利用引文网络评估文本长度对相关概念识别的影响

Jorge A. V. Tohalino

圣保罗大学数学与计算机科学研究所，圣卡洛斯，圣保罗州，巴西

Thiago C. Silva

巴西利亚天主教大学，巴西利亚，联邦区，巴西

Diego R. Amancio*

数学与计算机科学研究所，计算机科学系，圣保罗大学 ˜圣保罗 ˜圣卡洛斯，圣保罗州，巴西（日期：2023年1月18日）

arXiv:2301.06168v1 [cs.DL] 15 Jan 2023

摘要

在各种实际应用中，识别非结构化数据中最重要的概念至关重要。尽管已有大量方法被提出用于提取文本的主题，但只有少数研究探讨了文本长度对关键词提取（KE）方法表现的影响。本研究采用一种基于网络的方法，评估从论文摘要中提取的关键词是否与从全文论文中提取的关键词相符。我们使用社区检测方法，在引文网络中识别相关论文组成的群组。随后利用这些论文簇从摘要中提取关键词。结果表明，尽管我们在关键词提取方法中采用的各种社区检测方法取得了相近的准确率，但相关性分析显示，这些方法为每篇摘要生成了不同的关键词列表。我们还发现，所有考察的方法准确率都偏低。令人意外的是，文本聚类方法的表现优于所有基于引文的方法。研究结果表明，使用不同的信息来源提取关键词可能导致表现存在显著差异，而这一影响对于依赖识别相关概念的应用而言可能十分重要。

* diego@icmc.usp.br

I. 引言

随着互联网上大量文本内容日益丰富，对文本进行高效分析的需求也变得迫切。在线文本数据涵盖了各种规模和类型，包括书籍、百科全书和报纸。在过去几十年里，以短文本形式出现的用户生成内容也呈指数级增长。这类内容包括社交媒体消息、产品描述、在线评论以及研究论文 [21]。为了概括这些海量信息，关键词提取（KE）已成为一项重要的自然语言处理（NLP）应用。关键词提取旨在从给定文档中识别最具信息量且最相关的词语或主题 [39]。关键词是实用工具，可以帮助用户快速了解文本的整体内容。此外，关键词提取在多种自然语言处理应用中发挥着重要作用，包括文本分类、文档摘要、文档标注、推荐系统、语音识别等 [21, 22]。

关键词提取（KE）任务已成为众多研究的主题。这些研究大致可分为三类：统计方法、语言学/句法方法和基于图的方法。不同范式也被结合起来，用于监督分类：提取出的特征被输入机器学习算法，以执行二分类任务 [18]。尽管这些方法在处理长文本方面已显示出有效性，但将其应用于稀疏度较高的短文本时，仍面临重大挑战 [9]。

在短文本中识别关键词，尤其是科学手稿中的关键词，是一项重大挑战；当使用开放学术数据集，且仅提供标题和摘要作为文本信息来源时，这一挑战尤为突出 [17]。从短文本中提取关键词的挑战，尤其是在科学论文中，推动了一些方法的开发。一种旨在解决摘要稀疏性的方法，是使用聚类技术对摘要进行分组 [37]。具体做法是利用引文作为代理，判断论文之间的相似性，从而避免直接比较短文本。尽管使用了此类聚类方法和其他外部信息 [9, 21], ，但目前仍缺乏全面研究来比较从摘要和全文中提取的关键词之间的兼容性。准确的表示...

科学论文中所蕴含语义信息的表征在许多领域都至关重要，因为它是许多科学计量学研究的基础。因此，本研究旨在探讨以下研究问题：

1. 从摘要中提取的关键词与从相应全文论文中提取的关键词有多大程度的相似性？2. 不同社区检测方法提取的关键词集合是否一致？3. 与直接根据文本信息评估摘要相似度相比，使用引文是否能带来更优的表现？

我们采用聚类方法从摘要中提取关键词，并将其与从相应全文中提取的关键词进行比较。我们利用引文网络，评估了多种成熟的社区检测方法在识别相关论文组以进行关键词提取方面的表现。我们还评估了不依赖引文信息的聚类方法，包括基于神经嵌入的技术。

这项研究揭示了若干有趣的结果。研究发现，所有受评估的方法提取的关键词与全文中的关键词均存在显著差异。我们观察到，仅依赖文本信息的聚类方法优于基于引文网络的方法，这表明引文可能并非衡量语义相似度的最佳替代指标。此外，结果表明，尽管不同方法识别出的关键词集合存在差异，但受评估的各种社区检测策略表现相近。

总而言之，我们的研究结果表明，用于提取关键词的信息量会显著影响这项任务的表现。因此，在条件允许时，使用相似性网络的研究应考虑采用全文，以便从论文网络中提取有关主题的更可靠信息。

本文结构如下：第二节全面回顾关键词提取领域中最相关的研究。第三节概述了从短文本和长文本中提取关键词的方法论，并介绍所采用的数据集。主要结果将在

第四节中呈现并讨论。最后，我们将在第五节总结结论，并提出未来研究的潜在方向。

II. 相关工作

早期针对关键词提取问题的研究侧重于统计方法。词在文本中的空间分布被用来衡量词语的相关性 [7]。最简单的方法基于词频，将词频较高的词视为关键词。然而，这些方法没有考虑词序，因此，如果将文本打乱，毫无意义的文本版本仍会生成相同的关键词集合。随后，人们提出了结合词频与空间分布的方法，通过词聚类和熵来解决这一问题 [7, 29]。这些方法背后的理念是，重要词语通常集中在文本中某些特定部分，而这些部分正是主题所在。因此，无关词语会较为均匀地分布在文本中，而关键词的分布不均，并且往往会形成语义群组。对基于词频的方法的另一项改进是词频-逆文档频率方法，它根据某个词在一篇文本中的词频以及该词在整个数据集中的词频来衡量其重要性。这些方法的主要优势在于简单易用，且不需要外部语料库或语言知识。

基于图的方法也被用于对文本进行建模 [23]。多项研究将文档表示为词共现网络，以解决关键词提取问题：如果两个词在给定上下文中共同出现，它们便会相连 [40]。随后使用中心性指标为每个词分配重要性值。在 [20], 中，作者得出结论，网络指标能够成功提取与关键词提取任务相关的词。他们还指出，基于网络的方法无需使用外部语料库，且不受语言限制。使用词嵌入和较大的上下文也有助于提升共现网络在提取关键词时的质量 [40]。 [16], 提出了另一种方法，将社群检测方法应用于语义关系网络。作者使用维基百科来确定文档中各词之间的语义相关性。根据 [16], ，重要的词往往聚集在连接紧密的社群中，而这些社群与文档的主题相关。

为了解决短文本中的关键词提取问题，多项研究借助语义和背景知识。根据 [9], ，仅从词语中提取基本或直观的特征，不足以从短文本中找出关键词。在 [21], 中，作者指出，文本聚类方法也有助于解决短文本的语义稀疏性。这些技术能够将相关文本聚类，从而通过汇总同一簇中的文本提取更多语义信息。 [44]采用聚类算法识别每个簇中最相关的词，其假设是主题相似的文本包含相似的关键词。随后，他们对每个文本簇应用基于图的方法，并使用 PageRank算法提取关键词。

在 [39], 中，作者提出了一种基于每个词的信息量来选择关键词的方法。该分数在语料库、聚类和文档三个层面上计算。在语料库层面，信息量的计算考虑了所有文档；而在聚类层面，则在一组相关文本中计算词的重要性。随后，利用前述步骤的结果计算文档层面的信息量。这种方法在关键词提取方面表现良好。

关于基于图的技术，一些研究通过将词语之间不同的语义关系纳入词排序算法，并将其作为节点权重，改进了 TextRank [24] 。例如， [22] 使用维基百科作为外部知识库，而 [21] 则采用 Word2Vec 和 Doc2Vec 嵌入模型来计算词语之间的语义相似度。

尽管许多研究分别关注从短文本或长文本中提取关键词，但本文对从短文本和长文本中提取关键词的成熟方法进行了比较分析。我们重点考察从研究论文摘要中提取的关键词与从全文内容中提取的关键词之间的兼容性。

III. 材料与方法

所提出的关键词提取框架包含以下主要步骤：i）文本预处理；ii）网络构建；iii）社区检测；iv）短文本关键词提取；以及 v）长文本关键词提取。以下将对这些步骤进行总结，并

在图1中进行说明。

1. 文本预处理：此阶段包括文本处理和向量化步骤。第一步是移除停用词。对剩余词语进行词干提取，并采用词频-逆文档频率方法，将预处理后的文本转换为向量形式。关于所采用预处理步骤的更多详情，请参见第三节B。

2. 网络构建：我们首先构建了一个论文引文网络，用于短文本关键词提取。我们还将每篇论文的完整内容建模为词共现网络，用于从长文本中提取关键词。在第三节C中，我们将介绍创建这两种网络模型所需的步骤。

3. 社区检测：我们对引文网络应用社区检测方法，以找出相关论文簇（见第三节D）。

4. 短文本关键词提取：此阶段负责从短文本（论文摘要）中提取关键词。该阶段将使用上一步得到的簇。我们分别在社群内外计算每个词语的相关性。此外，我们还提出了两种基于词频-逆文档频率和K-Means算法的关键词提取方法。短文本关键词提取方法详见第三节E。

5. 长文本关键词提取：为了识别参考关键词，我们将论文的完整内容作为若干知名关键词提取方法的输入。我们评估了基于词频、词频-逆文档频率、熵、间歇性、BERT、Yake 和 TextRank [1, 5, 7, 14, 24] 的方法。我们还采用了基于共现网络和中心性指标的网络方法，为长文本提取关键词。这些网络通过中心性指标进行表征。所采用方法论的详细说明见第三节F。

A. 数据集

使用了以下两个数据集：

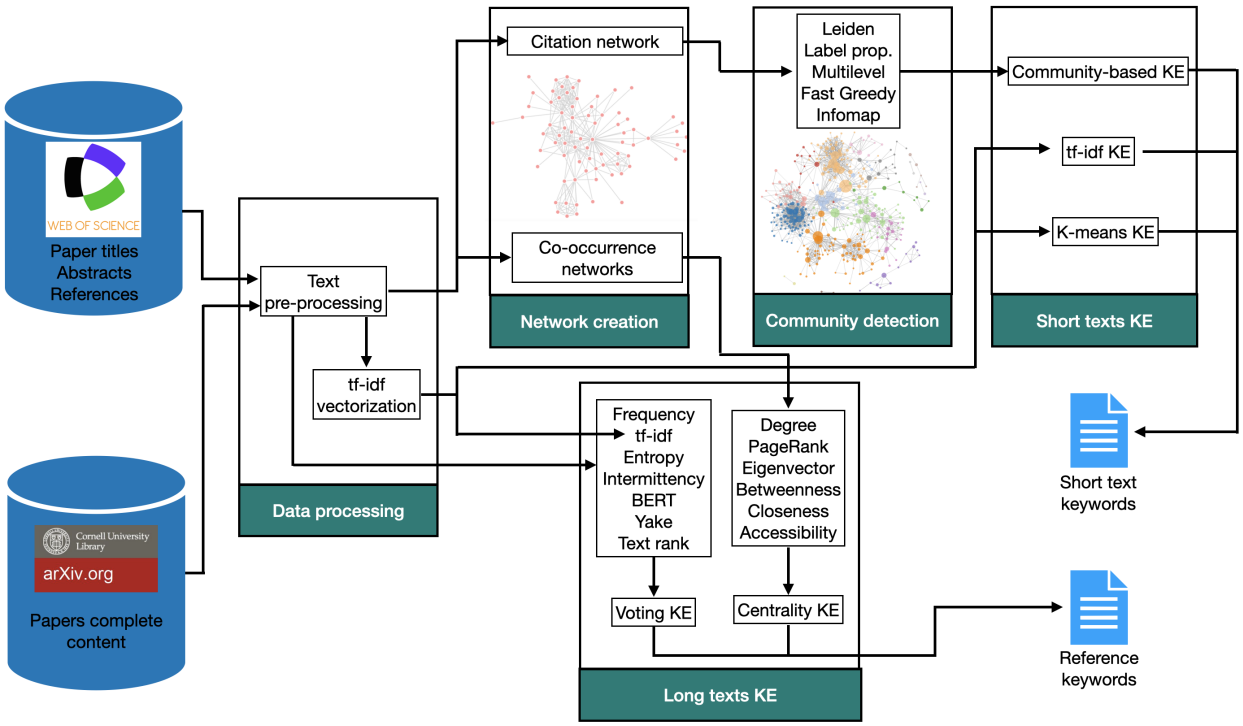


图 1：所采用的关键词提取系统包含多个阶段。首先对文本进行预处理。随后构建引文网络，并采用社区检测算法从短文本中提取关键词。为进行比较评估，我们实现了词频-逆文档频率和 K-Means 两种方法。这两种方法都不考虑引文来对论文进行聚类。为生成参考关键词，我们结合统计方法和传统关键词提取方法处理全文。此外，我们还评估了一种作为关键词提取替代方法的共现网络方法。

1. 短文本关键词提取数据集：我们使用了在 [37] 中收集的数据集。作者从复杂网络领域的 11,063 篇论文中检索了相关信息。数据来自 *Web of Science* （WOS）数据库 [27]。所选论文发表于 1991 年至 2013 年。对于每篇论文，作者都提取了标题、摘要和参考文献列表。后者用于构建引文网络。每篇论文的标题和摘要均作为关键词提取技术的输入。

2. 长文本关键词提取数据集：为了为每篇论文的摘要生成参考关键词列表，我们收集了每篇论文的全文内容，包括引言、方法、结果与讨论、结论以及附录部分。我们使用 arXiv数据库 [28] 的应用程序接口提取论文的完整内容。我们以每篇论文的标题和摘要作为关键词，通过 arXiv 应用程序接口进行自动检索。

表 I 总结了数据集的统计信息。所提供的信息是根据预处理后的版本计算得出的。

表 I：数据集的统计信息。 $|D|$ 表示文档数量。我们还列出了词元的平均数量（W_{avg}），以及词元数量分布的第一（W_{q1}）和第三（W_{q3}）四分位数。类似地， U_{avg} 表示平均词汇量，而 U_{q1}和 U_{q3} 分别表示词汇量分布的第一和第三四分位数。

数据集	**描述**	$\|D\|$	W_{avg}	W_{q1}	W_{q3}	U_{avg}	U_{q1}	U_{q3}
短文本	论文摘要	11,063	79.12	59.00	95.00	55.97	43.00	66.00
长文本	论文全文	1,982	2,020.58	1,170.25	2,405.00	517.64	394.00	595.00

B. 数据处理

此阶段包括三个步骤：数据准备、文本预处理和词频-逆文档频率向量化。数据准备步骤包括处理从 arXiv数据库中获取的论文（论文全文）。我们获取了每篇论文的 LaTeX 版本，因此必须移除所有 LaTeX 标签。我们还从清理后的文本中移除了作者名单、机构和致谢。全文论文分析纳入了以下部分：引言、相关工作、方法论（或材料与方法）、结果、讨论、结论和附录部分。

对数据集中的所有文本应用了文本预处理转换。我们首先移除了停用词和标点符号。然后，对剩余的词进行词干提取。此步骤旨在将每个词映射到其词根或词干 [30]。我们使用词频-逆文档频率技术将预处理后的文本转换为稀疏向量表示。为计算一个词 w的重要性，该技术会考虑 w 在单个文档中的内部频率。此外，还会将内部频率与 w 在数据集所有文档中的相对频率进行比较 [36]。文档 d 中 w的词频-逆文档频率表示计算如下

$$\text{tf-idf}(w,d) = \frac{f(w,d)}{n_d} \cdot \frac{\log N}{\log(N_w)}, \tag{1}$$

其中， $f(w,\ d)$表示 d中的词频 w， d中的词数为 n_d， N表示数据集中文档的总数，而 N_w 表示文档数量

其中 w 至少出现一次。

我们将每篇摘要的 tf-idf向量表示用作一种基于 K-Means 的短文本关键词提取方法的输入值。此外，我们还使用论文全文内容中每个词的 tf-idf权重，为一种长文本关键词提取方法（LKE）赋予重要性值。

C. 网络构建

鉴于论文摘要篇幅较短，无法从单篇文本中提取具有统计显著性的信息。因此，我们采用网络表示技术来提取补充信息，以增强关键词提取过程。通过分析网络的拓扑和结构属性，我们能够推断文本的特征，从而确定每个词的相对重要性。

本研究使用了两种不同的网络模型。为了将短文本聚类为相关论文组，我们在短文本关键词提取任务中使用了引文网络。在这种情况下，网络结构代表整个文档数据集。相反，从长文本中提取关键词时，则将每篇文本建模为词共现网络 [8, 19, 38]。

无权论文引文网络按照 [37] 中描述的方法论构建。引文网络旨在呈现一种语义相似性结构，不使用文本信息来建立论文之间的联系。最终得到的网络由 11,063 个节点和 94,472 条边组成。随后，利用该网络的社区结构以及标题和摘要信息，检测各网络社区中最重要的词。

论文的全文内容被建模为词共现网络。在该图模型中，每个节点代表一个词，两个节点之间的边基于两个词的邻域关系建立。我们采用了可纳入虚拟连接的方法，使相似的词能够相互连接。这一模型及其变体已应用于许多不同场景 [11, 13, 15, 31]。我们使用广为人知的中心性度量来描述这些网络，并根据词在网络中的结构重要性对其进行排序 [26]。

D. 社区检测

此阶段负责检测社群，即通过引用链接连接起来的论文簇。与网络中的其他节点相比，社群中的节点彼此之间的连接更为密集 [32]。识别大型网络中的社群是一项十分有用的任务。例如，属于同一社群的节点很可能具有若干共同属性。此外，发现的社群数量及其各自的特征有助于在分类任务中识别网络类别 [45]。识别社群也有助于理解网络的动态演化和组织结构 [12]。本文评估了以下方法：多层次、标签传播、Infomap、快速贪心和 Leiden 方法 [4, 10, 33, 35, 42]。附录中简要介绍了每种方法。

在社区检测方法的背景下，我们考察了基于社区的方法是否具有一致性，即能否生成定义明确且规模较大的社区。这是我们分析中的重要一步，因为小型社区可能导致性能低下 [37]。在论文引文网络中，大多数社区检测方法发现了 23 至 39 个论文社区，通常每个社区都包含 100 多篇论文。然而，Infomap 生成了 400 多个社区，其中大多数社区包含不到 10 篇论文。在计算每个词的相关性之前，我们决定过滤掉包含论文较少的社区。

E. 短文本关键词提取

这一步包括从预处理后的论文摘要中提取关键词。我们评估了一种基于网络社区的方法，该方法生成词重要性指数，用于对论文摘要中的每个词进行排序。为了进行比较，我们还评估了用于短文本关键词提取任务的词频-逆文档频率和基于 K-Means 的方法。

1. 基于社区的方法：我们利用论文引文网络中发现的社区结构，检测论文摘要中每个词的词重要性指数。该指数衡量某个词在一个社区内出现的相对频率，与其在引文网络其余文档中的频率之间的关系 [37]。为计算某个词 w的词重要性指数 I，我们首先计算该词在社区内的频率

。该数值即相对内部频率$F_{\alpha}^{(\mathrm{in})}$（$w$），其计算公式为

$$F_{\alpha}^{(\mathrm{in})}(w) = \frac{n_{\alpha}(w)}{|\alpha|}, \tag{2}$$

其中，$n_{\alpha}(w)$是包含 w 出现在社区中α的论文总数，$|\alpha|$ 表示与某个社区相关联的论文数 α。我们还计算了 w 在 α之外的相对频率 $F_{\alpha}^{(\mathrm{out})}(w)$，其计算方式如下：

$$F_{\alpha}^{(\mathrm{out})}(w) = \sum_{\gamma \neq \alpha} \frac{n_{\gamma}(w)}{N - |\alpha|}, \tag{3}$$

其中，N 是网络中的论文总数。然后，重要性指数$I(w)$计算为社区内相对频率与社区外相对频率之间的最大差值，即：

$$I(w) = \max_{\alpha} \left[F_{\alpha}^{(\mathrm{in})}(w) - F_{\alpha}^{(\mathrm{out})}(w) \right]. \tag{4}$$

我们计算了论文摘要中所有词的词重要性指数，然后将排名最高的词视为每篇摘要的相关关键词。

2. 基于词频-逆文档频率的方法：计算数据集中所有论文摘要的词频-逆文档频率值。对于每篇摘要，我们考虑其中各个词的词频-逆文档频率权重，并选取词频-逆文档频率值最高的词作为该摘要的相关关键词。

3. 基于 *K-Means* 的方法：此方法等同于基于社区的方法。区别在于，簇是通过K-Means算法 [34]获得的。为了获得簇，我们首先获取每篇摘要的嵌入。然后，我们评估了 K 的多个取值，以找到最优簇数。

F. 长文本关键词提取

从全文中获取的关键词，在评估从短文本中提取的关键词质量时被视为参考关键词。这里，我们将研究论文的完整内容作为输入文本。我们采用了文献中发现的若干方法，来提取文档

中的关键词。这些方法可分为两类：统计方法和基于图的方法：

- 统计和传统关键词提取方法：在这一步中，我们采用了关键词提取任务中常用的统计技术。这些方法会对文档中词语的统计分布进行适当分析。统计方法的主要目标是在不借助任何先验信息或外部信息的情况下，检测并排列文档中的相关词语 [7]。我们采用的方法基于词频、词语词频-逆文档频率、词熵、词语间歇性和 Yake。我们还评估了一种名为 TextRank 的基于图的方法，以及一种基于 BERT 词嵌入的方法。相关方法详见附录。

- 基于网络的方法：我们使用了一套全面的中心性度量，分析从论文全文内容中提取的词共现网络。网络测量有助于识别网络中最相关的节点 [26]。因此，可以根据节点的拓扑重要性对其排序，从而找出每篇文本中最重要的词 [41]。我们依据以下中心性指标，选取每篇文本中排名最高的节点（词）作为关键词：度、PageRank、中介中心性、特征向量中心性、接近中心性，以及在前两个层级计算的可达性 [43]。我们还采用了一种综合各项中心性指标结果的方法。该方法以下简称“投票系统”，即将多数网络测量方法选出的词作为每篇文本的相关关键词。

IV. 结果与讨论

我们的分析分为两节。第四节 A 描述了对数据集的统计分析。第四节 B 比较了从短文本和全文来源中提取的关键词。我们还分析了不同网络社区方法在这项任务中的表现。

A. 数据集分析与参考关键词的选择

本节首先通过分析论文摘要（短文本）与研究论文全文内容（长文本）之间的共有词数量，对数据集进行统计评估。这项分析是为每篇论文摘要生成一组黄金标准关键词的初步步骤。由于许多包含全文论文的数据集缺少人工专家选定的关键词，我们以每篇论文的全文内容（包括除摘要以外的所有章节）作为起点。我们采用关键词提取方法，从每篇论文的完整内容中提取参考关键词。不过，我们首先分析了每篇摘要与全文内容之间共有词的数量。在某些情况下，论文作者可能会在摘要中使用特定词语来表达主要观点，而在论文其余部分则改用其他词语、同义词或相似表达。因此，分析从全文中提取的信息是否与摘要内容相符就变得十分重要。

我们计算了摘要中有多少个词（w）也出现在论文全文中。数据集中这一数量的累积分布（即 $P(x \geq w)$）如图 2 所示。大量研究论文（80%）的摘要与论文全文之间有很多（40 个）共同词。50% 的论文至少有 50 个共同词。正如预期的那样，这意味着摘要中的大部分信息也出现在论文其余部分中。

现在，我们评估在关键词 中，有多少也出现在摘要中，这些关键词是通过全文内容分析 得出的。我们采用两种方法，根据每篇论文的全文提取参考关键词：统计方法和基于图的关键词提取方法。我们通过统计每种方法生成的关键词与构成论文摘要的词之间的交集数量，评估这两种方法。图 3 展示了每种方法得到的结果。根据摘要和全文的长度（见表 I），我们考虑恢复每种关键词提取方法生成的 5 至 50 个关键词。然后，我们统计这些关键词中有多少属于摘要。

与传统方法和统计方法相比，图 3(a) 所示结果表明，Yake、词频和词熵这些方法优于其他关键词提取技术。这些方法能够找到数量最多的常见词，

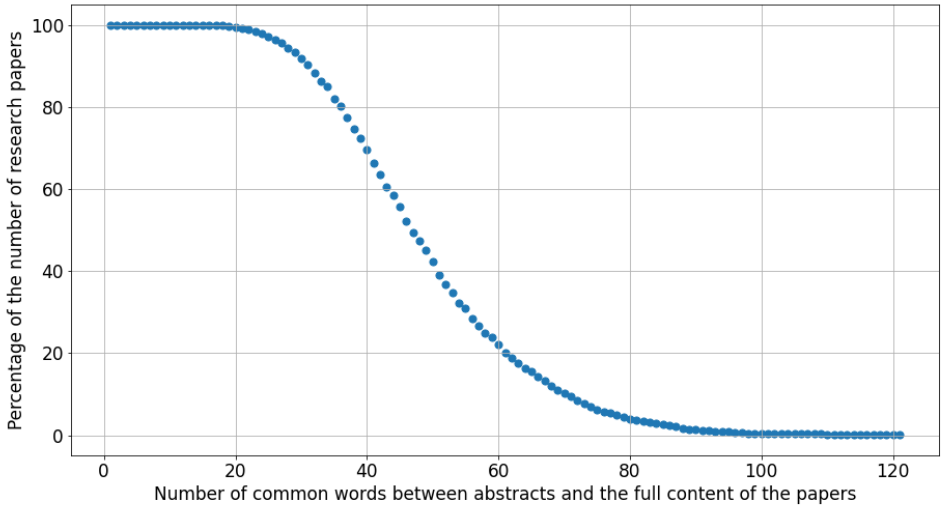
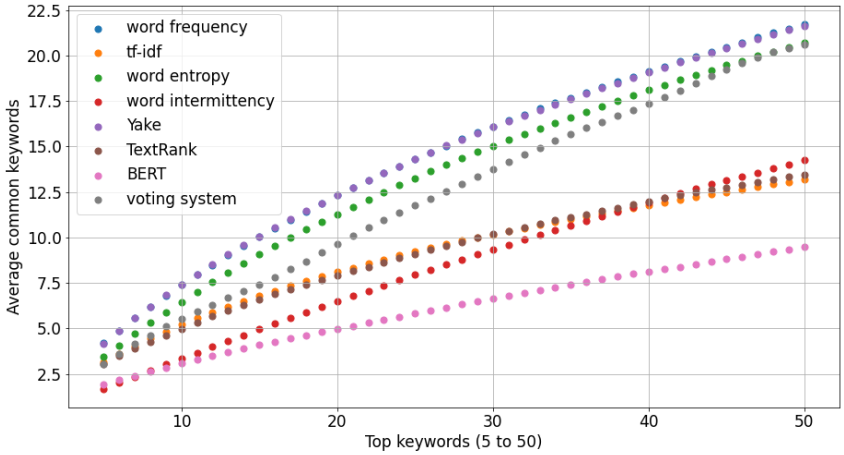


图 2：分析论文摘要与研究论文全文内容之间的共同词数量。横轴表示论文摘要与其对应全文内容之间的共同词数量。纵轴表示$P(x \geq w)$，即摘要与全文之间至少包含 w 个共同词的论文所占比例。

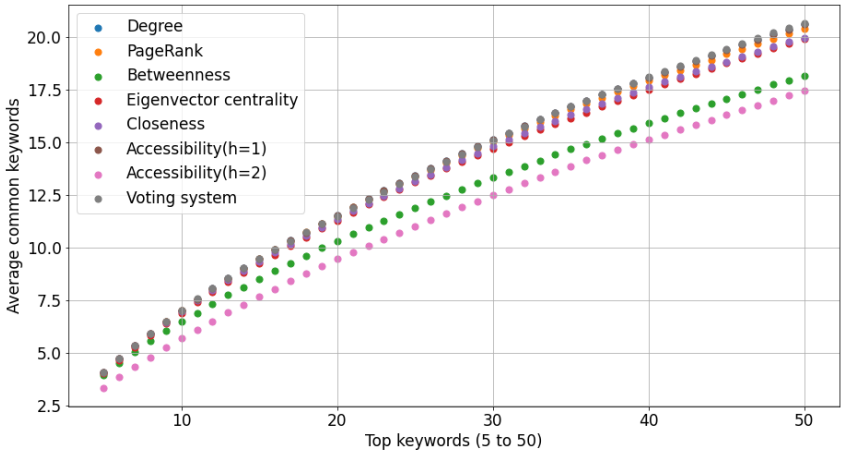
这些词出现在摘要中。投票系统方法并未取得最佳结果。基于词语间歇性和 BERT 的方法与论文摘要共有的词也较少。我们还评估了基于词共现网络和中心性度量的方法。图 3(b) 所示结果表明，几乎所有中心性指标的表现都相近。我们观察到，投票系统、节点度、PageRank 和可达性（在第一个层级计算）优于其他基于网络的方法。不过，与其他网络指标相比，其表现差异并不显著。

B. 从摘要中提取关键词

在本节中，我们分析从摘要中提取关键词的方法能否捕捉到分析全文内容时发现的关键词。我们采用准确率作为性能评估指标。方法的性能以参考关键词与短文本关键词提取方法生成的关键词之间的共有词数占参考关键词总数的比例来衡量。



(a)对研究论文全文内容采用传统和统计关键词提取方法进行分析。



(b)对研究论文全文内容采用基于网络的关键词提取方法进行分析。

图3：分析论文摘要中常见词与采用关键词提取方法从较长文本（即论文全文内容）中识别出的词之间的重叠情况。横轴表示在全文内容中提取出的关键词数量，纵轴表示检索出的关键词中同时出现在论文摘要中的关键词平均数量。

关键词。我们设定参数 N，表示评估中纳入的参考关键词数量。作为参考关键词（即从全文中获得的关键词），我们采用了上一节中观察到的性能最高的方法。

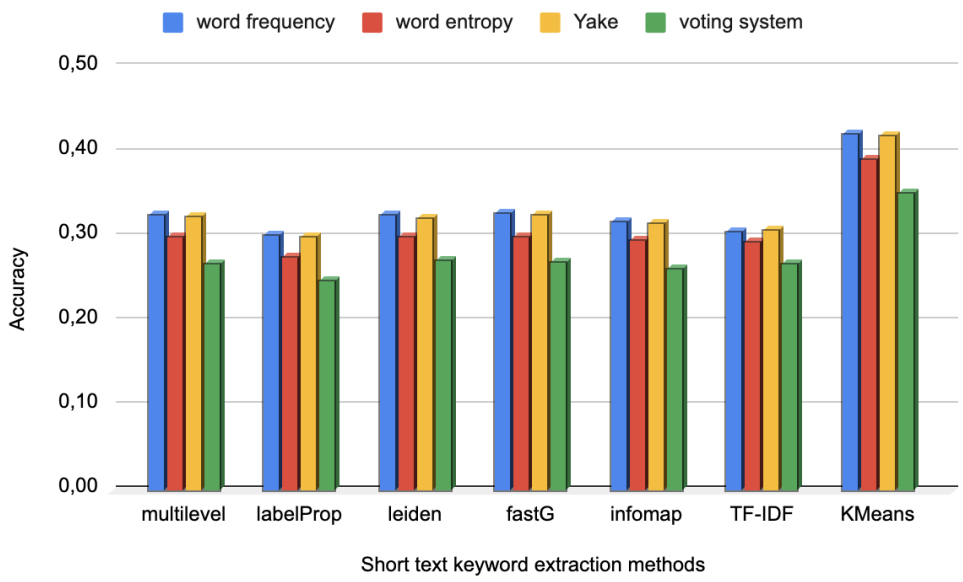
图 4(a) 展示了以

通过统计方法获得的词作为参考关键词进行的性能分析。图 4(a) 中的结果表明，基于社区的方法取得了相近的准确率，因为没有哪种方法明显优于其他方法。标签传播方法的表现略逊于其他方法。我们还发现，TF-IDF 方法的表现与其他基于网络社区的方法相近。令人意外的是，忽略引文信息、仅使用文本信息时，性能有所提升。K-Means 方法明显优于所有其他方法，某些情况下性能提升了 25%。附录中展示了针对参数 N 不同取值的完整分析。

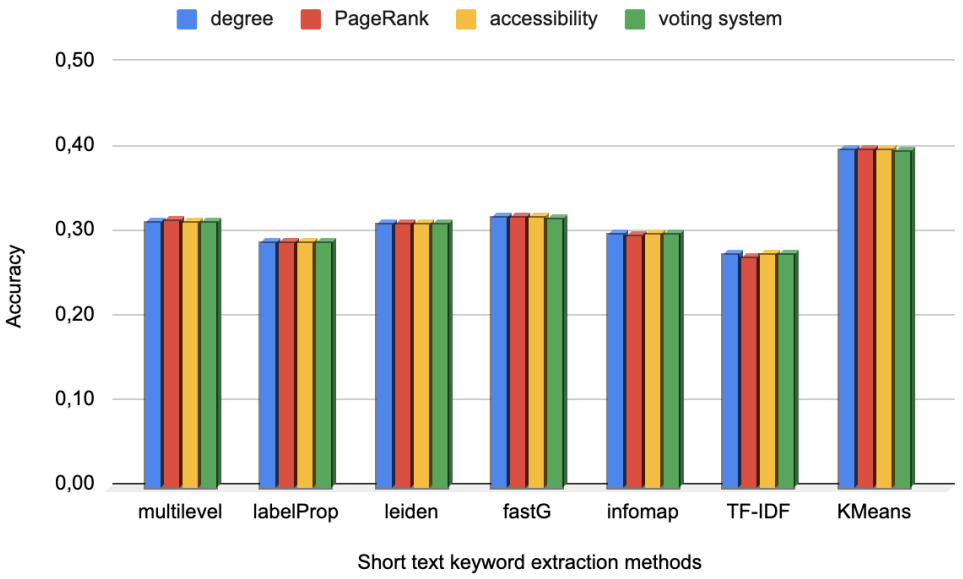
图 4(b) 展示了以基于（共现）网络的方法得到的关键词作为参考关键词时的性能分析。我们可以看到，所有纳入考察的网络中心性指标所呈现的性能几乎相同。这一结果与文本网络分析领域的文献一致，因为分析书面文本时，各中心性指标之间存在相关性。关于通过社区检测方法生成参考关键词，还值得注意的是，所有基于社区的方法在这项任务中的表现都相似。相反，通过词频-逆文档频率选择参考关键词时，性能最差。再一次，KMeans 方法取得了最佳表现。

社区检测方法表现相似的一种可能解释是，所有方法生成的划分相似，因此为每篇论文摘要选取了相同的一组关键词。为了检验这一假设，我们计算了不同社区检测方法生成的词语相关性排名之间的斯皮尔曼相关系数。结果（未显示）表明，这些方法实际上选取了不同的关键词集合。如果考虑词语的完整排名，斯皮尔曼相关系数通常低于 0.10。类似地，仅考虑排名前 30 的词语时，所有相关系数都低于 0.27。正如预期，在这两种情况下，Multilevel 和 Leiden 方法生成的排名之间的相关性最高 [4, 42]。

性能结果揭示了一些有趣的发现。首先，我们发现，仅凭引文信息识别关键词，无法使摘要和全文中找到的关键词达到最高匹配度。尽管引文已被用于许多场景中 [2], ，但观察到的低性能可能有一个原因：引文可能无法



(a) 将传统方法和统计方法作为参考关键词的性能分析。



(b) 将基于网络的方法作为参考关键词的准确性分析。

图 4：对短文本（论文摘要）的关键词提取方法评估所得准确率进行比较分析。我们从全文中提取了 $N = 30$ 关键词。

反映文本的语义相似度，从而影响社区检测方法的性能。事实上，一些研究指出，引文与内容相似性之间存在差异。例如， [2] 发现，引文与内容相似性并不一致，因为选择参考文献时，往往会忽略最相似的论文

对于论文。类似地，内容差异也被用于改进重现引文网络增长过程的模型，因为内容相似性也被用作模拟引文网络增长的重要特征 [46]。

虽然使用文本信息能够更好地从全文内容中提取关键词，但所得准确率仍低于 50%。这意味着，仅凭论文摘要中的聚类信息不足以还原论文的全文内容。这可能会影响许多基于关键词还原文本内容的研究。例如，在研究引文网络的性质时，通过关键词筛选论文可能会影响引文网络指标的稳定性。如果关键词术语的定义不够明确，用来筛选相关论文，那么在一个更大领域中可能会识别出数量不同、代表各子领域的社群。

从摘要和全文中提取的内容差异，也可能导致在科学学领域中产生不同的解读。例如，在文档相似性网络中，一篇论文的中心性可能很大程度上取决于使用的是摘要还是全文。如果在其他情境下研究此类网络，可能会得出不够稳健的结论。例如，比较由引文关联的论文之间的语义相似度时，结果可能因用于衡量语义相似度的文本量不同而有所差异 [2]。因此，若要依据对论文语义相似度的分析得出结论，考虑全文内容仍然很重要。

五、结论

从短文本中识别关键词是一项重大挑战。本文评估了广为人知的方法能否从摘要中提取出与基于全文内容的分析结果相符的关键词。由于摘要提供的语境有限，我们采用了利用引文语境将论文聚类为语义相似组的方法。此外，我们还使用了基于统计和 K-Means 算法的策略。我们通过多种技术从论文全文内容中获取参考关键词。

研究结果表明，K-Means算法这类简单方法的表现优于依赖从引文网络中提取的社群的方法。此外，结果显示，各种应用于

至引文网络 网络的社区检测方法表现相似，没有任何一种特定方法明显占优 .

总而言之，引文网络和不依赖引文的替代方法表现均不理想。这一结果表明，从摘要中提取的关键词与通过全面内容分析得到的关键词并不一致。因此，有必要开展进一步研究，以探讨观察到的差异是否会导致文档相似性网络分析结果出现偏差 [25]。

提升关键词提取聚类方法性能的一种潜在途径，是采用其他文本向量表示方法。融入文本嵌入（例如由 BERT 模型生成的嵌入） [3], 可能有助于有效表示文档。此外，在生成参考关键词时进行同义词处理也可能有所裨益。创建论文网络时，整合引文信息与文本信息也有望提升性能。

致谢

本研究部分经费由巴西高等教育人员培训协调委员会（CAPES）资助，经费代码 001。Diego R. Amancio 感谢巴西国家科学技术发展委员会（资助编号304026/2018-2、311074/2021-9）和圣保罗州研究基金会（资助编号20/06271-0）提供的经费支持。Thiago C. Silva（资助编号308171/2019-5、408546/2018-2）衷心感谢巴西国家科学技术发展委员会提供的经费支持。

附录

A. 社区检测方法

本节简要介绍本文采用的网络社区方法：

1. 多层次：在该算法中，每个节点最初都被分配到不同的社区。随后，将节点移至其对应邻居所在的社区中，并选择能够最大幅度提升模块度 [4]的社区。重复这一过程，直至节点对模块度的局部贡献不再增加。原始网络中的每个社区随后

都会被缩减为单个节点（同时保留相邻边的总权重），然后该方法继续进入下一层。当社区缩减为节点后，模块度得分无法再提高时，算法便会结束。

2. 标签传播： [33] 中提出的方法基于网络中的邻域连通性和信息扩散原理。该方法首先为网络中的每个节点分配唯一的社区标签，随后将这些标签传播到整个网络中。在每次迭代中，每个节点都会采用其直接邻域中最常见的标签。然后随机移除网络中的边，并在下一次迭代开始前以随机顺序更新节点。当节点达到共识时，算法便会停止。这里的共识是指每个节点都持有其邻居节点中的多数标签。

3. *Leiden*算法： [42], 中提出的Leiden算法，是对广泛使用的Multilevel 方法 [4]的改进。众所周知，后者存在一个缺点：经常会发现连接较弱的社群。相比之下，Leiden 方法旨在通过实施以下三个阶段确保社群连接紧密：(i) 节点局部移动（与 Multilevel 方法相同）；(ii) 划分细化；以及 (iii) 网络聚合。通过纳入这三个阶段，与Multilevel 方法相比，Leiden算法能够在显著缩短的时间内发现质量更高的簇。

4. 快速贪心： 该算法基于层次凝聚聚类，旨在优化模块度得分 [10]。该方法首先考虑一个仅由高度连接节点之间的边构成的子网络。随后，这种方法会评估随机选取的边，找出能提高子网络模块度的边，并将其聚合。重复此过程，直至模块度的增量改善变得微不足道。最后，通过识别子网络中的连通分量来获得各个社群。

5. *Infomap*： 该算法由 [35]提出，基于信息论。这种方法首先将网络编码为模块，以最大化

从原始网络中保留的信息量。随后，编码网络通过容量有限的信道传输。解码器的目标是尝试解码消息，并构建一组原始网络的可能候选项。候选项越少，传输的原始网络信息就越多。该算法还利用随机游走分析信息在网络中的流动。

B. 统计关键词检测

• 词频和 *TF-IDF* 方法（频率和 *TF-IDF*）：关键词提取最简单的技术之一是基于频率的方法，它将高频出现的词视为相关词。因此，按词频排名最高的词被视为关键词。为缓解基于频率的方法的局限性，我们还评估了 TF-IDF 方法。与基于频率的方法不同，TF-IDF 方法根据每个词在文档中以及整个数据集中的出现次数，为其词频赋予权重。在这种方法中，TF-IDF 值最高的词被视为关键词。

• 词熵（*W.E.*）：该方法利用香农熵分析给定文本中每个词的出现序列所包含的信息量 [7]。此技术需要将文本划分为 N 个片段，以计算每个词的熵。在本研究中，我们根据构成每篇文本的句子数量对论文文本进行划分。根据该方法，一个词的熵值越高，该词在文本中的分布越不均匀，因此其相关性也越高。该方法的一项主要优势是，无需使用大型文本语料库进行训练，只需输入文本即可。

• 词语间歇性（*W.I.*）：该指标考虑了词语的重要性与其空间分布之间的关系 [1]。先前研究发现，重要词语与文本的主题密切相关，且呈现出高度异质的分布。这类词语往往位于特定区域

的文本中，表现出较大的词频波动，并且通常会形成簇 [7]。相比之下，停用词等常见词语会随机分布在文档中，分布相对均匀。因此，正如 [7], 所提出的，可以通过对词语出现次数的分布进行统计分析，从给定文本中识别相关关键词。与词频和熵方法类似，该技术仅依据目标文本识别重要词语，无需外部信息。

- $Yake$：Yake方法从源文本中提取统计特征，以识别最相关的关键词 [5]。每个词项都会计算五项特征：(i) 大小写；(ii) 词语位置，为出现在文本开头的词赋予更高的重要性；(iii) 词频，为出现频率更高的词赋予更高的相关性；(iv) 词语与上下文的相关性，用于衡量目标词左右两侧出现的不同词项数量；以及 (v) 词 $difSentence$，用于衡量一个词在不同句子中出现的频率。随后，这些特征会合并为一个单一指标，为每个词赋予重要性权重。根据该得分，数值较低的词被视为相关关键词 [6]。

- *TextRank*（*TextR*）：TextRank方法由 [24], 提出，是一种基于图的方法，采用PageRank算法，广泛用于文本摘要和关键词提取任务。在这种方法中，文本被建模为词共现网络，节点由词表示；如果两个节点在窗口大小范围内共现，就会在它们之间建立边。原论文中，窗口大小设定为 2 到 10 个词。该方法采用PageRank算法为每个词排序，并选取排名靠前的词作为相关关键词。

- 基于*BERT*的方法：基于变换器的双向编码器表示（BERT）技术是一种先进的嵌入模型，通过稠密向量表示捕捉文档的语义内容 [14]。基于BERT的方法为文本中的每个n元语法生成词嵌入。随后，应用余弦相似度度量，找出与

原始文档最相似的词。排名靠前的相似词随后被视为每篇文档的相关关键词。

- 投票系统（*V.S.*）：为了提高长文本关键词提取器的准确率，我们综合了上述方法得到的结果。我们采用了一种基于多数关键词提取方法所找到的关键词的投票系统。

C. 基于准确性分析的完整结果

表 II：关键词提取方法评估所得的准确率

短文本（论文摘要）。此处，我们将相关词语视为参考关键词

这些词语是通过传统和统计方法从论文全文内容中提取的。 N 表示我们为短文本和长文本关键词提取方法找回的排名靠前的关键词数量

提取方法。

	词频				词频-逆文档频率			
方法	**N=10**	**N=20**	**N=30**	**N=40**	**N=10**	**N=20**	**N=30**	**N=40**
多层次	0.1578	0.2559	0.3269	0.3678	0.0740	0.1066	0.1448	0.1793
标签传播	0.1375	0.2312	0.3022	0.3504	0.0787	0.1033	0.1363	0.1724
Leiden算法	0.1496	0.2533	0.3260	0.3664	0.0819	0.1153	0.1463	0.1785
fastG	0.1465	0.2518	0.3290	0.3684	0.0713	0.1023	0.1419	0.1767
Infomap	0.1375	0.2320	0.3175	0.3647	0.1458	0.1561	0.1726	0.1951
词频-逆文档频率	0.2858	0.2889	0.3071	0.3350	0.3596	0.3223	0.2984	0.2755
K均值	0.4253	0.4213	0.4220	0.4146	0.2129	0.2106	0.2198	0.2279
	词熵				**词语间歇性**			
方法	**N=10**	**N=20**	**N=30**	**N=40**	**N=10**	**N=20**	**N=30**	**N=40**
多层次	0.1341	0.2294	0.3004	0.3460	0.0679	0.1291	0.1828	0.2233
标签传播	0.1165	0.2052	0.2769	0.3285	0.0608	0.1159	0.1663	0.2110
Leiden算法	0.1279	0.2293	0.3007	0.3455	0.0651	0.1288	0.1829	0.2228
fastG	0.1237	0.2232	0.3012	0.3456	0.0658	0.1241	0.1807	0.2218
Infomap	0.1259	0.2166	0.2967	0.3441	0.0715	0.1274	0.1811	0.2225
词频-逆文档频率	0.2477	0.2708	0.2945	0.3217	0.1093	0.1539	0.1868	0.2180
K均值	0.3596	0.3835	0.3929	0.3925	0.1335	0.1921	0.2280	0.2499
	Yake				**TextRank**			
方法	**N=10**	**N=20**	**N=30**	**N=40**	**N=10**	**N=20**	**N=30**	**N=40**
多层次	0.1553	0.2536	0.3242	0.3650	0.1138	0.1689	0.2101	0.2326
标签传播	0.1362	0.2290	0.2997	0.3481	0.0967	0.1594	0.2012	0.2260
Leiden算法	0.1509	0.2516	0.3226	0.3633	0.0986	0.1673	0.2090	0.2319
fastG	0.1448	0.2485	0.3256	0.3653	0.1172	0.1818	0.2184	0.2356
Infomap	0.1400	0.2309	0.3153	0.3616	0.0829	0.1503	0.2034	0.2302
词频-逆文档频率	0.2905	0.2937	0.3091	0.3353	0.1381	0.1492	0.1703	0.1951
K均值	0.4221	0.4190	0.4194	0.4121	0.2132	0.2337	0.2459	0.2479
	BERT				**投票系统**			
方法	**N=10**	**N=20**	**N=30**	**N=40**	**N=10**	**N=20**	**N=30**	**N=40**
多层次	0.0573	0.0924	0.1254	0.1501	0.1231	0.1925	0.2688	0.3283
标签传播	0.0710	0.0936	0.1221	0.1462	0.1108	0.1758	0.2485	0.3123
Leiden算法	0.0757	0.1092	0.1325	0.1518	0.1325	0.2003	0.2716	0.3274
fastG	0.0739	0.0965	0.1257	0.1494	0.1300	0.1941	0.2702	0.3280
Infomap	0.0596	0.0938	0.1261	0.1483	0.1115	0.1826	0.2633	0.3251
词频-逆文档频率	0.0899	0.1043	0.1207	0.1394	0.1891	0.2280	0.2687	0.3070
K均值	0.1804	0.1668	0.1688	0.1726	0.2761	0.3110	0.3531	0.3746

表 III：短文本（论文摘要）关键词提取方法评估所得的准确率。此处，我们将基于网络的方法在论文全文内容中找到的最重要词语作为参考关键词。 N表示我们为短文本和长文本关键词提取方法各自找回的排名靠前关键词数量。

	度				PageRank			
方法	**N=10**	**N=20**	**N=30**	**N=40**	**N=10**	**N=20**	**N=30**	**N=40**
多层次	0.1605	0.2490	0.3149	0.3539	0.1622	0.2498	0.3153	0.3514
标签传播	0.1324	0.2234	0.2903	0.3374	0.1329	0.2247	0.2899	0.3349
Leiden算法	0.1417	0.2421	0.3125	0.3522	0.1428	0.2429	0.3120	0.3494
fastG	0.1466	0.2496	0.3195	0.3549	0.1481	0.2510	0.3194	0.3523
Infomap	0.1256	0.2152	0.3002	0.3476	0.1247	0.2146	0.2979	0.3445
词频-逆文档频率	0.2529	0.2546	0.2757	0.3085	0.2516	0.2521	0.2725	0.3043
K均值	0.4047	0.3989	0.4010	0.3948	0.4070	0.3998	0.3999	0.3920
	中介中心性				**特征向量中心性**			
方法	**N=10**	**N=20**	**N=30**	**N=40**	**N=10**	**N=20**	**N=30**	**N=40**
多层次	0.1563	0.2281	0.2809	0.3127	0.1484	0.2385	0.3008	0.3385
标签传播	0.1255	0.2021	0.2571	0.2973	0.1253	0.2136	0.2787	0.3238
Leiden算法	0.1356	0.2202	0.2774	0.3099	0.1343	0.2315	0.2994	0.3370
fastG	0.1392	0.2272	0.2836	0.3134	0.1387	0.2373	0.3031	0.3393
Infomap	0.1125	0.1855	0.2596	0.3028	0.1239	0.2120	0.2921	0.3346
词频-逆文档频率	0.2291	0.2255	0.2387	0.2661	0.2490	0.2522	0.2726	0.3021
K均值	0.3828	0.3656	0.3574	0.3503	0.3865	0.3844	0.3847	0.3797
	接近中心性				**可达性 (h=1)**			
方法	**N=10**	**N=20**	**N=30**	**N=40**	**N=10**	**N=20**	**N=30**	**N=40**
多层次	0.1585	0.2438	0.3066	0.3426	0.1605	0.2490	0.3149	0.3539
标签传播	0.1317	0.2198	0.2855	0.3278	0.1324	0.2234	0.2903	0.3374
Leiden算法	0.1390	0.2373	0.3052	0.3411	0.1417	0.2421	0.3125	0.3522
fastG	0.1461	0.2446	0.3096	0.3431	0.1466	0.2496	0.3195	0.3549
Infomap	0.1209	0.2127	0.2932	0.3367	0.1256	0.2152	0.3002	0.3476
词频-逆文档频率	0.2478	0.2499	0.2717	0.3019	0.2529	0.2546	0.2757	0.3085
K均值	0.3936	0.3883	0.3891	0.3828	0.4047	0.3989	0.4010	0.3948
	可达性(h=2)				**投票系统**			
方法	**N=10**	**N=20**	**N=30**	**N=40**	**N=10**	**N=20**	**N=30**	**N=40**
多层次	0.1196	0.1943	0.2526	0.2904	0.1605	0.2481	0.3141	0.3535
标签传播	0.1048	0.1744	0.2338	0.2777	0.1324	0.2226	0.2909	0.3377
Leiden算法	0.1088	0.1869	0.2508	0.2894	0.1426	0.2408	0.3119	0.3522
fastG	0.1108	0.1909	0.2519	0.2899	0.1464	0.2487	0.3175	0.3548
Infomap	0.0997	0.1740	0.2419	0.2855	0.1255	0.2147	0.3001	0.3479
词频-逆文档频率	0.1977	0.2125	0.2344	0.2638	0.2536	0.2546	0.2759	0.3093
K均值	0.3162	0.3167	0.3236	0.3264	0.4039	0.3963	0.3985	0.3952

[1] D. R. Amancio. 探究用于建模短文本的复杂网络的拓扑属性。*PLOS ONE*, 10(2):e0118394, 2015.[2] D. R. Amancio, M. d. G. V. Nunes, O. N. Oliveira Jr, and L. da F. Costa. 运用复杂网络概念评估科学论文中的引文方法。*Scientometrics*, 91(3):827–842, 2012.[3] I. Beltagy, K. Lo, and A. Cohan. SciBERT：一种用于科学文本的预训练语言模型。*arXiv* 预印本 *arXiv:1903.10676*, 2019.[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. 大型网络中社区的快速展开。*Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008年10月。[5] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt. Yake!：使用多个局部特征从单篇文档中提取关键词。*Information Sciences*, 509:257–289, 2020.[6] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, and A. Jatowt. Yake!：与语料集无关的自动关键词提取器。载于欧洲信息检索会议，第806–810页。施普林格，2018年。[7] C. Carretero-Campos, P. Bernaola-Galv´an, A. Coronado, and P. Carpena. 改进短文本中的统计关键词检测：熵方法与聚类方法。*Physica A:Statistical Mechanics and its Applications*, 392(6):1481–1492, 2013.[8] N. Castro and M. Stella. 心理词典的多重网络结构会影响失语症患者的图片命名。*Journal of Complex Networks*, 7(6):913–931, 2019.[9] J. Chen, H. Hou, and J. Gao. 基于图的中文短文本关键词提取中的内部重要性因素。*ACM* 亚洲与低资源语言信息处理汇刊 *(TALLIP)*, 19(5):1–15, 2020.[10] A. Clauset, M. E. Newman, and C. Moore. 在超大型网络中发现社区结构。*Physical review E*, 70(6):066111, 2004.[11] E. A. Corrˆea Jr and D. R. Amancio. 使用词嵌入和复杂网络中的社区检测进行词义归纳。*Physica A: Statistical Mechanics and its Applications*, 523:180–190, 2019.

[12] N. Dakiche、F. B.-S. Tayeb、Y. Slimani 和 K. Benatchba。社交网络中的社群演化追踪：综述。*Information Processing & Management*，56(3):1084–1102，2019。[13] H. F. de Arruda、L. d. F. Costa 和 D. R. Amancio。通过复杂网络中的社区检测进行主题分割。*Chaos: An Interdisciplinary Journal of Nonlinear Science*，26(6):063120，2016。[14] J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova。BERT：用于语言理解的深度双向 Transformer 预训练。*arXiv*预印本 *arXiv:1810.04805*，2018。[15] H. Ferraz de Arruda、F. Nascimento Silva、V. Queiroz Marinho、D. Raphael Amancio 和 L. da Fontoura Costa。将文本表示为复杂网络：一种介观方法。*Journal of Complex Networks*，6(1):125–144，2018。[16] M. Grineva、M. Grinev 和 D. Lizorkin。从含噪声的多主题文档中提取关键词。在第*18*届万维网国际会议论文集中，第661–670页，2009。[17] A.-W. Harzing。学术出版领域的两位新秀：Crossref 和 Dimensions 与 Google Scholar、Microsoft Academic、Scopus 及 Web of Science 相比如何? *Scientometrics*，120(1):341–349，2019。[18] X. Jiang、Y. Hu 和 H. Li。一种关键词短语提取的排序方法。在第*32*届国际 *ACM SIGIR* 信息检索研究与开发会议论文集中，第756–757页，2009。[19] S. M. Joseph、S. Citraro、V. Morini、G. Rossetti 和 M. Stella。认知网络邻域量化了自杀遗书和 Reddit 心理健康社群中表达的情感。*Physica A: Statistical Mechanics and its Applications*，第128336页，2022。[20] S. Lahiri、S. ray choudhury 和 C. Caragea。利用搭配网络上的中心性度量进行关键词和关键词短语提取。2014年1月。[21] J. Li、G. Huang、C. Fan、Z. Sun 和 H. Zhu。通过 Word2Vec、Doc2Vec 和 TextRank 为短文本提取关键词。*Turkish Journal of Electrical Engineering and Computer Sciences*，27(3):1794–1805，2019。[22] W. Li 和 J. Zhao。利用维基百科提取短文本关键词的 TextRank 算法。在*2016*年第*3*届国际信息科学与控制工程会议（*ICISCE*）中，第683–686页。IEEE，2016。[23] J. Machicao、E. A. Corrˆea Jr、G. H. Miranda、D. R. Amancio 和 O. M. Bruno。作者身份

归属的判定基于类生命网络自动机。*PLOS ONE*，13(3):e0193703，2018。[24] R. Mihalcea 和 P. Tarau。TextRank：为文本理清脉络。在*2004*年自然语言处理实证方法会议论文集中，第404–411页，2004。[25] S. Minaee、N. Kalchbrenner、E. Cambria、N. Nikzad、M. Chenaghlu 和 J. Gao。基于深度学习的文本分类：全面综述。*ACM*计算调查（*CSUR*），54(3):1–40，2021。[26] M. Newman。网络。牛津大学出版社，2018。[27] https://clarivate.com/webofsciencegroup/solutions/web-of-science/.[28] https://arxiv.org/.[29] M. Ortu˜no、P. Carpena、P. Bernaola-Galv´an、E. Munoz 和 A. M. Somoza。自然语言和 DNA 中的关键词检测。*EPL*（欧洲物理学快报），57(5):759，2002。[30] R. Pramana、J. J. Subroto、A. A. S. Gunawan 等。句子相似度中词干提取和词形还原性能的系统性文献综述。在*2022*年第七届*IEEE*信息技术与数字应用国际会议（*ICITDA*）论文集中，第1–6页。IEEE，2022。[31] L. V. Quispe、J. A. Tohalino 和 D. R. Amancio。利用虚拟边提高共现文本网络的区分能力。*Physica A: Statistical Mechanics and its Applications*，562:125344，2021。[32] F. Radicchi、C. Castellano、F. Cecconi、V. Loreto 和 D. Parisi。网络中社区的定义与识别。美国国家科学院院刊，101(9):2658–2663，2004。[33] U. N. Raghavan、R. Albert 和 S. Kumara。用于检测大规模网络中社群结构的近线性时间算法。《物理评论*E*》，76:036106，2007年9月。[34] M. Z. Rodriguez、C. H. Comin、D. Casanova、O. M. Bruno、D. R. Amancio、L. d. F. Costa 和 F. A. Rodrigues。聚类算法：比较研究。*PLOS ONE*，14(1):e0210236，2019。[35] M. Rosvall 和 C. T. Bergstrom。复杂网络中的随机游走图揭示社群结构。美国国家科学院院刊，105(4):1118–1123，2008。[36] G. Salton 和 C.-S. Yang。自动索引中的术语值规范。《文献学杂志》，1973。[37] F. N. Silva、D. R. Amancio、M. Bardosova、L. d. F. Costa 和 O. N. Oliveira Jr。利用网络

科学与文本分析，以生成特定科学主题的综述。《信息计量学杂志》，10(2):487–502，2016。[38] M. Stella。多重网络可量化心理词典对灾难性概念失效、失语症退化和老化的稳健性。*Physica A: Statistical Mechanics and its Applications*，554:124382，2020。[39] M. Timonen、T. Toivanen、Y. Teng、C. Chen 和 L. He。基于信息量的短文档关键词提取。载于 *KDIR*，第 411–421 页，2012。[40] J. A. Tohalino、T. C. Silva 和 D. R. Amancio。利用虚拟边从建模为复杂网络的文本中提取关键词。*arXiv*预印本 *arXiv:2205.02172*，2022。[41] J. V. Tohalino 和 D. R. Amancio。使用多层网络进行抽取式多文档摘要。*Physica A: Statistical Mechanics and its Applications*，503:526–539，2018。[42] V. A. Traag、L. Waltman 和 N. J. Van Eck。从 Louvain 到 Leiden：保证社区紧密连接。《科学报告》，9(1):1–12，2019。[43] B. A. N. Traven،colo 和 L. d. F. Costa。复杂网络中的可达性。《物理学快报*A*》，373(1):89–95，2008。[44] X. Wan 和 J. Xiao。CollabRank：迈向单文档关键短语提取的协作方法。载于第*22*届国际计算语言学会议论文集（*Coling 2008*），第 969–976 页，2008。[45] X.-G. Wang。利用社区结构的网络分类方法。*Applied Math-ematics & Information Sciences*，9(3):1181，2015。[46] Q. Zhao 和 X. Feng。利用引文网络结构预测论文被引次数：一种深度学习方法。《信息计量学杂志》，16(1):101235，2022。