arXiv:2301.00565v1 [cs.CL] 2 Jan 2023

使用语义而非词语进行主题跟踪

Judicael POUMAY[1] 和 Ashwin ITTOO[1]

列日大学/列日高等商学院，卢夫雷街 14 号，4000 列日，比利时 {judicael.poumay, ashwin.ittoo}@uliege.be

摘要。 能够监测主题随时间的演变对企业而言极具价值。目前，所有现有的主题跟踪方法都通过匹配词语的使用来利用词汇信息。然而，迄今没有研究尝试使用语义信息进行主题跟踪。因此，我们探索了一种使用词嵌入的新颖基于语义的方法。结果表明，基于语义的主题跟踪方法与基于词汇的方法不相上下，但会犯下不同的错误。这表明，两种方法可能互为补充。

关键词： 主题跟踪 · 词汇 · 语义 · 主题模型

1 引言

在网络上大量文本中蕴藏着有意义的见解，这些见解有助于支持商业决策活动。主题建模方法能够从语料库中提取潜在主题 [4,10] ，并可用于发现这些见解。应用示例包括欺诈检测 [11],以及了解员工和客户满意度 [8,7]。通过跟踪提取出的主题随时间的变化，可以了解其演变过程或发现新兴主题。因此，我们专注于主题跟踪这一任务，其目标是关联在不同时间段提取出的同一主题实例。

过去已提出多种主题跟踪方法 [3,6,13,12,9]。这些方法采用 JS 散度 [13,12,9] 或在线主题模型 [3,6]等度量，它们依赖词汇信息来跟踪主题随时间的变化。

然而，迄今尚无研究尝试利用语义信息来跟踪主题随时间的变化。直观而言，基于语义的方法可能颇有前景，因为它们不依赖简单的表层形式，还能捕捉同义关系等概念。例如，对于一个关于“人工智能”的主题，随着时间推移，我们可能会观察到“机器学习”一词变得比“人工智能”更流行。然而，基于词汇的方法无法处理这种词汇漂移，也无法关联这些词语随时间的变化。相反，语义方法则能够捕捉这种词汇变化。此外，多次运行得到的主题-词分布并不稳定 [1], ，也就是说，一个主题最终生成的高频词往往会显著变化。这意味着，我们用于跟踪主题的词汇信息也不稳定，即使主题的整体语义保持不变。因此，基于语义的方法可能更为稳健。

因此，我们旨在研究如何利用语义信息进行主题跟踪，并将其与词汇信息进行比较。为此，我们的主要贡献是提出一种基于词嵌入的新型语义主题跟踪方法，称为语义散度（SD）。此外，我们还研究了层次主题建模背景下主题跟踪所面临的挑战。

2 背景与相关工作

2.1 主题建模

潜在狄利克雷分配 [4] 是首个传统主题模型。潜在狄利克雷分配的核心是一个贝叶斯生成模型，其中包含两个狄利克雷分布，分别对应文档-主题分布和主题-词分布。通过推断过程学习并优化这些分布，从而提取主题。潜在狄利克雷分配的主要弱点是要求用户预先指定要提取的主题数量。

自潜在狄利克雷分配以来，人们提出了更复杂的主题模型。尤其是 HTMOT [10]，它被提出用于同时建模主题层级和时间性。具体而言，HTMOT 会生成一棵主题树，其深度以及每个分支的子主题数量会在训练过程中动态确定。此外，HTMOT 还对主题的时间性进行建模，从而能够提取词汇相近但时间上不同的主题。

2.2 主题跟踪

主题跟踪是监测主题随时间演变的任务。它最初于 1998 年的一项试点研究中被定义为：将新闻报道流持续自动分类为已知或新主题。 [2]

目前，有两种通用框架在主题跟踪领域展开竞争。第一种是在线主题模型，它会逐步纳入新数据。 [3,6]在[3], 中，作者提出了在线潜在狄利克雷分配（Online-LDA），这是潜在狄利克雷分配（LDA）的一种变体，能够利用新文档更新模型，而无需访问先前处理过的文档。实际上，在线潜在狄利克雷分配假设时间被划分为若干片段，并在每个片段使用前一片段作为先验来训练一个 LDA 模型。作者通过向新闻流中人为注入新主题，证明了他们的系统能够发现新兴主题。他们在 NIPS 和 Reuters-21578 数据集上进行了实验。同样，在 [6], 中，作者提出了一种能够以在线方式动态确定适当主题数量的模型。他们在 20 Newsgroup 和 TDT-2 数据集上进行了实验。

第二条研究脉络关注如何关联在不同时间段独立提取的主题 [13,12,9]。在 [13], 中，作者使用了 2000 年至 2015 年间发表于各类期刊的约 30,000 篇论文摘要。随后，他们对每一年的数据分别应用潜在狄利克雷分配，并使用 Jensen-Shannon 散度（JS）衡量主题之间的相似度，从而将主题关联起来 [5]。在 [12]中，作者在

新闻文章上应用了类似的方法。不过，两者有所不同： [13] 仅将主题关联起来，而 [12] 则对主题进行聚类。这意味着，两个主题一旦关联，便会形成一个簇，之后的主题将与整个簇进行比较，而不只是与前一个主题进行比较。最后，在 [9], 中，作者还提出了一种将 JS 散度应用于科学论文的跟踪方法。不过，他们并未将关联限制为一对一映射，因此主题可以融合或拆分。上述所有论文都通过定性分析评估了其主题跟踪方法，并证明了相关技术的性能。

我们基于第二种流程，因为它能够更好地并行处理：各个时间切片可以独立处理。

3 方法

本节将介绍我们用于主题跟踪的方法。首先，我们将描述语料库和主题提取方法。接下来，我们将定义 SD 度量。最后，我们将介绍主题跟踪算法。

3.1 主题提取

为了开展实验，我们从 Digital Trends 的 [1]档案中抓取了 1 万篇 2019年到 2020年 的文章。该新闻网站主要关注科技新闻，涉及硬件、太空探索和 COVID-19 等主题。对于所有文章，我们提取了正文、标题、类别和时间戳。我们按照 HTMOT [10] 对语料库进行了预处理。

为提取主题层级（见图 1），我们使用了 HTMOT 主题模型[10] 。提取出的主题由一个词语列表和一个实体列表表示。

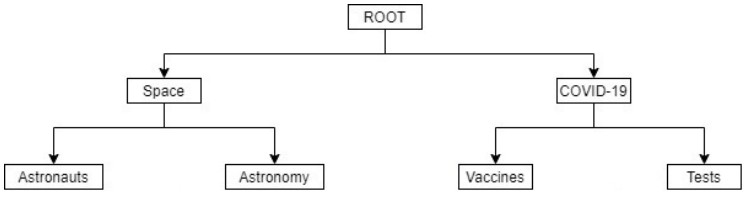


图 1. 主题层级示例

我们遵循 HTMOT [10] ，只关注提取出的主题的第一层和第二层。具体而言，作者观察到，层级越深，主题就越晦涩，越难以让代表普通受众的标注人员理解。因此，要评估主题树深层主题跟踪的正确性也更加困难。

[1] https://www.digitaltrends.com/

3.2 提出的语义散度度量

现在，我们将介绍一种新颖的主题跟踪方法，它不同于以往研究中传统采用的 JS 散度。我们将该方法称为“语义散度”，简称 SD。它利用词嵌入来衡量主题之间的距离。每个主题都将被赋予一个嵌入向量，该向量是该主题中排名靠前的词的嵌入向量按其概率加权求和的结果。然后，通过计算两个主题各自嵌入向量的余弦距离，得出它们之间的距离。我们将使用 FastText 作为词嵌入模型。FastText 有助于处理罕见词和词汇表外词汇。这一点至关重要，因为我们的预处理步骤包括词形还原，可能会产生拼写错误的词。因此，主题的嵌入定义如下：

$$emb(t) = \sum_{(w,p)\in t} p * FastText(w) \tag{1}$$

两个主题之间的语义散度定义如下：

$$SD(t_1, t_2) = cosine(emb(t_1), emb(t_2)) \tag{2}$$

其中， w 是主题 t 中的一个词， p 是该词的概率。

3.3 主题跟踪算法

最后，为了跟踪主题随时间的变化，我们在语料库上应用了 HTMOT。对于每一年（2019年和2020年），我们都获得了相应的主题树。然后，我们使用 JS 或 SD 计算了两年间所有主题对之间的距离。为此，我们使用每个主题排名前 100 的词和前 15 个实体来表示该主题。随后，我们对计算出的所有主题对进行排序，并迭代选择相似度最高的主题对（SD 或 JS 分数最低），确保每个主题只配对一次。最后，我们使用预先设定的阈值移除了分数较差的配对。

请注意，我们的方法未考虑结构信息。事实上，在层次主题建模的背景下追踪主题面临着另一个有趣的挑战：可能产生许多不同的树，而它们都同样正确。在一次运行中，我们可能提取出“太空”这一主题，其子主题可以归为“太空探索”和“天文学”。相反，在另一次运行中，我们可能会将“太空探索”和“天文学”提取为单独的主题，并各自带有自己的子主题。因此，很难利用主题树中包含的结构信息来追踪主题，因为不能指望它遵循特定的概念分类体系。

4 结果：JS 与 SD

本节将讨论基于语义的方法与传统的基于词汇的方法相比表现如何。

首先，我们研究了两种方法的重合情况，即两者都提取出的配对数量。我们发现，使用 JS 并将阈值设为 <0.4 时提取出了 111 对，而使用 SD 并将阈值设为 <0.1 时提取出了 121 对。这些阈值是根据经验观察设定的，但可能取决于所用的数据集。这 111 至 121 对可以分为三类（见图 2）。两种方法得出的相同配对有 72 对（占配对总数的 60% 至 65%）。例如，太空和电子游戏等主题很容易由两种方法在两个年份之间配对。这表明，我们的 SD 方法能够跨时间配对主题，且表现与 JS 相近。两种方法得出的不同配对有 39 至 49 对（占配对总数的 35% 至 40%），我们可以对这些配对进行评估。在这些不同的配对中，我们注意到，大多数情况下，一种方法（例如 SD）能够追踪／关联两个年份中的一对主题，而另一种方法（例如 JS）则无法做到，因为最佳配对的得分超过了阈值。这样一来，还剩下 10 对不同的配对，它们可以根据各自共同的 2019年或 2020年主题进行配对（见图 2）。

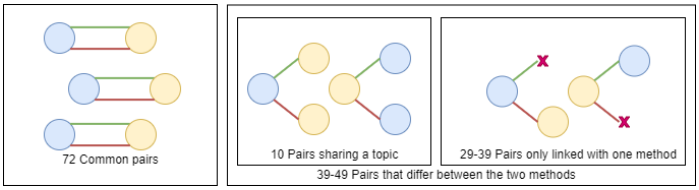


图2。 两种方法提取的配对可分为三类。圆圈表示主题，颜色表示年份（2019年为蓝色；2020年为黄色）。连线颜色表示所用的方法（JS 为红色；SD 为绿色）。这三类分别是：1）两种方法都提取出的配对（72）。2）有所不同但共享一个主题的配对（10）。例如，JS 提取了配对 4-D，而 SD 提取了 4-E。3）仅由一种方法建立关联的主题配对（29-39）。

为了比较这两种追踪方法的表现，我们决定通过问卷调查比较这 10 组由 JS 和 SD 提取的主题配对。具体来说，每道题都会给出一个初始主题，并向标注人员展示 JS 和 SD 的配对结果，请他们判断哪一种更好。此外，我们还请标注人员按 1 到 5 分的量表提供信心评分。我们总共收到了 38 份答卷，答卷者来自一个专注于回答问卷调查的小型在线社区[2]。问卷调查可在 GitHub 上找到 [3]。

查看问卷调查结果（表1）可知，SD 略胜于 JS，54% 的标注人员更倾向于前者。此外，我们还注意到，标注人员对自己的评估很有把握，平均信心评分为 3.3。有趣的是，

[2] https://www.reddit.com/r/SampleSize/[3] https://github.com/JudicaelPoumay/TopicTrackingPaper

答案之间存在很大差异。有些主题显然更适合其中一种方法（问题3和问题5），而另一些则不那么明确（问题1、问题2和问题4）。

表1. “选择 SD”列对应于选择 SD 配对为最佳配对的标注人员所占百分比。

问题 选择		
Q1	42.1% (22)	3.2
Q2	63.2% (14)	2.6
Q3	21.1%（30）	3.7
Q4	65.8% (13)	3.5
Q5	78.9% (8)	3.5
平均值	54%	3.3

例如，图3对应问题1。它展示了如何使用 JS 和 SD 将一个2019年主题与2020年主题进行配对。首先，我们可以注意到，两种方法记录的配对间距离都接近阈值。具体而言，JS 配对为 0.29，SD 配对为 0.09（JS 的阈值为 $= 0.4$，SD 的阈值为 0.1）。这很合理，因为两种方法都会提取优质配对（JS/SD 值较低的配对）。其次，2019年的主题涉及社交媒体数据安全，而选定的2020年主题则涉及：

– 与 JS 配对时的社交媒体。– 与 SD 配对时的数据安全。

因此，这两种配对似乎都合适，这或许可以解释标注人员为何犹豫不决。具体而言，16 人认为 SD 配对更好，而 22 人认为 JS 配对更好。他们对这个问题的置信度为 5 分中的 3.2 分。

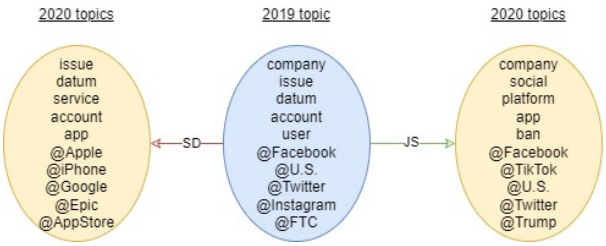


图 3. 同一 2019年主题下 SD 与 JS 之间不同配对方式的首个示例。

同样，图4对应问题5，展示了如何根据这两种方法，将另一个2019年的主题进行配对。这里，2019年的主题是关于网络安全的，而选定的2020年主题则是关于：

– 与 JS 配对时的数据安全。– 与 SD 配对时的
网络安全主题。

此外，SD 选择的主题是 JS 选择主题的一个子主题，这体现了在层级化设置中进行主题跟踪的难度。事实上，很难将一个主题与其子主题区分开来，尤其是在该子主题占主导地位时，因为父主题是其各个子主题的总和。在这种情况下，标注人员的意见更加一致，38 人中有 30 人认为 SD 配对更好。他们对这一问题的置信度为 5 分制中的 3.5 分。

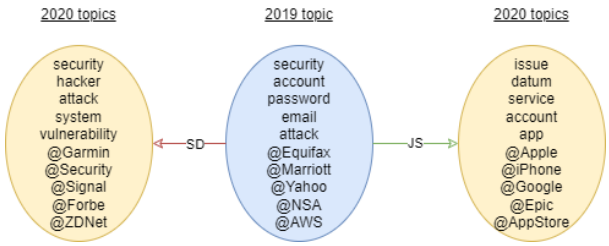


图 4. 同一 2019年主题中 SD 与 JS 采用不同配对方式的第二个示例。

因此，我们认为 JS 和 SD 是两种截然不同的方法，且二者各有优势。JS 由词汇驱动，因此可能最适合用于关联那些往往具有稳定且明确词汇的主题，例如法律文件中的主题。另一方面，SD 由语义驱动，可能更适合用于关联具有更强词汇变异性的主题。词汇变异性可能源于词汇漂移，即术语的流行程度随时间发生变化；也可能源于不使用标准词汇的非正式文本，例如推文。因此，我们认为，在主题跟踪方面，SD 不仅能与 JS 竞争，还能对其形成补充。

5 结论

本文提出了一种新颖的基于语义的主题跟踪方法（SD）。我们证明，该方法的性能可与基于词汇的最先进方法（JS）相媲美。这验证了我们的假设：语义信息对于跟踪主题具有重要价值。

此外，我们还讨论了在主题层级中进行主题跟踪所面临的挑战。首先，主题与其子主题可能难以区分，这使主题跟踪更具挑战性。其次，层级越深的主题越晦涩，因此更难评估其跟踪质量。最后，主题层级可能存在多种同样正确的排列方式，这使得难以利用结构信息进行主题跟踪。

我们相信，我们的工作将有助于未来关于主题跟踪混合方法的研究，例如整合词汇和语义信息。

参考资料

1. Agrawal, A., Fu, W., Menzies, T.：主题建模有什么问题？以及如何利用基于搜索的软件工程解决这些问题。Information and Software Technology **98**，74–88（2018）2. Allan, J., Carbonell, J.G., Doddington, G., Yamron, J., Yang, Y.：主题检测与跟踪试点研究最终报告（1998）3. AlSumait, L., Barbar´a, D., Domeniconi, C.：在线潜在狄利克雷分配：用于文本流挖掘的自适应主题模型及其在主题检测与跟踪中的应用。载于：2008年第八届IEEE国际数据挖掘会议。第3–12页。IEEE（2008）4. Blei, D.M., Ng, A.Y., Jordan, M.I.：潜在狄利克雷分配。J. Mach. Learn. Res. **3**（null），993–1022（2003年3月）5. Dagan, I., Lee, L., Pereira, F.：基于相似度的词义消歧方法。载于：第35届计算语言学协会年会暨第8届计算语言学协会欧洲分会会议。第56–63页。计算语言学协会，马德里，西班牙（1997年7月）。https://doi.org/10.3115/976909.979625，https: //aclanthology.org/P97 1008 6 - . 6. Fan, W., Guo, Z., Bouguila, N., Hou, W.：利用层次贝叶斯非参数模型进行基于聚类的在线新闻主题检测与跟踪。载于：第44届ACM SIGIR信息检索研究与发展会议论文集。第2126–2130页（2021）7. Ibrahim, N.F., Wang, X.：面向在线零售服务改进的文本分析方法：来自Twitter的证据。Decision Support Systems **121**，37 50 2019 10 1016 2019 03 – ( )。https://doi.org/https://doi.org/ . /j.dss. . .002, https://www.sciencedirect.com/science/article/pii/S0167923619300405 8. Jung, Y., Suh, Y.：挖掘员工心声：一种从在线员工评论中识别和分析工作满意度因素的文本挖掘方法。Decision Support Systems **123**，113074（2019）。https://doi.org/https://doi.org/10. 1016/j.dss.2019.113074, https://www.sciencedirect.com/science/article/ pii/S0167923619301034 9. Liu, H., Chen, Z., Tang, J., Zhou, Y., Liu, S.：绘制技术演化路径：一种用于动态主题检测与跟踪的新型模型。Scientometrics**125**(3)，2043–2090（2020）10. Poumay, J., Ittoo, A.：HTMOT：层次主题随时间建模（2021）。https://doi.org/10 03104 11 48550 2112 . / arXiv. . . Wang, Y., Xu, W.：结合深度学习与基于潜在狄利克雷分配的文本分析来检测汽车保险欺诈。Decision Support Systems **105**，87 95 2018 10 1016 2017 11 – ( )。https://doi.org/https://doi.org/ . /j.dss. . .001, https://www.sciencedirect.com/science/article/pii/S0167923617302130 12. Xu, G., Meng, Y., Chen, Z., Qiu, X., Wang, C., Yao, H.：在线新闻文本主题检测与跟踪研究。IEEE Access **7**，58407–58418（2019）13. Zhu, M., Zhang, X., Wang, H.：一种基于潜在狄利克雷分配的主题演化模型：来自信息科学期刊的证据。载于：2016年建模、仿真与优化技术及应用国际会议（MSOTA 2016）论文集。第49–54页（2016）