-news-18846 bydata版本）（Rennie，2005）。荷兰语数据集为 *Wilhelmus* 数据集（Kestemont, Stronks, DeBruin, & Winkel，2017年）。

表3列出了所用四个数据集的一些统计信息。BBCNews 数据集包含 2,225 篇文档，分属五个类别之一。BBC体育数据集包含 731 篇文档，分属五个类别之一。20 Newsgroups 数据集包含 18,846 篇文档，分属 20 个类别之一。该数据集划分为训练集（60%）和测试集（40%）。我们使用该数据集的一个子集来评估信息检索效果。我们从四个类别（comp.graphics、rec.sport.hockey、sci.crypt 和 talk.politics.guns）的训练集中随机选取 600 篇文档，并从这四个类别的测试集中选取 400 篇文档。该*Wilhelmus* 数据集包含 186 篇文档，分为六个类别。

为了预处理三个英语数据集，我们将所有字符转换为小写，删除标点符号、数字和停用词，并进行词形还原。随后，

表 3：数据集特征。

类别	Data
商业	510
娱乐	386
政治	417
体育	511
技术	401

(a) BBC新闻的特征。

类别	Data
田径	101
板球	124
足球”	265
橄榄球”	147
网球	100

(b) BBC体育的特征。

类别	训练数据	测试数据
comp.graphics	141	100
rec.sport.hockey	164	99
sci.crypt	161	106
talk.politics.guns	134	95

(c) 20新闻组的特征。

类别	Data
datheen	35
marnix	46
heere	23
haecht	35
fruytiers	33
coornhert	14

(d) Wilhelmus 数据集的特征。

忽略频率低于 10 的词项。此外，我们会删除 20 Newsgroups 数据集中不需要的部分，例如邮件头（包括“From:”和“Reply-To:”等字段及其后的电子邮件地址），因为这些内容与信息检索几乎无关。荷兰语 *Wilhelmus* 数据集已预处理为标签—词形对。

由于*Wilhelmus* 和 BBCSport 数据集的文档数量相对较少，我们使用留一交叉验证（LOOCV）评估*Wilhelmus* 数据集，并使用五折交叉验证评估 BBCSport 数据集，以比较潜在语义分析和对应分析（Gareth、Daniela、Trevor 和 Robert，2021年）。BBCNews 数据集被随机划分为训练集（80%）和验证集（20%）。

在本研究的信息检索部分，验证集中的每篇文档都作为一个查询，且已知该文档的类别。训练集中与查询属于同一类别的文档，就是该查询的相关文档。

3.5 评估

我们比较潜在语义分析和对应分析各四个版本的平均精度均值，以探究这些方法在初始维度贡献发生变化时的信息检索表现（Kolda 和 O’leary，1998）。平均精度均值的计算方法如下：

- 通过计算查询向量与文档集合中每个文档向量之间的相似度来进行评估。我们使用三种相似度度量：欧几里得距离、点积相似度和余弦相似度。由于欧几里得距离是采用对应分析的一项重要动机，我们报告欧几里得距离的结果，并仅在正文中报告点积相似度和余弦相似度的部分结果，其余结果则见补充材料。

• 对于欧几里得距离，文档根据其与查询向量的相似度按升序排列（对于点积相似度和余弦相似度，则按降序排列）；因此，排在第一位的文档相似度最高。

• 精确率-召回率点由文档的有序列表推导得出。对于给定的查询，表 4根据文档是否相关以及是否被检索到，将有序列表中的文档划分为四种类型：

$\mathbf{C}$ = 有序列表中相关文档的集合，即与查询属于同一类别的文档

$\mathbf{D}$ = 有序列表中检索到的文档集合。也就是说，当返回 10 个文档时，检索到的文档集合由有序列表中的前 10 个文档组成。

表4：检索到的文档和相关文档。

	相关	不相关
已检索到	$\mathbf{C} \cap \mathbf{D}$	$\overline{\mathbf{C}} \cap \mathbf{D}$
未检索	$\mathbf{C} \cap \overline{\mathbf{D}}$	$\overline{\mathbf{C}} \cap \overline{\mathbf{D}}$

设 $|.|$ 表示一个集合中的文档数量。则精确率和召回率定义为

$$\text{precision} = \frac{|\mathbf{C} \cap \mathbf{D}|}{|\mathbf{D}|} \tag{11}$$

and

$$\text{recall} = \frac{|\mathbf{C} \cap \mathbf{D}|}{|\mathbf{C}|}. \tag{12}$$

因此，精确率定义为检索到的相关文档数量与检索到的文档总数之比，召回率定义为检索到的相关文档数量与相关文档总数之比。对于给定的查询，集合 $\mathbf{C}$ 是固定的。集合$\mathbf{D}$ 并非固定；如果我们返回前i 个文档，那么 $\mathbf{D}$ 由有序列表中的前 i 个文档组成。因此，对于给定的 i，我们可以得到一组精确率（见公式（11））和召回率（见公式（12））配对。我们将 i 的取值从 1 设为 l （有序列表中的文档数量），从而得到 l 组精确率-召回率配对。

• 然后，在 11 个召回率（0、0.1、 $\cdots$、1.0）下计算 11 个伪精确率，其中召回率为 x 时的伪精确率，是召回率从 x 到 1 之间的最大精确率。例如，召回率为 0.2 时的伪精确率，是召回率从 0.2 到 1 之间的最大精确率。

• 该查询的平均精确率是通过对 11 个伪精确率取平均得到的。

• 平均精度均值是所有查询的平均值。

MAP 值越高，表示性能越好。

4 欧氏距离的结果

4.1 比较潜在语义分析与对应分析在信息检索中的表现

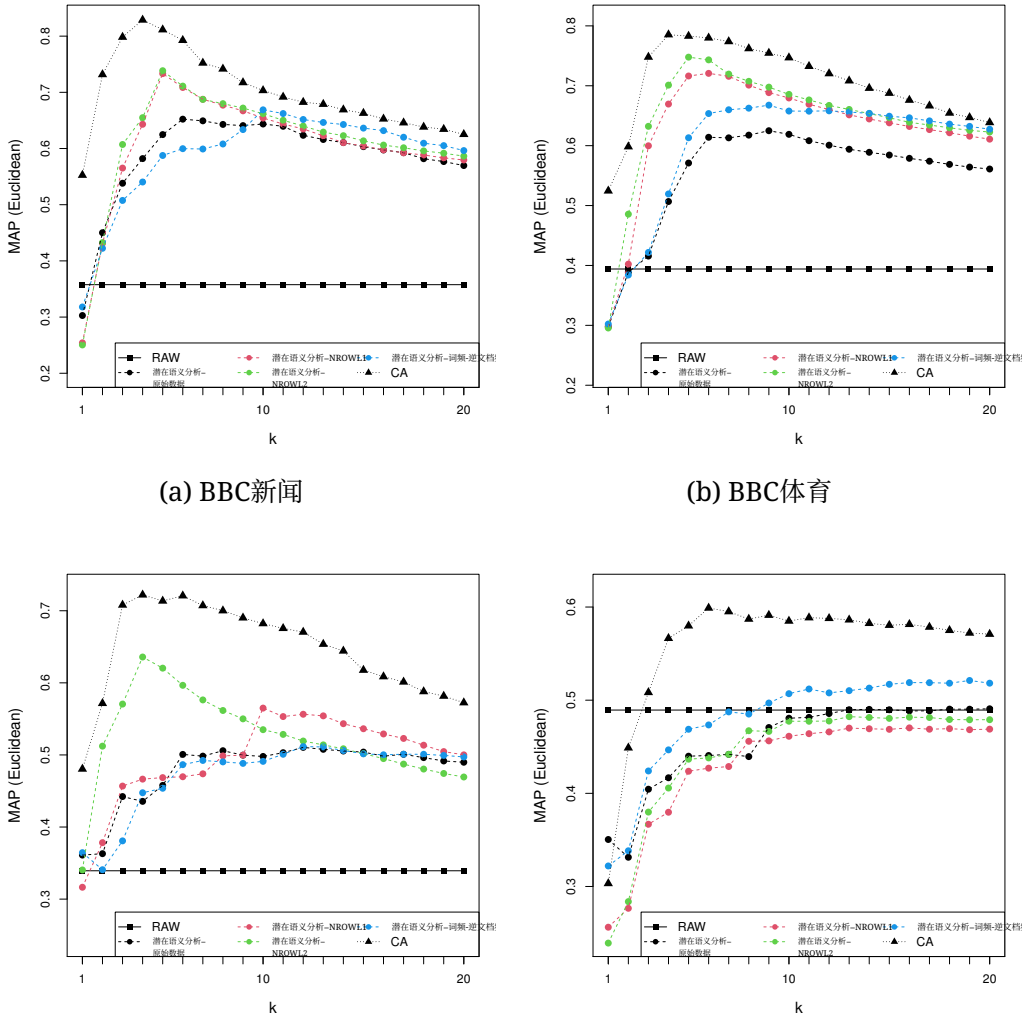
4.1.1 标准加权指数为 $\alpha = 1$ 时，四种版本的潜在语义分析和对应分析的 MAP 随维度数量的变化

我们首先考察潜在语义分析和对应分析在平均精度均值方面的性能，采用它们的标准用法，即不改变加权指数 α，即 $\alpha = 1$。不预先使用潜在语义分析或对应分析，而直接在文档-词项矩阵上进行的词项匹配，称为原始数据。我们预计，与 Qi 等人(2023年即将刊于 Natural Language Engineering的研究结果一致，潜在语义分析和对应分析的性能将优于原始数据，而对应分析的性能将优于潜在语义分析的四个版本。

图 3 展示了平均精度均值随维度数量 k 的变化情况，涵盖潜在语义分析的不同加权方案以及对应分析。我们仅展示前 20 个维度，因为所有曲线通常在维度达到 20 后都会下降。维度数最多为 100 的图见补充材料。表 5 列出了潜在语义分析的四个版本及对应分析在四个数据集上达到最佳平均精度均值时的维度数，以及相应的平均精度均值。根据图 3 和表 5，我们得出以下结论：

- 潜在语义分析和对应分析的平均精度均值都高于原始数据；使用完整维度的矩阵时，原始数据呈现为一条直线。
- 潜在语义分析和对应分析的性能都取决于维度数量 k。总体而言，平均精度均值随着 k 增加而上升，达到峰值后开始下降。对应分析的峰值出现在 $k = 4$。在对应分析中，与噪声相比，用于计算平均精度均值的信息在前四个维度中不断增加。在 $k \geq 5$的分量中，噪声占据主导地位，有用信息则相对不足，因此从此处开始，平均精度均值下降。
- 对应分析的平均精度均值显著优于潜在语义分析的四个版本，这与 Qi 等人 (2023年即将刊载于 Natural Language Engineering)的结论一致。他们指出，对于文档-词项矩阵，对应分析的性能优于潜在语义分析。这是因为潜在语义分析和对应分析对边际的处理方式不同。边际提供了与执行查询无关的信息。在对应分析中，边际被去除，因此，相较于噪声的信息相对量（我们非正式地称之为信息-噪声比）在对应分析中远高于潜在语义分析。这解释了对应分析为何具有更高的平均精度均值。
- 四种版本的潜在语义分析，其峰值通常出现在比对应分析更高的维度k 。这是因为在固定$\alpha = 1$时，边际对查询而言属于噪声；在潜在语义分析中，这种噪声在前几个维度中发挥着重要作用。因此，对应分析的峰值更早出现，也可以用其更高的信息噪声比来解释。
- 四种潜在语义分析方法的效果并不相同。在全部四个数据集中，采用加权方案都能显著提升潜在语义分析的性能。这些改进——

相较于 LSA-RAW 的效果取决于数据。综合四个数据集来看，LSA-NROWL2 的表现最佳；但对于*Wilhelmus* 数据集，LSA-NROWL1 和 LSA-NROWL2 的平均精度均值略低于 LSA-RAW。



(a) BBC新闻　(b) BBC体育

(c) 20新闻组　(d) Wilhelmus

图3：平均精度均值随维度数量的变化 k （采用标准坐标）。

表5：采用最佳维度数量时的平均精度均值 k。最佳值以粗体显示。

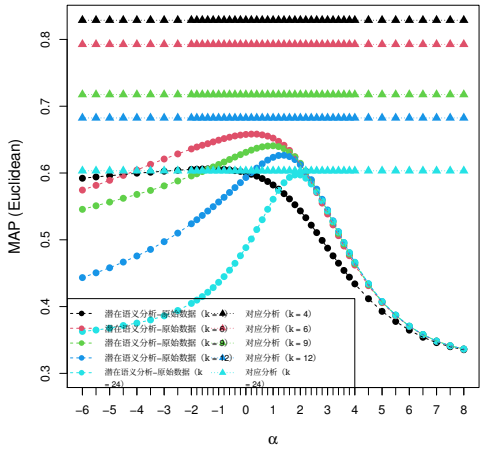
	BBC新闻		BBC体育		20新闻组		Wilhelmus	
	k	平均精度均值k	k	平均精度均值 k	MAP	k 值	平均精度均	
RAW		0.358		0.394	0.339		0.489	
LSA-RAW	6	0.652	9	0.625	12	0.510	24	0.492
LSA-NROWL1	5	0.733	6	0.721	10	0.565	16	0.470
LSA-NROWL2	5	0.738	5	0.748	4	0.636	13	0.482
LSA-TFIDF	10	0.669	9	0.668	12	0.512	19	0.521
CA	4	**0.829**	4	**0.785**	4	**0.722**	6	**0.599**

4.1.2 不同维度数量下，LSA 的平均精度均值随加权指数 α 变化的情况，与对应分析的平均精度均值对比

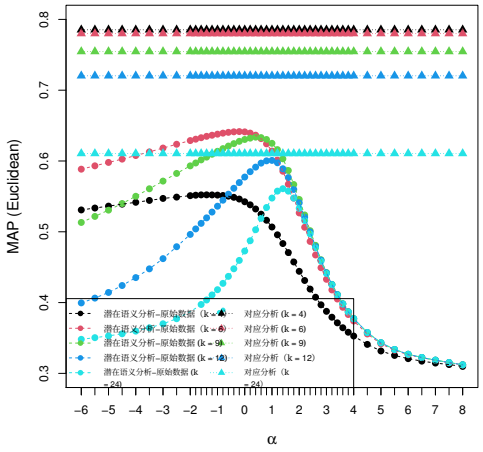
在第4.1.1节中，我们发现，就平均精度均值而言，对应分析的表现优于四种潜在语义分析版本，其中潜在语义分析采用了常见的加权指数 $\alpha = 1$。本节中，我们研究改变 α时，LSA-RAW 的性能是否会提升。

图 4 展示了维度数为 α 时，LSA-RAW 的平均精度均值随$k = 4$变化的情况，维度数分别为 6、9、12、和 24。作为对比，我们还报告了第 4.1.1 节中对应分析在这些维度数下的平均精度均值。选择这些 k 值，是因为表 5中这些维度数对 LSA-RAW 和对应分析而言都是最优的。表6 展示了最优的 α 及其对应的平均精度均值，这是图 4的精简版本。我们从图 4 和表 6中得出以下结论：

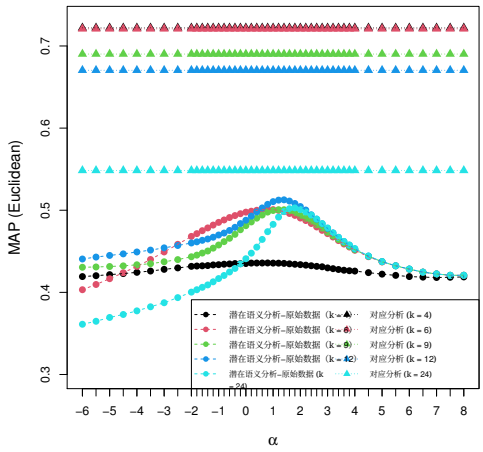
- 尽管通过调整 α，潜在语义分析-原始数据的性能有所提升，但对应分析的表现仍优于潜在语义分析-原始数据。
- 对于潜在语义分析-原始数据，整体平均精度均值会随着 α 的变化先上升后下降。这意味着，调整 α 有可能提升潜在语义分析-原始数据的性能。
- 平均精度均值的提升幅度很小。以 BBC新闻 数据集为例。在第 4.1.1 节中，我们发现，当 $\alpha = 1$，且 $k = 6$时，平均精度均值达到最优值 0.652。表 6 显示，对于 $\alpha = 0.2$，平均精度均值上升至 0.658。显然，对于 6 个维度，当 $\alpha = 0.2$时，就平均精度均值而言，信息-噪声比达到最优。对于 $\alpha = 0.2$，后续维度（6 个维度中的后几个维度）上的距离增大，而初始维度上的距离减小。这意味着，采用 α =0.2 时，边际影响最大的初始维度所造成的影响会减弱。这与 Bullinariaand Levy (2012年) 的结果一致，该结果表明，减少初始维度能够提升性能。
- 此外，LSA-RAW 的最优加权指数 α 取决于数据，且通常会随着 k而增加。这与 Caron（2001年）的结果一致。随着维度数量变化，最优加权指数 α 的变化是所研究的特定维度数量下信息与噪声之比所致。例如，对于 BBCNews 数据集，最优维度数量为 6；当维度数量更大时，最优α 加权指数会增大。加权指数增大 α 表明，较早维度上的距离对信息检索更为重要，因此会降低后续维度的作用。



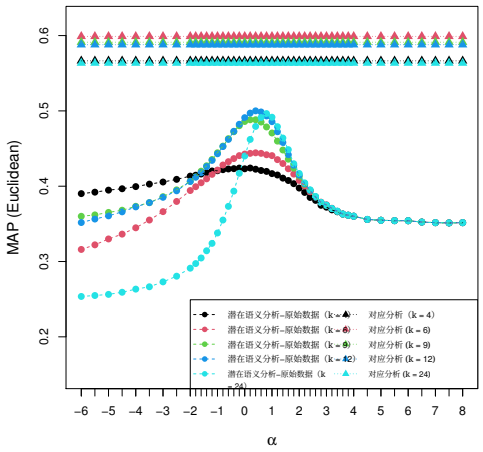
(a) BBC新闻



(b) BBC体育



(c) 20新闻组



(d) Wilhelmus

图4：LSA-RAW 的平均精度均值随 α 变化，以及对应分析的平均精度均值随 k变化 .

表6：LSA-RAW 在 α 下采用最优加权指数时的平均精度均值，以及对应分析在 $k = 4$、6、9、12、和 24 维度下的平均精度均值。粗体值为最佳结果。

	BBC新闻		BBC体育		20新闻组		Wilhelmus	
	α	平均精度均值	α	平均精度均值	α	MAP	α 值	平均精度均
LSA-RAW ($k = 4$)	-1.4	0.606	-1.4	0.552	0.8	0.436	0.2	0.424
LSA-RAW ($k = 6$)	0.2	0.658	-0.2	0.642	0.8	0.501	0.4	0.444
LSA-RAW ($k = 9$)	1	0.641	0.4	0.634	1.2	0.501	0.4	0.488
LSA-RAW ($k = 12$)	1.4	0.627	1	0.601	1.4	0.513	0.4	0.500
LSA-RAW ($k = 24$)	1.8	0.597	1.4	0.561	1.8	0.503	0.8	0.496
对应分析 ($k = 4$)		**0.829**		**0.785**		**0.722**		0.566
对应分析 ($k = 6$)		0.793		0.780		0.721		**0.599**
对应分析 ($k = 9$)		0.717		0.755		0.690		0.591
对应分析（$k = 12$)		0.682		0.720		0.670		0.588
对应分析（$k = 24$)		0.603		0.611		0.548		0.563

4.2 通过加权调整对应分析

4.2.1 对应分析中原始文档-词项矩阵各元素的加权

对原始文档-词项矩阵的元素进行加权，是提升潜在语义分析在信息检索中性能的有效方法。这里，我们探讨这一方法是否同样适用于对应分析。与图 3类似，图 5 展示了不同对应分析加权方案下，平均精度均值随 k 变化的情况。图 3 中的对应分析在图 5中称为 CA-RAW；对于对应分析/CA-RAW，这两幅图中的结果相同。对于对应分析的四个版本，表 7 展示了达到最优平均精度均值时的维度以及平均精度均值。我们从图 5和表 7得出以下结论：

- 总体而言，对原始矩阵各元素进行加权有时能提升对应分析的性能，但相较于 CA-RAW，提升幅度较小，且取决于数据。

与潜在语义分析相比，通过对原始矩阵的元素加权来提升对应分析在信息检索中的性能更为困难，原因有二：（1）CA-RAW 的平均精度均值本来就相对较高；（2）CA-RAW 已对原始文档-词项矩阵的元素进行加权，因为这本身就是该技术不可或缺的一部分（公式（5））。

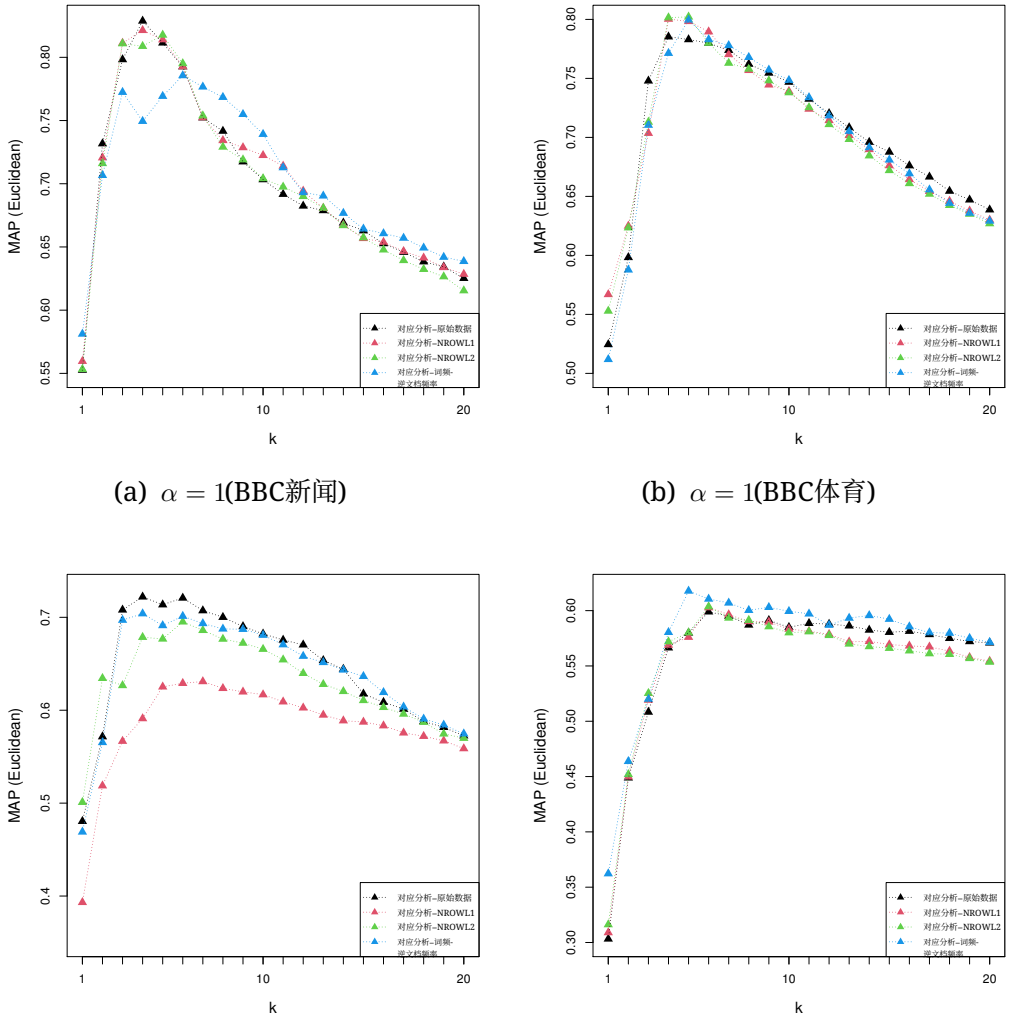
表7：四种对应分析版本在最优维度数 k 下的平均精度均值。最佳值以粗体显示。

	BBC新闻		BBC体育		20新闻组		Wilhelmus	
	k	平均精度均值	k	平均精度均值	k	MAP	k	平均精度均值
CA-RAW	4	**0.829**	4	0.785	4	**0.722**	6	0.599
CA-NROWL1	4	0.821	4	0.800	7	0.631	6	0.603
CA-NROWL2	5	0.818	5	**0.802**	6	0.695	6	0.604
CA-TFIDF	6	0.786	5	0.800	4	0.704	5	**0.618**

4.2.2 CA 的平均精度均值随加权指数的变化 α

本节介绍加权指数为 α 的对应分析。与图 4 类似，图 6 展示了在维度数为 α 时，CA-RAW 中的平均精度均值随 k =4、6、9、12、和 24 的变化。表 8 展示了最优的 α 及其对应的平均精度均值，这是图 6 的精简版本。根据图 6 和表 8，我们得出以下结论：

- 对于对应分析，整体平均精度均值随 α 先升高后降低。这意味着，改变 α 有可能提升对应分析的性能。
- 通过调整 α 来提高平均精度均值，具体效果取决于数据和维度。
- 如果将表 6 中的最大值与表 8中的最大值进行比较，几乎看不出有明显的提升。

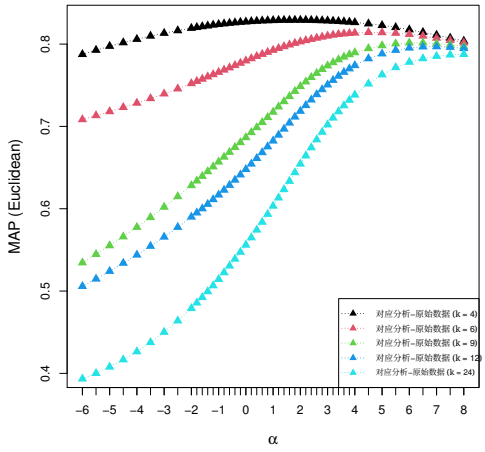


(a) $\alpha = 1$(BBC新闻)　　(b) $\alpha = 1$(BBC体育)

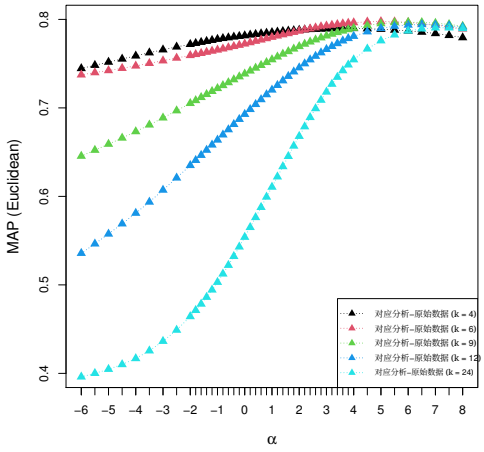
(c) $\alpha = 1$(20新闻组)　　(d) $\alpha = 1$(Wilhelmus)

图5：标准坐标下四种对应分析方法的平均精度均值随维度数量的变化情况 k 。

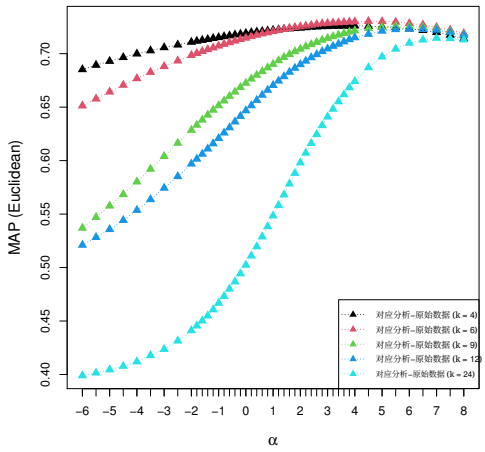
将表 8 与表 6中的 LSA-RAW 部分进行比较，CA-RAW 的最优 α 几乎总是大于 LSA-RAW，且几乎总是大于 1。也就是说，CA-RAW 要获得其最大平均精度均值，需要比 LSA-RAW 更大的 α 。因此，与潜在语义分析相比，对应分析通过更加重视初始维度而有所改进。潜在语义分析与对应分析的重要区别在于，潜在语义分析涉及边际，而对应分析则不涉及。因此，我们推断，潜在语义分析中的边际对初始维度有相当大的贡献；然而，它们与信息检索无关（属于“噪声”）。另一方面，对应分析有效地消除了这些无关信息。



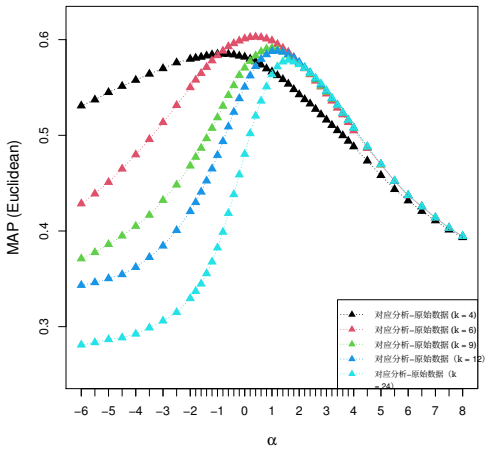
（a）BBC新闻



（b）BBC体育



（c）20新闻组



（d）Wilhelmus

图6：在不同的 k 值下，CA-RAW 的平均精度均值随 α 变化的情况。

表 8：在 $k = 4$、6、9、12、和 24 下，CA-RAW 采用最优 α 时的平均精度均值。粗体值为最佳结果。

	BBC新闻		BBC体育		20新闻组		Wilhelmus	
	α	平均精度均值	α	平均精度均值	α	MAP	α	平均精度均值
CA-RAW ($k = 4$)	2	**0.829**	3.6	0.790	4	0.726	-1	0.585
CA-RAW ($k = 6$)	4.5	0.814	5	**0.798**	4.5	**0.730**	0.4	**0.603**
CA-RAW ($k = 9$)	6.5	0.802	6	0.797	5.5	0.726	1	0.591
CA-RAW ($k = 12$)	7	0.797	6.5	0.794	6	0.723	1.2	0.588
CA-RAW ($k = 24$)	8	0.788	7.5	0.791	7	0.715	1.6	0.579

4.3 探讨在潜在语义分析和对应分析的最优维度数下，平均精度均值随 α 的变化

图7 展示了在最优维度数下，平均精度均值随 α 的变化。与第 4.1.1 节中讨论 $\alpha = 1$ 时类似，我们可以针对每个 α 求得相应的最优k (图中未显示)以及对应的平均精度均值(图中显示)。表 9 展示了最优 α、最优 k和对应的平均精度均值，是图 7 的简明版本。根据图 7 和表 9，我们可以看到

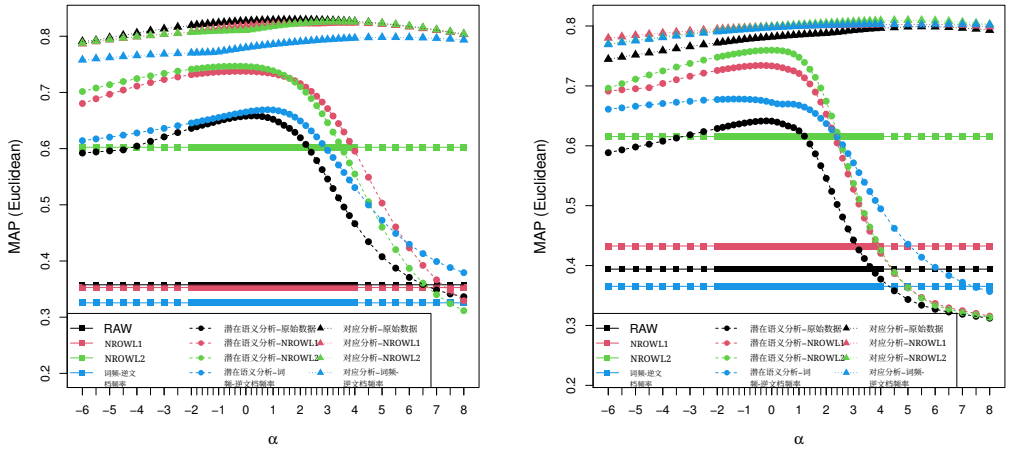
- 在最优 α 和最优维度数 k下，对应分析方法始终优于潜在语义分析方法和词项匹配方法。
- 在最优维度数下 k 对原始文档-词项矩阵的元素进行加权，可以提升对应分析的性能；不过，提升幅度较小且取决于数据。
- 与维度 $k = 4$、6、9、12、和24类似，在最优维度数 k 下，平均精度均值也会随 α 先升高后降低。因此，调整 α 可能会提升潜在语义分析和对应分析的性能。
- 对于不同的数据集，在最佳维度数量 k 下，最佳 α 差异很大。与潜在语义分析相比，对应分析需要更大的 α，才能在最佳维度数量 k下达到最佳性能。这表明，对应分析比潜在语义分析更重视初始维度。

表9：在最佳 α 和最佳维度 k下的平均精度均值。粗体值为组内最佳值；带下划线的值为总体最佳值。

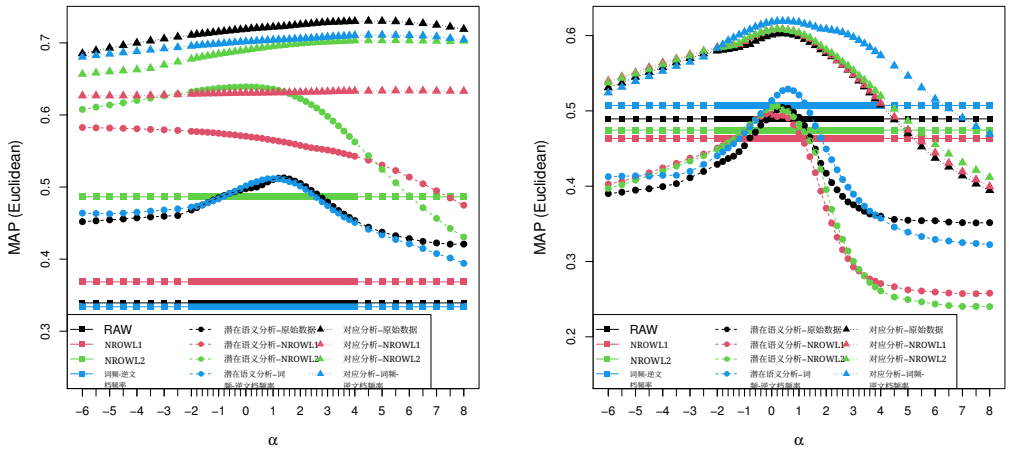
	BBC新闻			BBC体育			20新闻组			Wilhelmus		
	α	k	平均精度均值	α	k	平均精度均值	α	k	平均精度均值	α	k	平均精度均值
RAW			0.358			0.394			0.339			0.489
LSA-RAW	0.2	6	0.658	-0.2	6	0.642	1.4	12	0.513	0.4	13	0.505
CA-RAW	2	4	**<u>0.829</u>**	5.5	7	**0.799**	4.5	6	**<u>0.730</u>**	0.4	6	**0.603**
NROWL1			0.353			0.433			0.368			0.463
LSA-NROWL1	-0.4	5	0.738	-0.4	5	0.734	-6	10	0.583	-0.2	8	0.496
CA-NROWL1	3.6	5	**0.824**	5.5	5	**0.803**	5.5	7	**0.634**	0.2	6	**0.609**
NROWL2			0.602			0.615			0.486			0.474
LSA-NROWL2	-0.4	5	0.747	0	5	0.760	0	4	0.639	0.2	10	0.506
CA-NROWL2	3.6	5	**0.826**	5	5	**<u>0.810</u>**	5.5	6	**0.703**	0.2	6	**0.609**
词频-逆文档频率			0.326			0.365			0.334			0.507
LSA-TFIDF	0.8	10	0.669	-1.4	6	0.678	1	12	0.512	0.6	16	0.529
CA-TFIDF	5	6	**0.798**	6.5	7	**0.803**	5	6	**0.711**	0.6	5	**<u>0.619</u>**

5 点积相似度和余弦相似度的结果

在第 4 节中，我们展示了使用欧几里得距离作为相似度度量的结果。这里，为便于比较，我们给出点积相似度和余弦相似度的结果。



(a) BBC新闻　　(b) BBC体育



(c) 20新闻组　　(d) Wilhelmus

图7：最优维度数量下，平均精度均值随 α 变化的情况。

为节省篇幅，我们仅展示对应于欧几里得距离（表 5）、点积相似度（表 10）和余弦相似度（表 11）的表格。点积相似度和余弦相似度的其他表格和图示见补充材料。

点积相似度和余弦相似度的结果都得出了与欧几里得距离相同的结论。不过，就平均精度均值而言，余弦相似度的性能优于欧几里得距离和点积相似度。我们在第 4 节展示了欧几里得距离的结果，因为：（1）在调整加权指数的背景下，它更容易解释 α：随着 α 的增加，初始维度上行点（列点）之间的欧几里得距离相对于后续维度上的距离会增大；（2）在文献中，欧几里得距离是解释对应分析的首选方式（事实上，我们从未见过用余弦相似度或点积相似度来解释对应分析的情况）。

表10：点积相似度的最佳维度数量下的平均精度均值 k。粗体值为最佳值。

	BBC新闻		BBC体育		20新闻组		Wilhelmus	
	k	平均精度均值	k	平均精度均值	k	MAP	k	平均精度均值
RAW		0.543		0.534		0.454		0.342
LSA-RAW	14	0.580	15	0.536	36	0.488	90	0.343
LSA-NROWL1	5	0.751	7	0.663	14	0.601	80	0.404
LSA-NROWL2	5	0.744	6	0.693	4	0.672	34	0.475
LSA-TFIDF	32	0.613	22	0.651	60	0.490	100	0.421
CA	4	**0.896**	7	**0.865**	6	**0.788**	12	**0.642**

表11：关于余弦相似度的最佳维度数量下的平均精度均值 k。粗体值为最佳值。

	BBC新闻		BBC体育		20新闻组		Wilhelmus	
	k	平均精度均值	k	平均精度均值	k	MAP	k	平均精度均值
RAW		0.602		0.615		0.487		0.474
LSA-原始数据	6	0.779	9	0.766	12	0.654	22	0.484
LSA-NROWL1	5	0.817	6	0.775	12	0.651	13	0.481
LSA-NROWL2	5	0.825	5	0.795	4	0.745	13	0.482
LSA-词频-逆文档频率	10	0.796	9	0.819	12	0.665	19	0.531
CA	4	**0.902**	7	**0.878**	6	**0.794**	12	**0.652**

6 结论与讨论

我们将四种版本的潜在语义分析与对应分析进行了比较，发现就平均精度均值而言，对应分析的表现始终优于潜在语义分析。随后，我们比较了不同维度数量范围下，权重指数为 α 时的 LSA-RAW 与对应分析。尽管通过为 α选择合适的值可以提升潜在语义分析的表现，对应分析的表现始终优于潜在语义分析。

接下来，我们将原始文档-词项矩阵的不同加权因子应用于对应分析（CA）。我们发现，对原始矩阵进行加权有时能提升对应分析的性能，但相较于 CA-RAW，提升幅度较小，且取决于数据。随后，我们调整了对应分析中的加权指数 α。对于对应分析，MAP 随着 α的变化先上升后下降。调整加权指数 α 有可能提升对应分析的性能。然而，调整 α 所带来的性能提升取决于数据和维度。

对于 LSA，使用 $\alpha = 1$的标准坐标时，坐标各行之间的欧几里得距离近似于分解矩阵各行之间的欧几里得距离。对于 CA，坐标各行之间的欧几里得距离近似于分解矩阵各行之间的 χ^2–距离。与标准坐标相比， $\alpha < 1$ 对初始维度的强调程度较低。相反， $\alpha > 1$

与标准坐标相比，对初始维度的强调程度较高。CA 的最佳 α 几乎总是大于 LSA 的最佳值，并且几乎总是大于 1。

最后，我们研究了在维度数量达到最优时，MAP 随 α 变化的情况。结果再次表明，对应分析的表现优于潜在语义分析。尽管在维度数量达到最优时，最优的 α 取决于数据，但对应分析的最优 α 通常远大于潜在语义分析的最优值。

Bullinaria and Levy (2012年) 认为，潜在语义分析中的初始维度往往无法提供关于语义的最有用信息，且容易受到“噪声”的干扰。上述结果表明，与潜在语义分析相比，对应分析更加重视初始维度。潜在语义分析与对应分析的主要区别在于，前者包含边际，而后者不包含边际（Qi 等人，2023年即将刊于 Natural Language Engineering）。因此，我们推断，边际对潜在语义分析的初始维度有很大影响。这些边际与信息检索无关。对应分析有效地消除了这些无关信息。

本文重点研究了使用欧几里得距离时对应分析和潜在语义分析的性能。我们还针对点积相似度和余弦相似度进行了相同的实验。两者的结果与欧几里得距离的结果几乎相同。余弦相似度的表现优于欧几里得距离和点积相似度。本文聚焦于欧几里得距离，原因有二：（1）在调整 α的背景下，它更容易解释：随着 α增大，初始维度上的行点（列点）之间的欧几里得距离相对于后续维度上的距离会增大；（2）此前从未在对应分析中使用过点积相似度和余弦相似度，因此聚焦于欧几里得距离，更有利于将结果与现有文献进行比较。

根据实验结果和分析，我们提出以下三条实践建议：

1. 使用对应分析代替潜在语义分析；使用对应分析可视化数据。

2. 如果信息检索是关键问题，计算平均精度均值时应使用余弦相似度，而非欧几里得距离和点积相似度。

3. 如果平均精度均值方面的最优性能并非关键，便无需对对应分析所用原始文档-词项矩阵的元素进行加权，也无需优化 α 对应分析的性能，从而节省时间。否则，这两种加权方式可视为提升对应分析性能的潜在方法。

潜在语义分析有多种应用。未来工作中，尝试用对应分析替代潜在语义分析来改进这些应用，将会很有意思。

致谢

作者 Qianqian Qi 获得中国国家留学基金管理委员会的资助。

参考文献

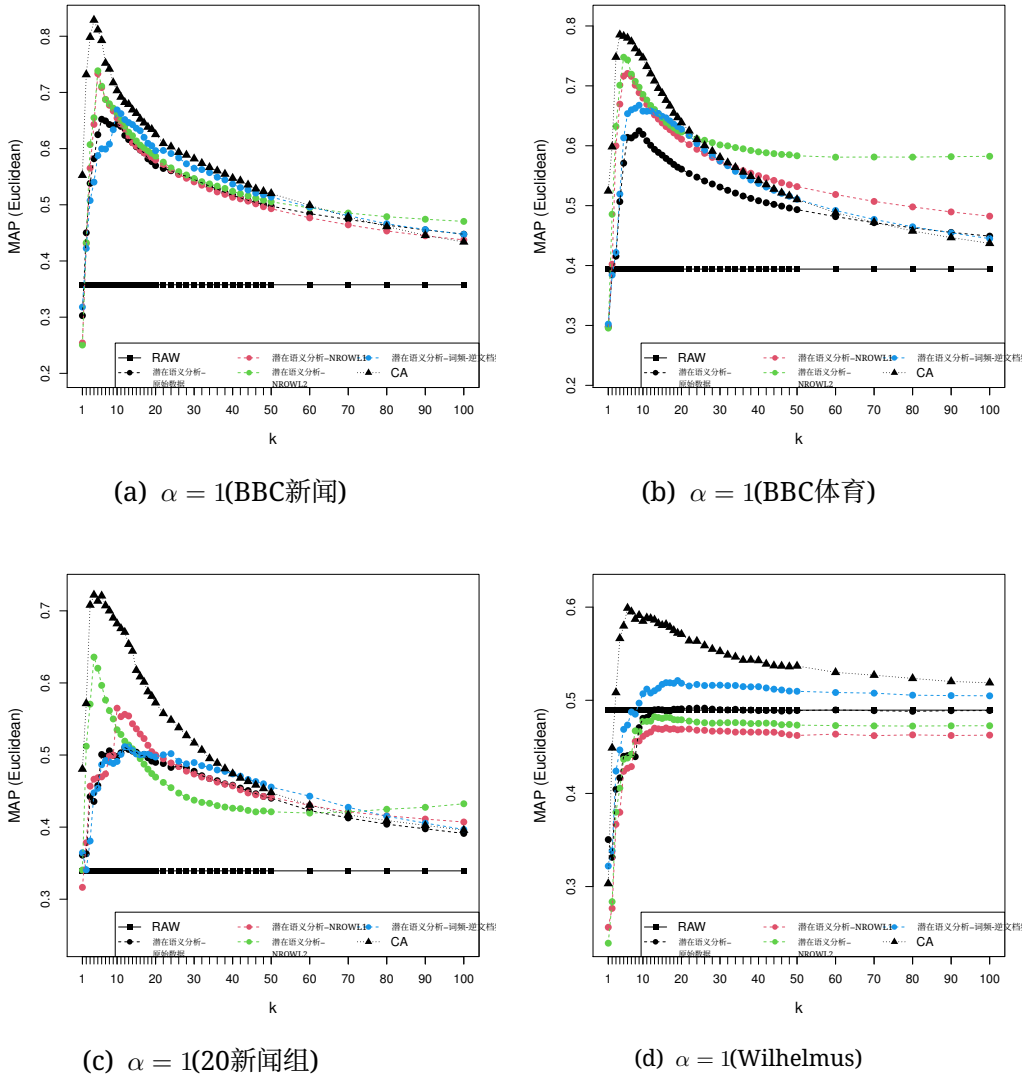
Aggarwal, C. C.（2018）。文本机器学习。施普林格。Al-Qahtani, M., Amira, A., & Ramzan, N.（2015）。一种用于电子健康系统的高效信息检索技术。载于《2015年国际系统、信号与图像处理会议（IWSSIP）》（第257–260页）。Altszyler, E., Sigman, M., Ribeiro, S., & Slezak, D. F.（2016）。小型语料库中潜在语义分析与 Word2Vec 嵌入的比较研究：梦境数据库案例研究。预印本，网址：https://arxiv.org/abs/1610.01520。Arenas-M´arquez, F. J., Martinez-Torres, R., & Toral, S.（2021）。对在线评论进行卷积神经编码，以识别 Tripadvisor 上的旅游团体类型主题。《信息处理与管理》，58(5)，102645。Azmi, A. M., Al-Jouie, M. F., & Hussain, M.（2019）。AAEE——阿拉伯语学生作文自动评估。《信息处理与管理》，56(5)，1736–1752。Bacciu, A., Morgia, M. L., Mei, A., Nemmi, E. N., Neri, V., & Stefa, J.（2019）。使用失真和潜在语义分析检测 Twitter 账号的机器人及性别。载于 Clef。Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E.（2009）。古怪的万维网：一个由经过语言学处理的超大型网络爬取语料库组成的集合。《语言资源与评估》，43(3)，209–226。Beh, E. J., & Lombardo, R.（2021）。对应分析导论。John Wiley & Sons。Bellaouar, S., Bellaouar, M. M., & Ghada, I. E.（2021）。主题建模：科学出版物中潜在语义分析与潜在狄利克雷分配的比较。载于《2021年第四届数据存储与数据工程国际会议》（第59–64页）。Berry, M. W., Dumais, S. T., & O'Brien, G. W.（1995）。运用线性代数实现智能信息检索。《SIAM评论》，37(4)，573–595。Bullinaria, J. A., & Levy, J. P.（2012）。从词语共现统计中提取语义表示：停用词表、词干提取和奇异值分解。《行为研究方法》，44(3)，890–907。Caron, J.（2001）。潜在语义分析评分实验：最优秩与基。载于《SIAM计算信息检索研讨会论文集》（第157–169页）。Chang, C.-Y., Lee, S.-J., Wu, C.-H., Liu, C.-F., & Liu, C.-K.（2021）。利用词语语义概念检测文本文件中的抄袭。《信息检索期刊》，24，298–321。Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R.（1990）。基于潜在语义分析的索引。《美国信息科学学会会刊》，41(6)，391–407。Drozd, A., Gladkova, A., & Matsuoka, S.（2016）。词嵌入、类比与机器学习：超越 king-man+ woman=queen。载于《COLING 2016：第26届国际计算语言学会议论文集：技术论文》（第3519–3530页）。

Duan, L., Gao, T., Ni, W., & Wang, W. (2021年). 基于潜在语义和显式评分的混合智能服务推荐。《国际智能系统杂志》, 36(12), 7867–7894. Dumais, S. T. (1991). 改进从外部来源检索信息的效果。《行为研究方法、仪器与计算机》, 23(2), 229–236. Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988年). 利用潜在语义分析改进文本信息的获取。载于 SIGCHI计算机系统中的人为因素会议论文集（第 281–285 页）。Dzisevič, R., & Šešok, D. (2019年). 使用不同特征提取方法进行文本分类。载于2019年电气、电子与信息科学开放会议（eSTREAM）论文集（第 1–4 页）。Gabriel, K. R. (1971年). 矩阵的双标图形显示及其在主成分分析中的应用。《生物统计学》, 58(3), 453–467. Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2021 年). 统计学习导论：及其在 R 中的应用。施普林格。Greenacre, M. J. (1984年). 对应分析的理论与应用。学术出版社。Greenacre, M. J. (2017年). 对应分析实践。CRC出版社。Greene, D., & Cunningham, P. (2006). 核文档聚类中对角占优问题的实用解决方案。载于第 23 届国际机器学习会议论文集（第 377–384 页）。Guo, J., Cai, Y., Fan, Y., Sun, F., Zhang, R., & Cheng, X. (2022). 第一阶段检索的语义模型：综合综述。ACM 信息系统汇刊（TOIS）, 40(4), 1–42. Gupta, H., & Patel, M. (2021年). 使用 LSA 和基于句子的主题建模与 BERT 进行文本摘要的方法。载于2021年国际人工智能与智能系统会议（ICAIS）论文集（第 511–517 页）。Hassani, A., Iranmanesh, A., & Mansouri, N. (2021年). 使用非负矩阵分解和潜在语义分析进行文本挖掘。《神经计算与应用》, 1–22. Horasan, F. (2022). 基于潜在语义索引的推荐系统混合协同过滤。《阿拉伯科学与工程杂志》, 1–15. Horasan, F., Erbay, H., Varcın, F., & Deniz, E. (2019年, 02). 潜在语义分析中的交替低秩矩阵近似。《科学编程》, 2019, 1–12. Hou, R., & Huang, C.-R. (2020年). 汉语区域变体与文体变体分类：一种基于可比平衡语料库的对应分析方法。Natural Language Engineering, 26(6), 613–640. Hu, X., Cai, Z., Franceschetti, D., Penumatsa, P., Graesser, A., Louwerse, M., . . . Tutor- ing Research Group (2003). LSA：第一维度和维度加权。载于认知科学学会年会论文集（第 25 卷）。Kestemont, M., Stronks, E., De Bruin, M., & Winkel, T. d. (2017年). 检索日期：2021年7月17日，网址：https://github.com/mikekestemont/anthem。Kolda, T. G., & O’leary, D. P. (1998). 一种用于潜在

语义索引信息检索的半离散矩阵分解。ACM 信息系统汇刊, 16(4), 322–346. Levy, O., Goldberg, Y., & Dagan, I. (2015年). 从词嵌入中汲取经验以改进分布相似度。计算语言学协会汇刊, 3, 211–225. Liu, T., Ungar, L., & Sedoc, J. (2019年). 通过概念器取反对词向量进行无监督后处理。载于 AAAI 人工智能会议论文集（第 33 卷，第 6778–6785 页）。Mohan, G. B., & Kumar, R. P. (2022年). 文本摘要主题建模综合调查。微电子与电信工程, 231–240. Mu, J., & Viswanath, P. (2018年). 除顶部之外：一种简单有效的词表示后处理方法。第六届国际学习表征会议，ICLR 2018——会议论文集。¨ ¨
Osterlund, A., Odling, D., & Sahlgren, M. (2015年). 分布式语义模型中潜变量的分解。载于2015年自然语言处理实证方法会议论文集（第 227–231 页）。Paralı, U., Zontul, M., & Ertu˘grul, D. C¸. (2019 年). 使用词项 - 文档矩阵的简化行阶梯形进行信息检索。互联网技术期刊。Parhizkar, R. (2013). 欧几里得距离矩阵：性质、算法与应用。洛桑联邦理工学院计算机与通信科学学院，瑞士洛桑。Patil, A. (2022). 使用 Kubeflow，通过 LSA 和词频-逆文档频率进行文档词语显著性分析，以实现信息检索。载于《专家云与应用》（第 335–348 页）。施普林格。Phillips, T., Saleh, A., Glazewski, K. D., Hmelosilver, C. E., Lee, S., Mott, B., & Lester, J. C. (2021年). 比较小规模教育数据文本分类的自然语言处理方法。载于第 11 届学习分析与知识国际会议伴随论文集。Qi, Q., Hessen, D. J., Deoskar, T., & Van der Heijden, P. G. M. (2023年，即将刊于 Natural Language Engineering). 文档-词项矩阵的潜在语义分析与对应分析比较。预印本见 https://arxiv.org/abs/2108.06197。Rennie, J. (2005年). 20 Newsgroups 数据集。检索日期：2022年4月21日，网址：http://qwone .com/~jason/20Newsgroups/。Rezende, J. M. d., Rodrigues, I. M. d. C., Resendo, L. C., & Komati, K. S. (2022年). 结合自然语言处理技术以及 LSA、Word2Vec 和词移距离算法，对专利文档进行技术预测与相似性分析。技术分析与战略管理, 1–22. S´egu´ela, J., & Saporta, G. (2011年). 潜在语义分析与对应分析的比较。载于 Carme 2011 对应分析及相关方法国际会议。Suleman, R. M., & Korkontzelos, I. (2021年). 扩展潜在语义分析以管理

其句法盲点。《Expert Systems with Applications》, 165, 114130。Van Dam, A., Dekker, M., Morales-Castilla, I., Rodr´ıguez, M. ´A., Wichmann, D., & Baudena, M. (2021年). 对应分析、谱聚类和图嵌入：在生态学和经济复杂性中的应用。《Scientific Reports》, 11(1), 1–14。Yin, Z., & Shen, Y. (2018). 论词嵌入的维度。《神经信息处理系统进展》, 31。Zhang, W., Yoshida, T., & Tang, X. (2011年). TF*IDF、潜在语义索引和多词在文本分类中的比较研究。《Expert Systems with Applications》, 38(3), 2758–2765。

附录 A：欧几里得距离



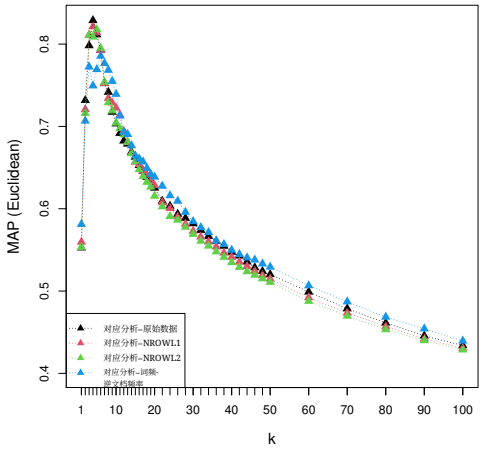
(a) $\alpha = 1$(BBC新闻)

(b) $\alpha = 1$(BBC体育)

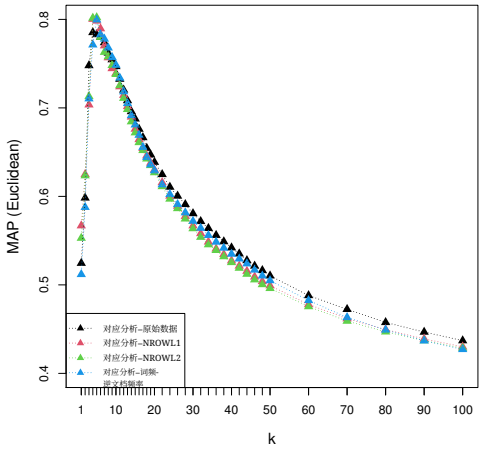
(c) $\alpha = 1$(20新闻组)

(d) $\alpha = 1$(Wilhelmus)

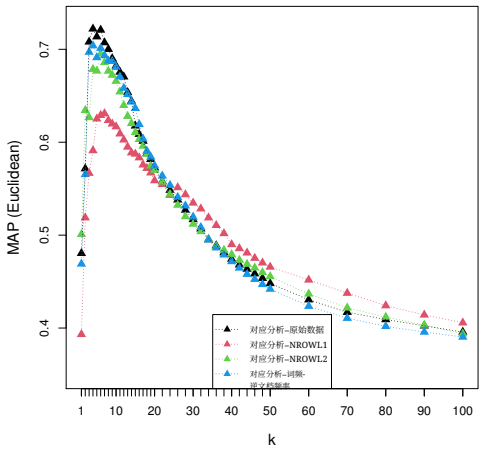
Figu图 A.1：标准坐标下，平均精度均值随维度数量的变化 k es.



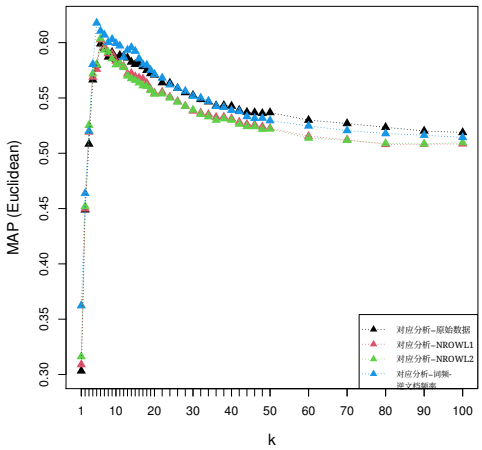
(a) $\alpha = 1$(BBC新闻)



(b) $\alpha = 1$(BBC体育)



(c) $\alpha = 1$(20新闻组)



(d) $\alpha = 1$(Wilhelmus)

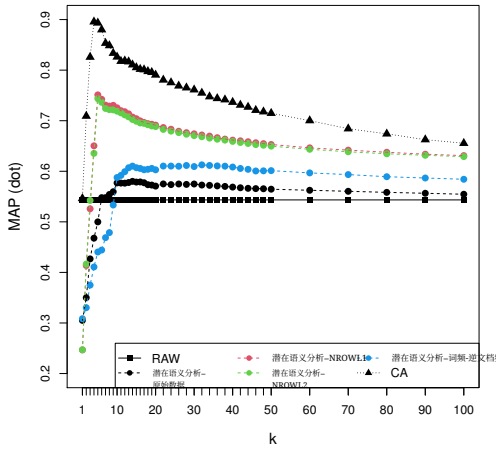
图 A.2：采用标准坐标时，四种对应分析版本的平均精度均值随维度数量的变化情况 k 。

补充材料 B：点积相似度

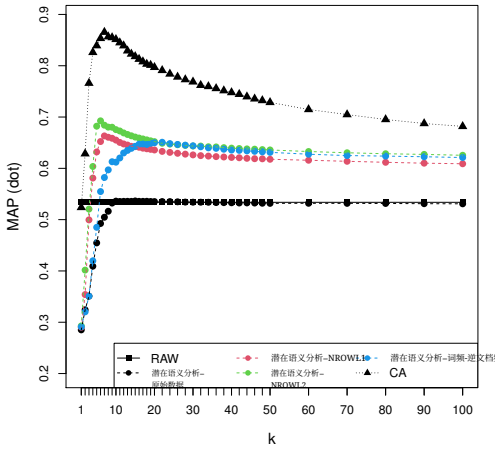
我们按照正文进行了相同的实验，但使用点积相似度而非欧几里得距离作为相似度测量方法。结果得出的结论与正文中使用欧几里得距离所得结论一致。

B. 1 比较潜在语义分析与对应分析在信息检索中的表现

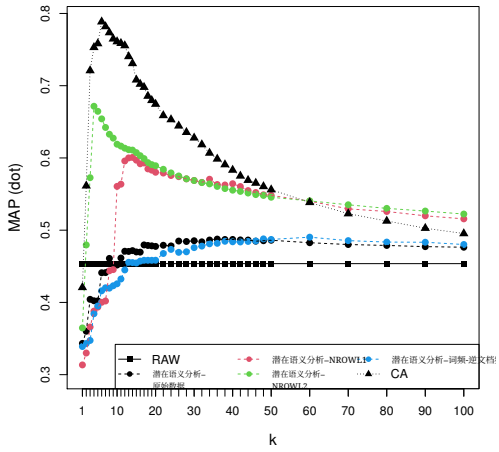
B.1.1 标准加权指数为 $\alpha = 1$ 时，四种潜在语义分析版本及对应分析的平均精度均值随维度数量的变化



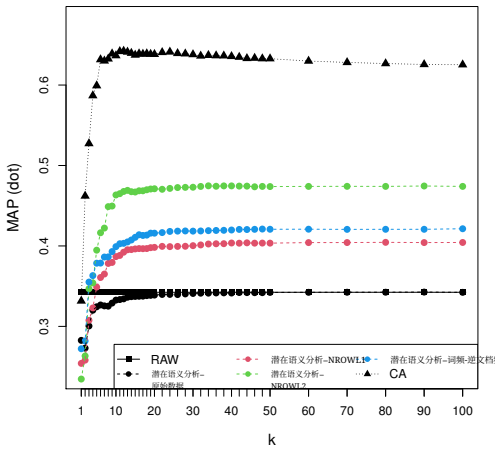
(a) BBC新闻



(b) BBC体育



(c) 20新闻组

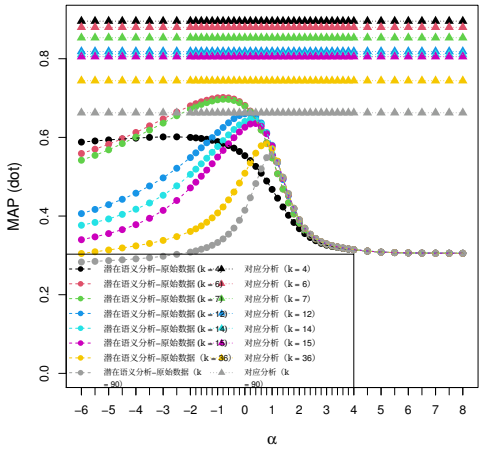


(d) Wilhelmus

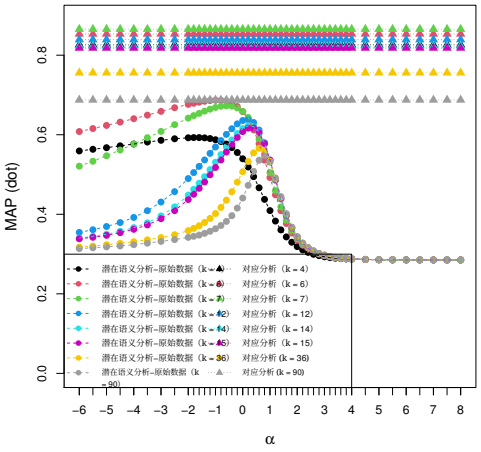
Fig图 B.1：标准坐标下平均精度均值随维度数量的变化 s.

B.1.2 加权指数 α 下平均精度均值随维度数量的变化：潜在语义分析的平均精度均值与对应分析的平均精度均值

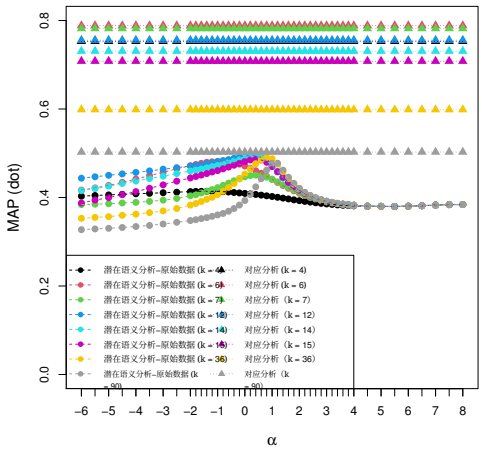
图 B.2 展示了 MAP 随以下维度数量的变化情况：关于 LSA-RAW 的 α 以及关于对应分析的 MAP： $k = 4$、6、7、12、14、15、36、和 90。



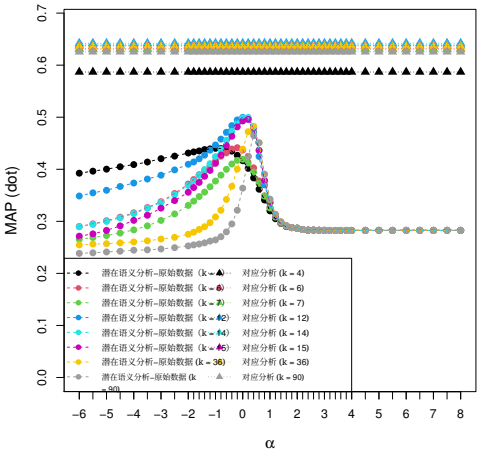
(a) BBC新闻



(b) BBC体育



(c) 20新闻组



(d) Wilhelmus

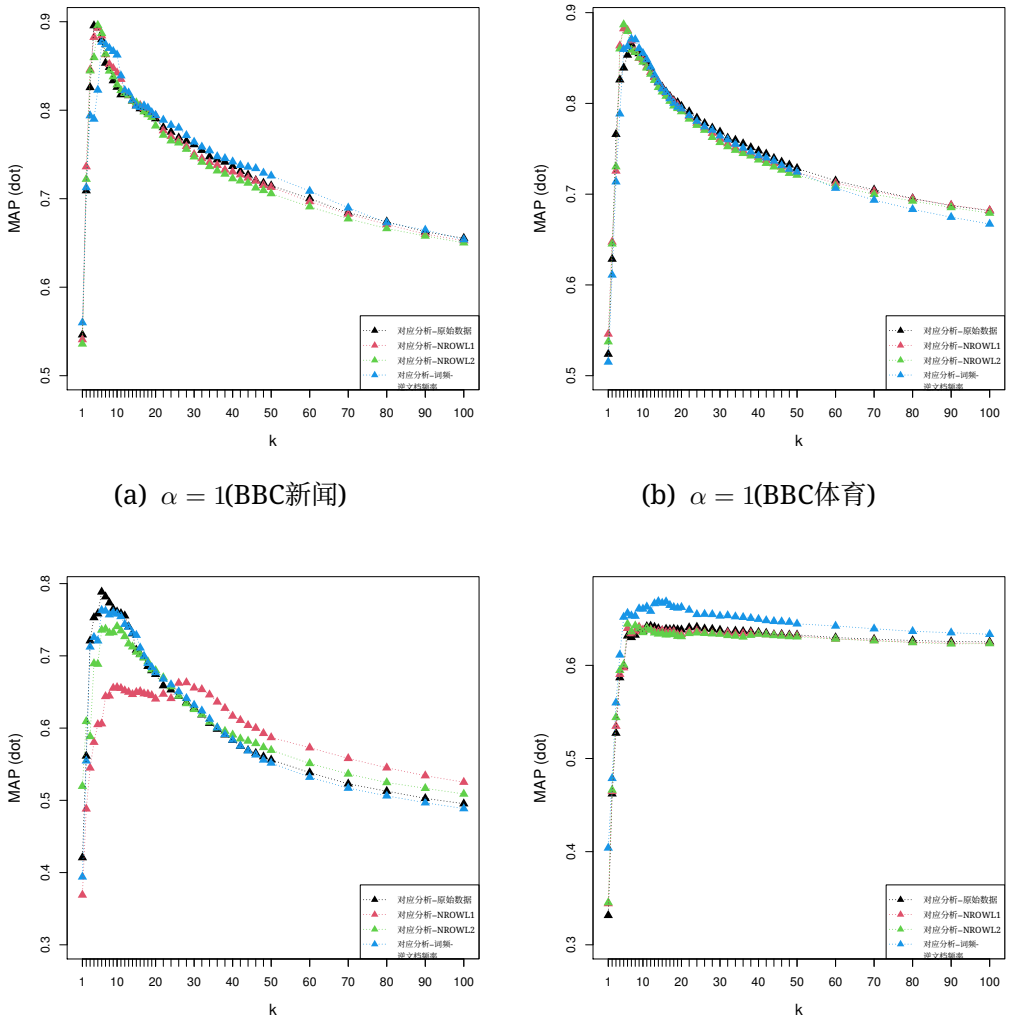
图 B.2：LSA-RAW 的平均精度均值随 α 变化，以及不同 k 值下对应分析的平均精度均值。

表 B.1：关于点积相似度，在 $k = 4$、6、7、12、14、15、36、和 90 个特征下，LSA-RAW 使用最优加权指数 α 时的平均精度均值，以及对应分析的平均精度均值。粗体数值为最佳结果。

	BBC新闻		BBC体育		20新闻组		Wilhelmus	
	α	平均精度均值	α	平均精度均值	α	MAP	α	MAP
LSA-RAW ($k = 4$)	-2.5	0.601	-1.8	0.593	-1.4	0.414	-1	0.440
LSA-RAW ($k = 6$)	-0.8	0.701	-1	0.687	-1	0.481	-0.2	0.441
LSA-RAW ($k = 7$)	-0.8	0.697	-0.6	0.673	0.4	0.450	0	0.419
LSA-RAW ($k = 12$)	0	0.660	0.2	0.638	0.2	0.495	0	0.500
LSA-RAW ($k = 14$)	0.2	0.646	0.2	0.623	0.4	0.492	0.2	0.500
LSA-RAW ($k = 15$)	0.4	0.635	0.4	0.617	0.4	0.486	0.2	0.496
LSA-RAW ($k = 36$)	0.8	0.584	0.6	0.566	0.8	0.490	0.4	0.482
LSA-RAW ($k = 90$)	1	0.556	0.8	0.547	1.2	0.478	0.4	0.465
对应分析 ($k = 4$)		**0.896**		0.826		0.753		0.587
对应分析 ($k = 6$)		0.880		0.853		**0.788**		0.632
对应分析 ($k = 7$)		0.853		**0.865**		0.782		0.630
对应分析 ($k = 12$)		0.818		0.839		0.756		**0.642**
对应分析 ($k = 14$)		0.811		0.823		0.731		0.640
对应分析 ($k = 15$)		0.805		0.818		0.708		0.637
对应分析 ($k = 36$)		0.744		0.756		0.599		0.637
对应分析 ($k = 90$)		0.663		0.687		0.503		0.626

B. 2 提升对应分析在信息检索中的性能

B. 2.1 对应分析中原始文档-词项矩阵的加权方案



(a) $\alpha = 1$(BBC新闻)　(b) $\alpha = 1$(BBC体育)

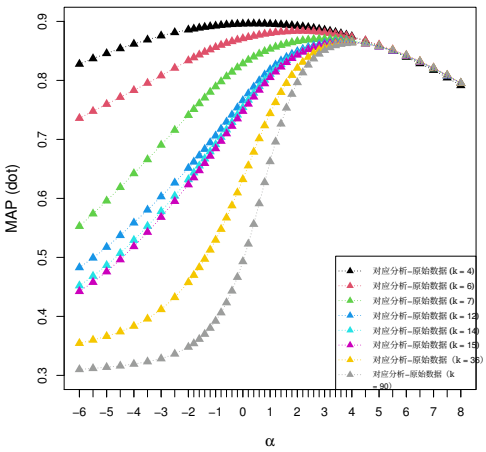
(c) $\alpha = 1$(20新闻组)　(d) $\alpha = 1$(Wilhelmus)

图 B.3：四种对应分析标准坐标版本的平均精度均值随维度数量变化的函数。 k

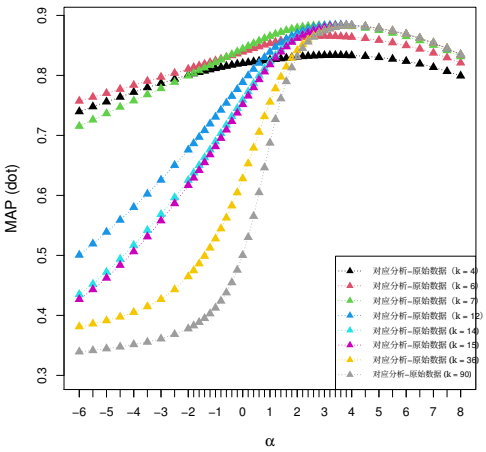
表 B.2：四种对应分析版本在点积相似度下，采用最优维度数量时的平均精度均值。粗体值为最佳结果。 k

	BBC新闻		BBC体育		20新闻组		Wilhelmus	
	k	平均精度均值	k	平均精度均值	k	MAP	k 值	平均精度均
CA-RAW	4	**0.896**	7	0.865	6	**0.788**	12	0.642
CA-NROWL1	5	0.893	5	0.882	28	0.663	9	0.641
CA-NROWL2	5	**0.896**	5	**0.887**	10	0.741	6	0.644
CA-TFIDF	6	0.877	7	0.871	6	0.763	14	**0.669**

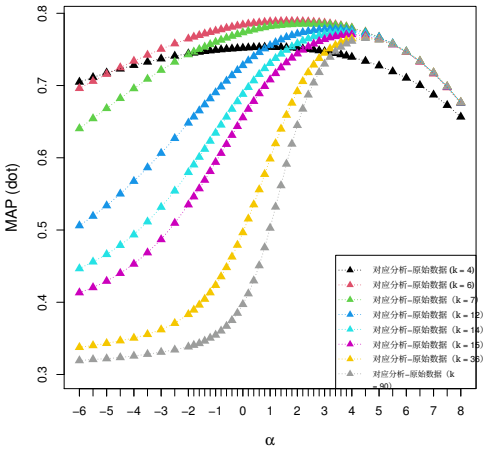
B. 2.2 对应分析中的加权指数 α



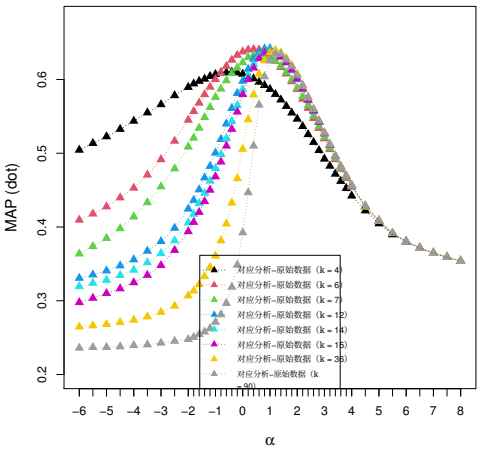
（a）BBC新闻



（b）BBC体育



（c）20新闻组



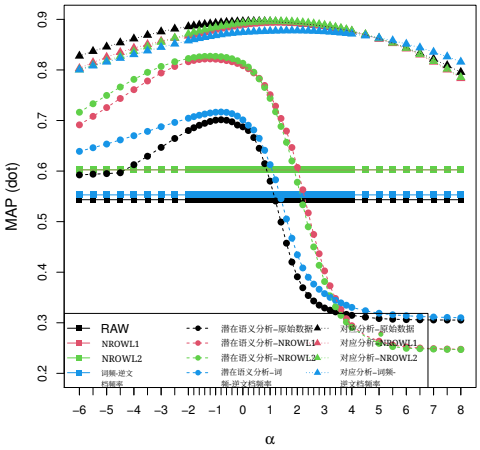
(d) Wilhelmus

图 B.4：CA-RAW 在不同 k 值下的平均精度均值随 α 的变化。

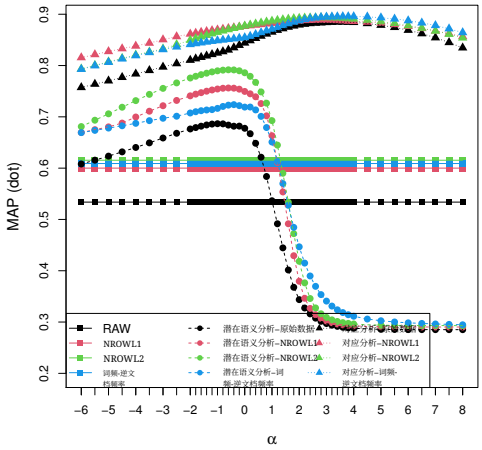
表 B.3：在 $k = 4$,6,7,12, 14,15, 36,和 90 的点积相似度下，CA-RAW 的最优 α 平均精度均值。粗体值为最佳值。

	BBC新闻		BBC体育 20新闻组				Wilhelmus	
	α	平均精度均值	α	平均精度均值	α	MAP	α	MAP
对应分析 ($k = 4$)	0.4	**0.896**	3.4	0.834	0.8	0.753	-0.4	0.610
对应分析 ($k = 6$)	2.2	0.884	3	0.866	1.6	**0.789**	0.4	0.641
对应分析 ($k = 7$)	3.2	0.870	3	0.883	2.4	0.785	0.6	0.635
对应分析 ($k = 12$)	3.8	0.866	3.6	**0.884**	3.4	0.779	1	**0.642**
对应分析 ($k = 14$)	3.8	0.866	3.8	0.883	3.8	0.774	1	0.640
对应分析 ($k = 15$)	3.8	0.865	3.8	0.883	4	0.771	1	0.637
对应分析 ($k = 36$)	4	0.864	4	0.883	4.5	0.767	1.2	0.640
对应分析（$k = 90$）	4	0.863	4	0.883	4.5	0.765	1.2	0.635

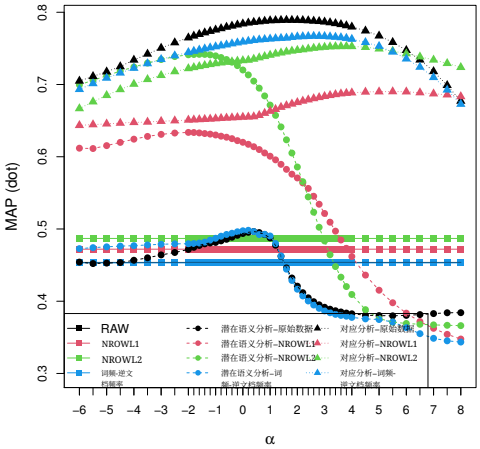
B.3 探究在潜在语义分析和对应分析的最佳维度数下，平均精度均值随 α 变化的情况



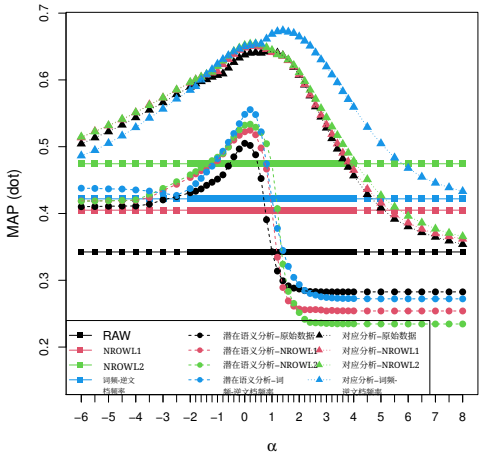
(a) BBC新闻



(b) BBC体育



(c) 20新闻组



(d) Wilhelmus

图 B.5：在最优维度下，平均精度均值随 α 变化的函数。

表 B.4：关于点积相似度，在最优 α 和最优维度 k 下的平均精度均值。粗体值为组内最佳；带下划线的值为总体最佳。

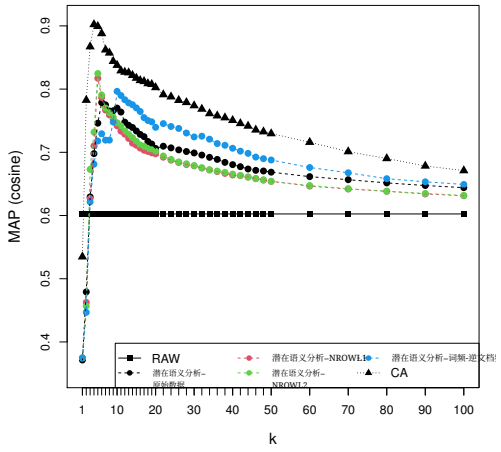
	BBC新闻			BBC体育			20新闻组			Wilhelmus		
	α	k 值	平均精度均	α	k 值	平均精度均	α	k	平均精度均值	α	k 值	平均精度均
RAW			0.543			0.534			0.454			0.342
LSA-RAW	-0.8	6	0.701	-1	6	0.687	0.4	17	0.496	0	13	0.505
CA-RAW	0.4	4	**0.896**	3.6	10	**0.885**	1.6	6	**0.789**	1	12	**0.642**
NROWL1			0.603			0.600			0.472			0.405
LSA-NROWL1	-1.2	5	0.822	-0.6	5	0.756	-2	12	0.634	0.2	11	0.525
CA-NROWL1	1.4	5	**0.893**	2.8	6	**0.889**	5.5	32	**0.690**	0.2	6	**0.651**
NROWL2			0.602			0.615			0.487			0.474
LSA-NROWL2	-1.2	5	0.827	-0.6	5	0.792	-1.6	4	0.742	0.2	10	0.534
CA-NROWL2	1.2	5	**0.896**	3	6	**0.893**	3.8	10	**0.753**	0.4	6	**0.654**
词频-逆文档频率			0.553			0.609			0.453			0.422
LSA-TFIDF	-0.8	10	0.717	-0.4	9	0.724	0.2	24	0.498	0.2	16	0.555
CA-TFIDF	2	6	**0.878**	4	10	**0.896**	2.6	9	**0.767**	1.4	16	**0.674**

补充材料 C：余弦相似度

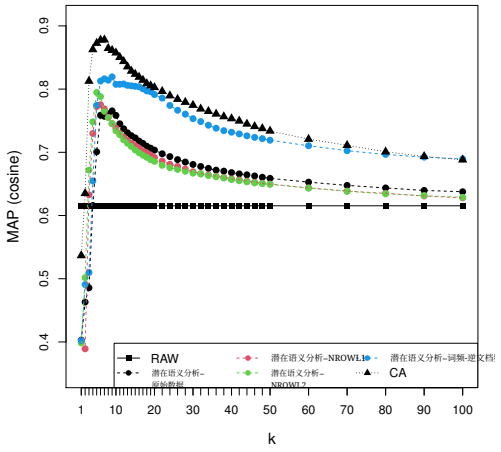
我们进行了与正文相同的实验，但使用余弦相似度而非欧几里得距离作为相似度测量方法。结果与正文的趋势一致，并得出了相似的结论。

C. 1 比较潜在语义分析与对应分析在信息检索中的表现

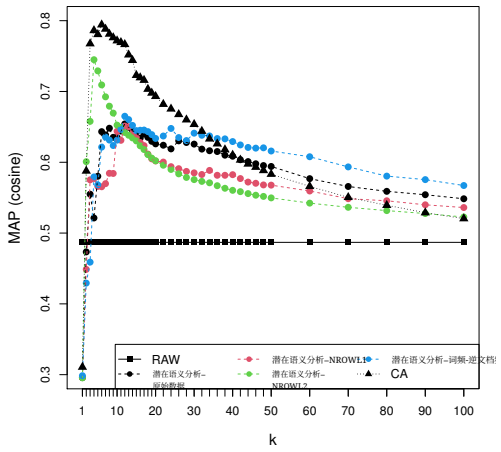
C.1.1 标准加权指数为 $\alpha = 1$ 时四种潜在语义分析版本及对应分析的平均精度均值随维度数量的变化



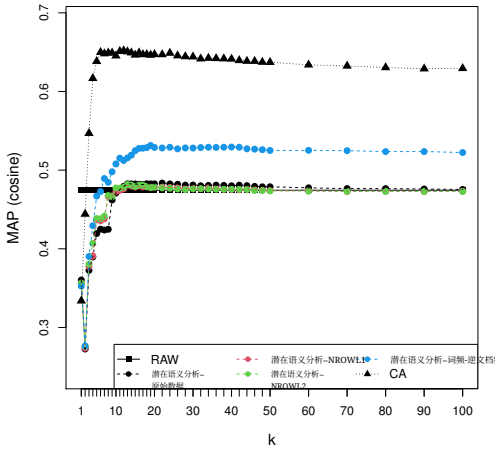
(a) BBC新闻



(b) BBC体育



(c) 20新闻组

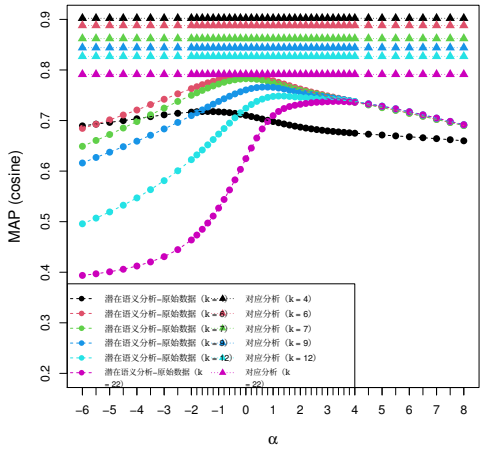


(d) Wilhelmus

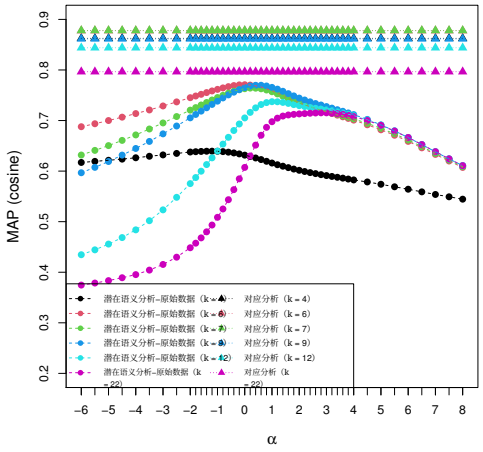
Fig图 C.1：标准坐标下，平均精度均值随维度数量的变化 s.

C. 1.2 潜在语义分析的加权指数 α 与对应分析的平均精度均值随维度数量的变化

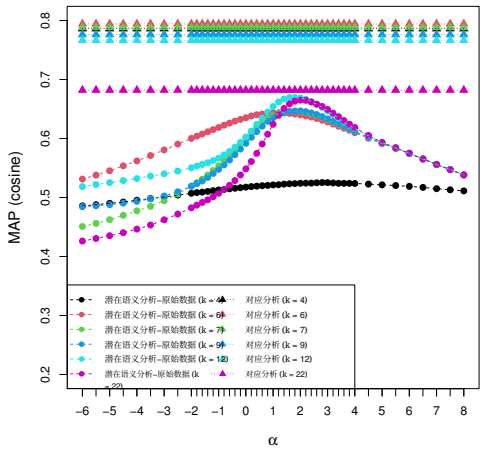
图 C.2 显示了平均精度均值随 α 变化：潜在语义分析-原始数据与对应分析的平均精度均值随维度数量的变化： $k = 4$、 6、 7、 9、 12、和 22。



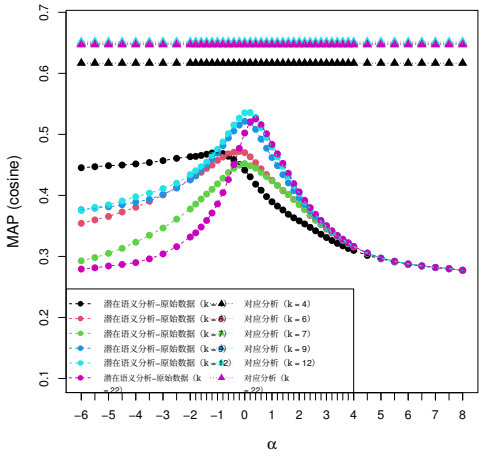
(a) BBC新闻



(b) BBC体育



(c) 20新闻组



(d) Wilhelmus

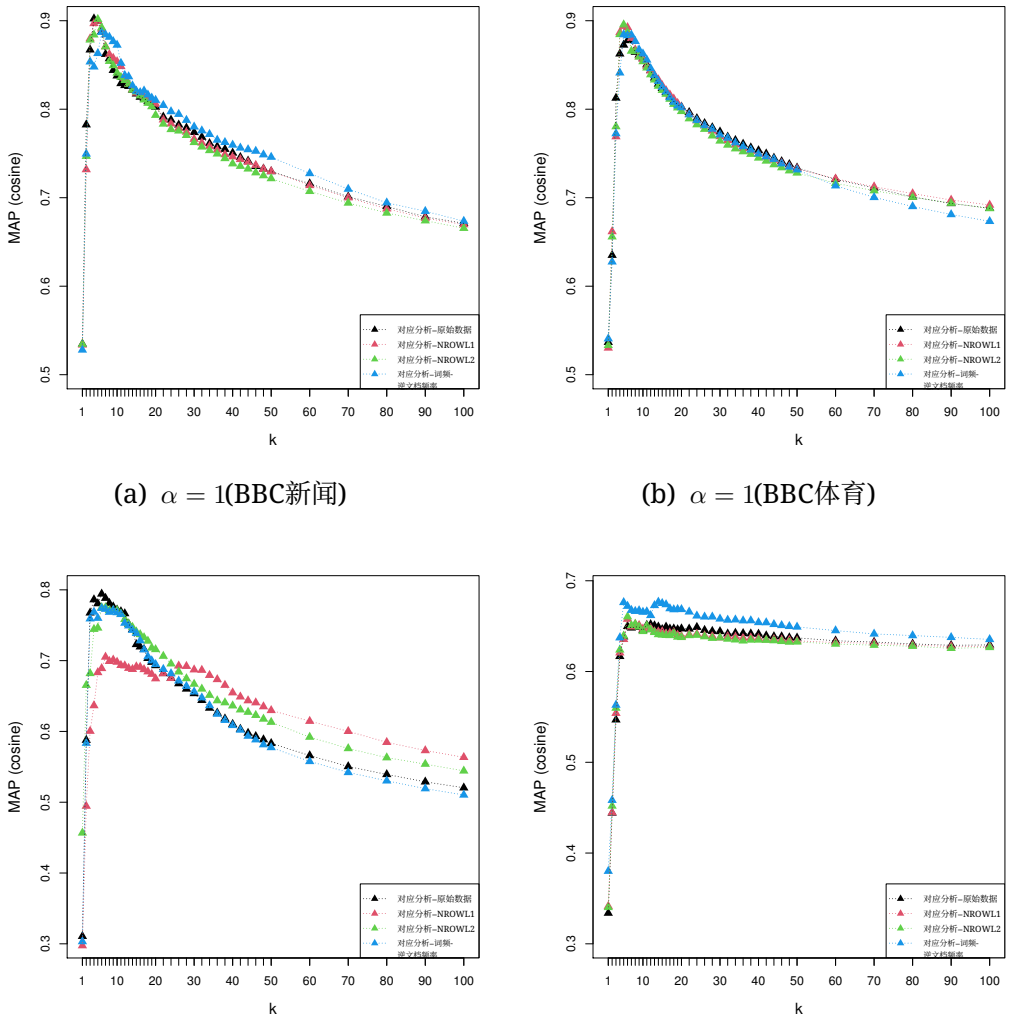
图 C.2：LSA-RAW 的平均精度均值随 α 变化的情况，以及对应分析在不同 k 值下的平均精度均值。

表 C.1：LSA-RAW 在最优加权指数 α 下的平均精度均值，以及对应分析在 $k = 4$, 6,7,9,12, 和 22 个余弦相似度值下的平均精度均值。粗体数值为最佳结果。

	BBC新闻		BBC体育		20新闻组		Wilhelmus	
	α	平均精度均值 α		平均精度均值 α		MAP	α	MAP
LSA-RAW ($k = 4$)	-1.4	0.718	-1.2	0.639	2.8	0.525	-1	0.470
LSA-RAW ($k = 6$)	0	0.788	0	0.771	1.2	0.644	-0.2	0.472
LSA-RAW ($k = 7$)	0	0.783	0.2	0.764	1.6	0.646	0	0.452
LSA-RAW ($k = 9$)	0.8	0.766	0.6	0.770	2	0.647	0	0.521
LSA-RAW ($k = 12$)	1.2	0.748	1	0.737	1.8	0.670	0.2	0.536
LSA-RAW ($k = 22$)	3.4	0.738	2.8	0.715	2	0.665	0.4	0.525
对应分析 ($k = 4$)		**0.902**		0.863		0.786		0.617
对应分析 ($k = 6$)		0.888		0.878		**0.794**		0.650
对应分析 ($k = 7$)		0.862		**0.878**		0.788		0.648
对应分析 ($k = 9$)		0.844		0.861		0.776		0.649
对应分析 ($k = 12$)		0.827		0.844		0.767		**0.652**
对应分析（$k = 22$）		0.791		0.796		0.682		0.647

C.2 利用对应分析提升信息检索性能

C. 2.1 对应分析所用原始文档-词项矩阵的加权方案



(a) $\alpha = 1$(BBC新闻)　(b) $\alpha = 1$(BBC体育)

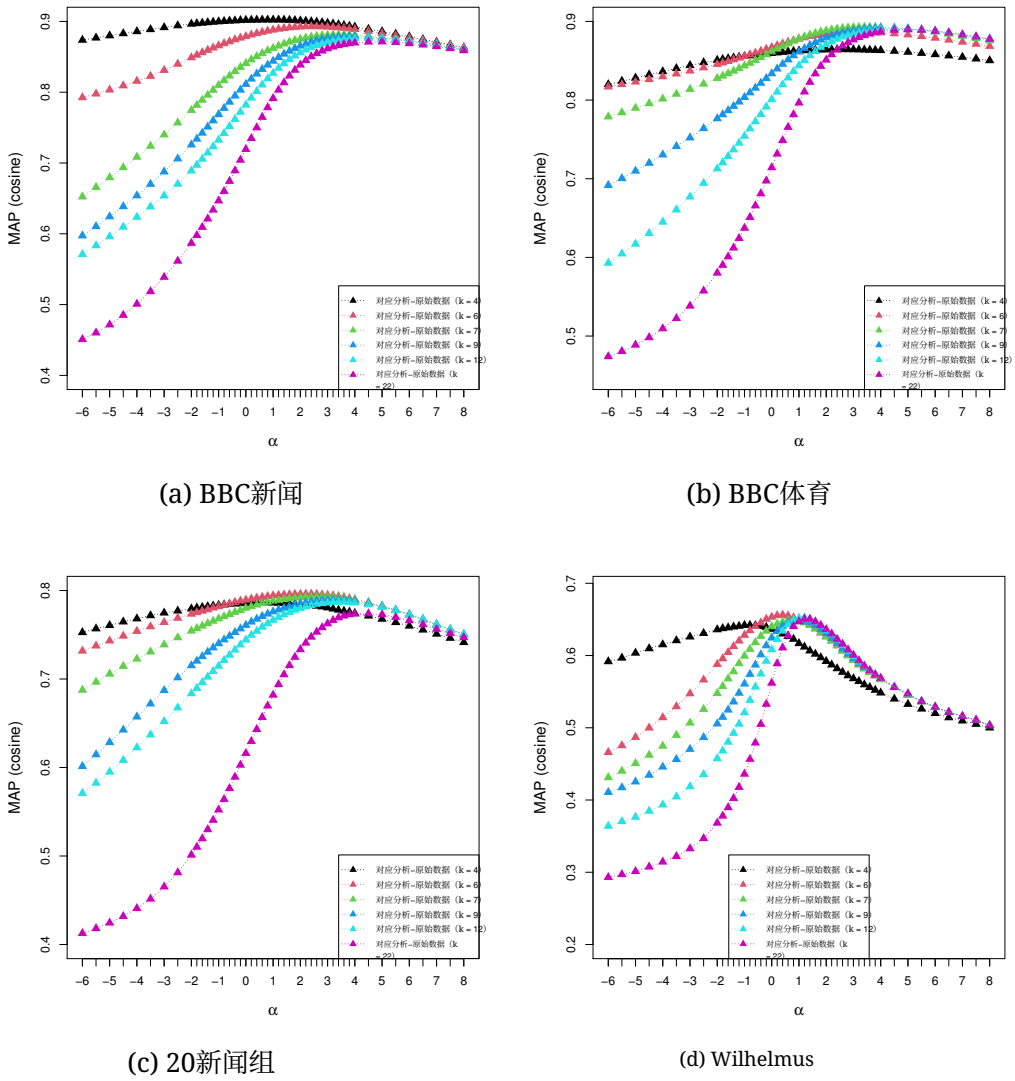
(c) $\alpha = 1$(20新闻组)　(d) $\alpha = 1$(Wilhelmus)

图 C.3：采用标准坐标时，对应分析四种版本的平均精度均值随维度数量的变化。k

表 C.2：关于余弦相似度，CA 四个版本在最佳维度数 k 下的平均精度均值。粗体值为最佳值。

	BBC新闻		BBC体育		20新闻组		Wilhelmus	
	k	平均精度均值	k	平均精度均值	k	MAP	k 值	平均精度均
CA-RAW	4	**0.902**	7	0.878	6	**0.794**	12	0.652
CA-NROWL1	5	0.900	5	0.894	7	0.705	6	0.658
CA-NROWL2	5	0.902	5	**0.896**	6	0.775	6	0.660
CA-TFIDF	6	0.887	5	0.884	6	0.774	14	**0.677**

C. 2.2 对应分析中的加权指数 α



(a) BBC新闻　(b) BBC体育

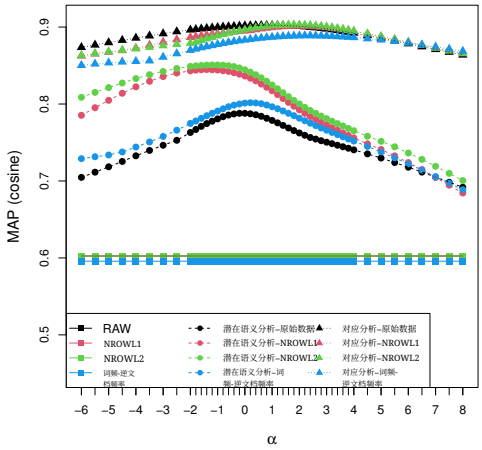
(c) 20新闻组　(d) Wilhelmus

图 C.4：CA-RAW 的平均精度均值随 α 变化的情况，其中 k 取不同值。

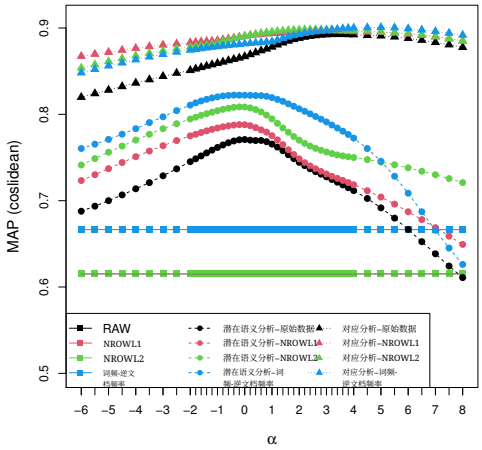
表 C.3：关于余弦相似度，在 $k = 4$,6,7,9,12, 和 22 下，CA-RAW 的最优 α 对应的平均精度均值。粗体数值为最佳结果。

	BBC新闻		BBC体育		20新闻组		Wilhelmus	
	α	平均精度均值	α	平均精度均值	α	MAP	α	MAP
对应分析 ($k = 4$)	0.8	**0.902**	2.6	0.864	0.8	0.786	-0.8	0.642
对应分析（$k = 6$）	2.4	0.893	3.2	0.887	2.4	**0.796**	0.4	**0.656**
对应分析（$k = 7$）	3.6	0.881	3.4	**0.893**	2.8	0.793	0.8	0.649
对应分析（$k = 9$）	3.8	0.879	4	0.892	3.4	0.789	1	0.649
对应分析（$k = 12$）	4.5	0.876	4.5	0.890	3.6	0.787	1.2	0.652
对应分析（$k = 22$）	5	0.871	5	0.890	4.5	0.774	1.2	0.651

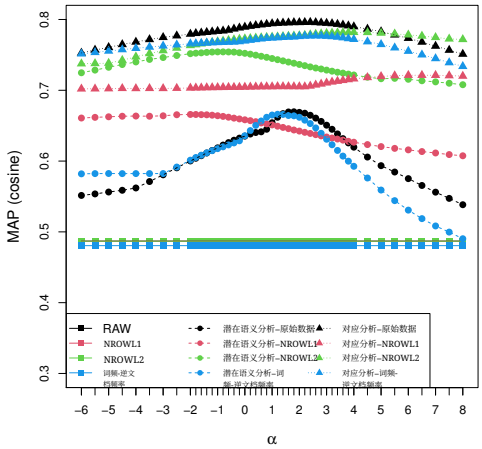
C. 3 探究在潜在语义分析和对应分析的最佳维度数下，平均精度均值随 α 变化的情况



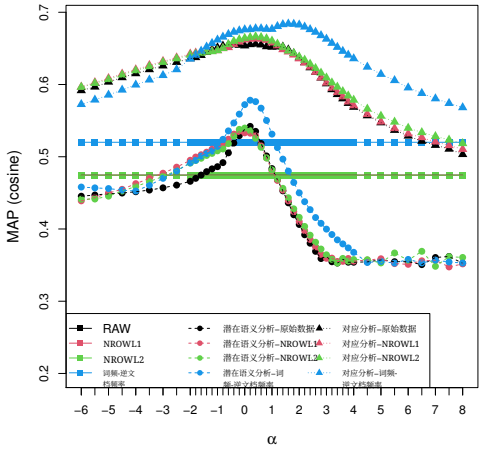
(a) BBC新闻



(b) BBC体育



(c) 20新闻组



(d) Wilhelmus

图 C.5：最优维度下，平均精度均值随 α 的变化。

表 C.4：在余弦相似度下，最优 α 和最优维度 k 对应的平均精度均值。粗体数值为组内最佳值；带下划线的数值为总体最佳值。

	BBC新闻			BBC体育			20新闻组			Wilhelmus		
	α	k 值	平均精度均	α	k	平均精度均值	α	k 值	平均精度均	α	k 值	平均精度均
RAW			0.602			0.615			0.487			0.474
LSA-RAW	0	6	0.788	0	6	0.771	1.8	12	0.670	0.2	13	0.542
CA-RAW	0.8	4	**0.902**	3.4	7	**0.893**	2.4	6	**0.796**	0.4	6	**0.656**
NROWL1			0.602			0.615			0.487			0.474
LSA-NROWL1	-1.4	5	0.845	-0.2	5	0.788	-2	12	0.666	-0.2	8	0.535
CA-NROWL1	2	5	**0.901**	2.8	6	**0.897**	6.5	32	**0.721**	0.4	6	**0.664**
NROWL2			0.602			0.615			0.487			0.474
LSA-NROWL2	-1	5	0.851	-0.2	5	0.809	-1	4	0.754	0.0	10	0.540
CA-NROWL2	1.8	5	**0.903**	2.6	5	**0.898**	4	10	**0.782**	0.4	6	**0.666**
词频-逆文档频率			0.596			0.667			0.481			0.520
LSA-TFIDF	0.2	10	0.802	-0.4	6	0.822	1.2	12	0.666	0.2	16	0.578
CA-TFIDF	2.4	6	**0.889**	5	10	**0.901**	2.6	9	**0.777**	1.6	14	**0.684**