主题提取和选择:在线讨论论坛一览

贝尔纳多·佩雷拉·努内斯* ，亚历山大·梅拉 ，Ricardo Kawase† 信，马可·A·卡萨诺瓦 ，吉尔达·海伦娜·B·德坎波斯‡
息学系 - PUC-Rio - 里约热内卢,RJ - 巴西 {bnunes, acaraballo, casanova}@inf.puc-rio.br † L3S
研究中心,汉诺威莱布尼茨大学,德国
kawase@l3s.de
‡ 教育部 - PUC-Rio - 里约热内卢,RJ - 巴西 gilda@ccead.puc-rio.br

摘要 论坛在知识创造过程中起着关键作用,为用户提供交流思想和协作的手段。然而,教育论坛以及其他一些在线教育环境经常受到主题中断的影响。由于内容主要由参与者（在我们的例子中是学习者)制作,因此一个人或几个人可能会改变讨论的进程。因此,重新调整论坛主题的讨论主题通常是由导师或主持人执行的任务。为了支持学习者和导师和谐地协调与特定讲座或课程相关的论坛讨论,在本文中,我们提出了一种结合语义技术和统计方法来查找和展示在线讨论论坛中要讨论的相关主题的方法。我们与学生、教授和大学工作人员一起调查了我们的主题提取和选择方法的结果。结果表明该方法具有潜在的可用性,并且在实际学习场景中具有潜在的适用性。

一、简介

过去十年,万维网已成为信息和知识的重要来源。独立用户和社区的多样性和参与度促进了丰富内容的创造和传播,这些内容可通过不同的沟通渠道（如社交媒体、实时频道、博客、论坛等)和格式（如文本、音频和视频)获取。

尤其是在线论坛在知识创造过程中发挥了关键作用,为用户提供了交流思想、形成观点、定位自己和协作的手段。在线论坛的重要性体现在维基百科1 上,其中每个维基百科条目都有一个基于论坛的页面2,该页面依靠用户的协作、讨论、共识和集体努力来保持维基百科不断更新和整理。

由于用户参与论坛所产生的好处,大多数在线课程都结合了教育材料和在线讨论论坛。然而,尽管论坛明显利用了集体智慧的创造[10],但对用户参与度的评估却相当困难。根据学生数量和职位数量,人工评估变得不切实际。以前的工作解决了评估学生参与质量的问题[8],[7]。然而,他们没有考虑一组特定的主题是否在特定学科的线程中得到解决。

此外,在线讨论论坛中的不同背景可能会将讨论引向不可预见的方向,需要外部支持来重新调整论坛主题的讨论主题。这项任务通常由导师或主持人执行。但是,正如我们在本文中所示,平均而言,50% 的论坛与不同的受众或导师/主持人讨论特定主题,涵盖不同的主题。

尽管论坛讨论通常不同,但必须解决一组特定主题才能实现课程目标。因此,如果给定的论坛没有涵盖一组预期的主题,那么对参加不同论坛（同一主题)的学生的评估可能会受到阻碍,因为所获得的知识取决于论坛中讨论的主题。

在本文中,我们结合语义技术和统计方法来查找、展示和推荐在线论坛中要讨论的相关主题。简而言之,在语义工具的帮助下,所提出的方法首先执行命名实体识别 (NER) 和主题提取,然后采用统计方法选择和排序最具代表性的主题。该方法输出特定论坛中讨论的最具代表性的主题以及一组建议讨论的主题。我们使用了巴西大学的 97 个在线论坛来验证和评估我们的方法。

本文的主要贡献在于：(i)对导师/主持人进度进行高水平评估；(ii)主题推荐；(iii)论坛覆盖范围；(iv)在学生参与特定主题的论坛后均衡知识获取；(v)对所讨论的主题提供更好的概述。

本文的其余部分安排如下。第二部分讨论并比较了相关研究。第三部分描述了在我们的背景下使用论坛的情况和动机。

第四部分介绍主题提取和选择

1http://www.wikipedia.org
2http://en.wikipedia.org/wiki/Help:使用讨论页 _

方法。第五节和第六节分别介绍了我们方法的评估设置和结果。最后,第七节讨论了我们的成果并概述了未来的工作。

二.相关工作

Li 和 Wu [6] 结合情绪分析和文本挖掘方法来检测特定时间范围内的热点论坛。他们的方法可以帮助用户对在线论坛中两极分化的消息组做出决策和预测。尽管没有在热点论坛中进行主题提取,但提取的每个主题的情感极性信息将有助于用户理解在讨论中如何处理给定主题。

丛等人。 [1] 提出了一种基于标记顺序模式（LSP)和基于图的传播模型在在线论坛中查找问答对的方法。虽然疑问句的模式是使用词性标签创建的,但答案是使用 KL 发散语言模型检测和排名的。同样,我们的方法是对他们的方法的补充,因为我们的方法将作为根据主题查找问题和答案的过滤器。相反,我们的方法将通过识别在线讨论中的关键帖子来受益于这种方法。

Scaffidi 等人的研究显示,在线论坛在学生技能发展中发挥着关键作用。[9]他们的研究重点关注促进新手开发人员之间讨论和协作的帖子类型。在线论坛中的用户行为研究有助于促进用户之间的积极互动,从而促进集体知识的构建。我们相信,引入新的话题供此类用户社区讨论可以引发新的讨论,从而产生新的知识。

德桑克蒂斯等人。 [4] 提供了关于电子学习场所的有趣讨论,例如视频会议教室、在线社区和小组讨论空间。尽管每个地点都会影响特定群体的学习过程,但它们都有一个共同点,都需要带来新的讨论,以促进参与者知识的发展。例如,在线社区通常比私人小组讨论空间持续时间更长,因为提出新问题的新参与者可以随时加入。因此,为了维持小组讨论,推荐新的讨论主题将促进参与者之间更长时间的互动和知识更新。

显然,在线讨论也可以由负责提出新话题和新问题的导师推动。先前的研究 [2] 表明,有辅导的场所可以提高参与者的记忆力和表现。在本文中,我们使用该工具协助导师解决与讨论相关的新主题。

三、动机

为了说明我们研究的动机,我们描述了两种场景,其中在线讨论论坛的参与者会从我们的方法中受益。这两种情况都源于巴西一所大学的工作人员评估在线讨论论坛的需求。

由于在线讨论论坛在学习过程中至关重要,因此大多数在线课程都利用其来实现特定目标。然而,评估学生对论坛的参与度并不是一项简单的任务,而且由于帖子数量众多,它可能变得不切实际。因此,为了保持教学质量和学生体验,大学工作人员需要一种工具来跟踪讨论进度。

大学工作人员描述的第一个场景是,导师为了论坛的流畅性而不断忽视相关话题的讨论。然而,尽管讨论流程至关重要,但导师在主持论坛时必须确保讨论特定主题,同时保持讨论流程。因此,大学工作人员对论坛分析感兴趣,以检查讨论中是否涵盖了特定主题。通过这种方式,他们可以确保所有参与者都有相似的经历和学习情况,可以为接下来的活动做出贡献。

如果一组主题未涵盖,他们希望进行干预并延长论坛关闭时间或创建新的论坛线程来讨论缺失的主题。

作为第一种情景的结果,第二种情景旨在通过可能有助于学生参与讨论的建议来促进讨论。由于多种原因,一些论坛缺乏互动,必须鼓励学生参与。通过这种方式,大学工作人员认为推荐工具将促进讨论并有助于实现论坛的目标。

因此,当前的工作可以帮助大学教职员工和学生更好地了解论坛中发生的情况,以采取正确的行动并创造可以改善学生学习体验的学习情境。

四.主题提取和选择

在本节中,我们将介绍连贯流程链的主要步骤,该流程链从语义和统计上选择给定在线讨论论坛中最相关的讨论主题。流程链由以下三个步骤组成:(i) 实体提取和丰富;(ii) 主题提取;(iii) 主题选择。

A.实体提取和丰富

在处理在线讨论论坛时,我们本质上是在处理非结构化数据,这反过来又阻碍了数据操作和文本中原子元素的识别。为了缓解这个问题,采用了信息提取（IE)方法,例如命名实体识别（NER)和名称解析。这些工具自动从非结构化信息中提取结构化信息

数据并链接到链接开放数据云（LOD)中的外部知识库,例如 DBpedia3 。

例如,在使用 IE 工具处理以下句子后：“我同意巴拉克·奥巴马的观点,应该调查整个事件。”,实体“巴拉克·奥巴马”被注释并归类为人物,并链接到 DBpe-dia 资源 http://dbpedia.org/resource/Barack Obama,其中可以找到关于他的结构化信息。

—

我们使用 DBpedia Spotlight 工具4来提取和丰富论坛主题中帖子中找到的实体。DBpedia Spotlight 会添加带有围绕论坛帖子中的原子元素（实体)的语义信息的标记。请注意,只要我们拥有可靠的实体存储库（例如 DBpedia 或 Freebase5)和适当的注释工具（例如 Spotlight）,我们的方法是独立于语言的。

B.主题提取

以上一步中找到的实体作为起点,主题提取步骤首先遍历实体关系以找到实体的更一般表示,即主题。

实体通常以 (Sub-ject, Predicate, Object) 形式的 RDF (Resource Description Framework) 三元组表示,其中每个三元组代表一个事实,而谓词则命名了主语和宾语之间的关系。例如,三元组为 (“Barack Obama”, “isPresidentOf”, “United States of America”)。此外,一组 RDF 三元组形成有向标记图,其中节点是一组主语和宾语,边由谓词表示。

因此,对于帖子中每个提取和丰富的实体,我们通过谓词 dcterms:subject 探索它们的关系,根据定义 6代表实体的主题。从这个意义上说,为了检索主题,我们通过 DBpedia SPARQL 端点7使用 RDF 的 SPARQL 查询语言,在 DBpedia 层次结构中向上导航以检索实体与其主题之间更广泛的语义关系。正如以下 SPARQL 查询所示,我们使用谓词 skos:broader。

前缀 dcterms:<http://purl.org/dc/terms/> 前缀 skos:<http://www.w3.org/2004/02/skos/core#>

选择不同的?l1?l2?l3?l4 其中{

<entity_uri> dcterms:subject ?l1 . ?l1 skos:broader ?l2 . ?l2 skos:broader ?l3 . ?l3 skos:broader ?l4

限制 1000；

3http://www.dbpedia.org
4http://dbpedia-spotlight.github.io/demo/
5http://www.freebase.com
6http://dublincore.org/documents/2012/06/14/dcmi- terms/?v=terms#elements-subject
7http://pt.dbpedia.org/sparql - 葡萄牙语的 DBpedia SPARQL 端点。

变量实体 uri 表示我们感兴趣的从论坛主题的帖子中提取的主题的实体,而变量 l1 到 l4 表示将从实体中检索的主题。

因此,给定一个实体,可以通过谓词 dcterms:subject 和 skos:broader 检索实体的主题。后一个谓词用于获得主题的更一般的表示。该策略将帮助我们找到最能涵盖论坛主题的主题。

请注意,一个实体/概念可以在 DBpedia 的层次类别的不同级别中找到,因此这种方法会让我们检索不同类别级别的主题。但是,与 [5] 一样,我们利用不同级别中主题的共现来找到最具代表性的主题（参见第 IV-C 节）。

C. 主题选择

最后,在这最后一步中,我们从属于论坛主题的帖子中选择最具代表性的主题。为此,我们依靠 tf-idf（词频 - 逆文档频率)得分来统计衡量论坛主题中主题的重要性。

通常,tf-idf 用于信息检索和文本挖掘,以衡量单词对集合中文档的重要性。然而,在本文中,我们调整了此指标,以考虑从帖子中提取的实体和主题,而不是单词。

因此,为了选择最具代表性的主题,我们计算两次 tf-idf 分数,一次针对从论坛主题中提取的实体（即集合中最具代表性的实体）,另一次针对从实体中提取的主题（参见第 IV-B 节）。

基本上,为了计算词频 (tf),我们计算实体 e 在帖子 p ∈ P 中出现的次数。

至于逆文档频率 (idf),我们通过除以帖子总数 |P| 来计算 (idf) 分数包含实体|Pe|的帖子数量,参见等式 1。 1.

$$tf\ id\ f(e, p, P) = tf(e, p) \times id\ f(e, P) \quad (1)$$

其中 tf 是帖子中术语的原始频率 tf(e, p) = f requency(e, p),idf 是集合 P 中实体的常见性/稀有性的度量,由下式给出

以下方程:$id\ f(e, P) = \log(\frac{|P|}{|Pe|})$。

计算每个实体的 tf-idf 分数后,选择最具代表性的实体。从选定的实体中,根据第 IV-B 节中描述的过程提取主题。

掌握主题后,我们计算从实体中提取的主题的 tf-idf 分数,并对它们进行递减排名。同样,选择给定论坛主题的最重要的代表性主题。请注意,代表论坛的主题数量由用户选择（在我们的示例中,是前 10 个相关主题）。最后,选择排名靠前的主题来代表论坛主题。

五、评估设置

在我们的研究过程中,使用来自在线讨论论坛的真实数据对我们的方法进行全面评估。我们的方法使用 97 个在线论坛进行了评估,其中包含巴西大学远程教育部门提供的总共 10,785 个匿名帖子。所有选定的论坛主题的评估至少发生两次。此外,每位教授还对自己举办的论坛中建议的主题进行了评估。

我们的主要目标包括全面评估基于以前的在线讨论论坛所推荐的主题以及对涵盖论坛讨论的选定主题的评估。为此,我们向远程教育系的 11 名学生、4 名教授和 3 名协调员提交了 3 份问卷,以收集对所提方法的不同观点和看法。

问卷分为三类问题,即感知有用性、感知易用性和附加建议。问题基本遵循 Davis [3] 提出的技术接受模型 (TAM),该模型可以说是最具影响力的“技术接受理论”。问卷采用 5 点同意和频率的李克特量表。

简而言之,该理论指出,衡量用户采用新技术的意图有两个关键方面,即感知有用性和感知易用性。感知有用性（PU)是指“一个人相信使用某个特定系统会提高他或她的工作绩效的程度”,而感知易用性（PEOU)是指“一个人相信使用某个特定系统会提高他或她的工作绩效的程度”。特定的系统将是免费的”[3]。

每份问卷分为 6 个 PU 问题、6 个 PEOU 问题和另外 3 个意见挖掘问题,我们要求参与者提供进一步的建议。请注意,对于大学工作人员,主题是通过两个随机选择的论坛主题进行评估的,因为他们不参与论坛。因此,论坛中讨论的主题列表和每个论坛主题的建议主题列表可供他们评估。

由于他们是大学的工作人员,因此如果需要更多信息,他们也可以参与论坛讨论。

六.结果

调查问卷的结果如图 1 所示。误差条形图显示,所有参与者都对所提出的主题、结果的影响和适用性表现出高度积极的看法。特别是,与其他级别的参与者相比,教授的接受度稍高一些。 PU 的内部一致性系数 Cronbach s α 为 0.65,PEOU 的内部一致性系数 α 为 0.72,表明结果具有良好的可靠性。这些

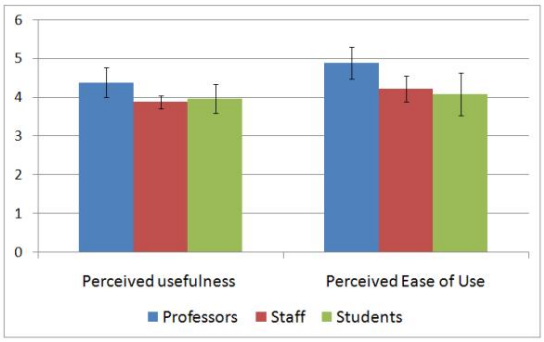

图 1.有关感知有用性和感知易用性的调查问题的误差线。

问卷结果表明我们提出的主题提取和选择方法的潜在可用性。

关于问卷中包含的建议,我们观察到最具争议的问题是推荐的主题是否应该向教授、学生或两者开放。所有教授都建议这些主题应该只对他们开放。所有工作人员都建议,主题应该对双方都开放。

有趣的是,学生们并没有达成共识。
虽然大多数人（64%)同意建议的主题应该向教授和学生开放,但 36% 的人认为主题建议应该只向教授开放。

我们认为,争议的起因是各组参与者的背景不同,对主题的理解也不同。教职员工并没有有效地参与在线论坛,他们认为讨论的主题应该来自教授和学生之间的协议。另一方面,教授认为工具应该直接向他们提供主题推荐,这实际上反映了他们控制周围人的需要。最后,学生意见的分歧在于,一些学生仍然怀疑在线教育论坛在没有适当管理的情况下能否顺利发展。

与向学生和大学工作人员提供的调查问卷不同,教授的调查问卷还有一个额外的问题,即其他教授是否可以从建议的主题中受益。结果表明,75% 的教授强烈同意其他教授会利用建议的主题。

最后,所有参与调查的大学工作人员和教授（强烈)同意,如果不同的论坛讨论相同的主题,将有助于对学生的评估。同样,所有大学工作人员（强烈)同意,拟议的方法将有助于评估教授在论坛中讨论的主题的覆盖范围。尽管如此,88% 的参与者同意这种方法的使用应该是可选的。

七.讨论与展望

我们提出了一种自动生成代表远程学习环境中的论坛主题的主题的方法。基本上,我们将语义和统计技术结合在一个连贯的流程链中,以提取、选择和排名论坛最相关的主题。

我们的实验表明,大多数教授、大学工作人员和学生愿意在未来的论坛中使用我们提出的方法。此外,75% 的教授表示其他教授将从建议的主题中受益。

通过查看 97 个论坛主题样本,我们证实,平均而言,不同论坛讨论的同一主题中 50% 的主题是不同的。这种情况引起了人们对论坛讨论的主题和学生的后期评估的担忧。按理说,不同论坛中讨论同一主题的学生应该有类似的经历并学习相同的主题。

因此,提供一种方法来概述不同论坛中讨论的主题将有助于大学工作人员(例如课程协调员)快速介入被忽视的主题的论坛。

理论上,使用所提出的方法可以更好地控制论坛中的教学内容,从而确保质量。在实践中,情况可能有所不同,一些受访者出于使用所提出方法的目的而产生了一些考虑。

首先要考虑的是教授指导论坛的自由。由于每位教授都有自己的教学风格,并且在处理某一主题时也可能有不同的观点,因此担心必须在论坛上讨论特定主题可能会降低一些教授的创造力和参与度。另一方面,助理教授也可以利用建议的主题来指导论坛。

关于使用所提出的方法的另一个考虑因素是学生对主题的使用,以防他们也可以获得主题建议。同时,主题建议可能会引发洞察力或使一些学生更加自信,而其他学生可能只坚持建议的主题并抑制对其他各种相关主题的讨论。

由于空间限制而提出的最后一个考虑因素取决于论坛的类型和所教授的课程。

在许多课程中,主题会随着时间的推移而变化,使用以前论坛线程中自动建议的主题可能会阻碍讨论流程。尽管主要主题将在未来的讨论中保留下来,但建议主题的列表必须不时手动更新。

总的来说,所提出的方法旨在帮助大学工作人员、教授和学生更好地了解论坛中正在讨论的内容,从而使教授能够采取更明智的行动保持讨论流程,改善学生的体验并确保主题覆盖范围。

我们的方法还为大学工作人员提供了评估论坛覆盖范围、跟踪学生在不同论坛中学习的内容以及在某些情况下检测论坛中讨论的主题的偏差的可能性。

我们认为,在在线课程中是否采用所提出的方法取决于课程的教学设计。课程的设置对于确定必须使用哪些方法或工具至关重要。

对于未来的工作,我们计划扩展接受外部主题建议的方法。例如,参与课程的教授也可以在讨论中添加主题。

此外,我们还计划创建一个 Moodle 插件。

参考

[1] G. Cong、L. Wang、C.-Y. Lin、Y.-I. Song 和 Y. Sun。从在线论坛中查找问答对。第 31 届 ACM SIGIR 国际信息检索研究与开发会议论文集,SIGIR 08,第 467-474 页,美国纽约州纽约,2008 年。ACM。

[2] JA Cottam、S. Menzel 和 J. Greenblatt。《辅导以保留学生的知识》。《第 42 届 ACM 计算机科学教育技术研讨会论文集》(SIGCSE 11),第 213–218 页,美国纽约州纽约,2011 年。ACM。

[3]FD戴维斯。感知有用性、感知易用性以及用户对信息技术的接受程度。 MIS 季刊,第 319-340 页,1989 年。

[4] G. DeSanctis, A.-L. Fayard, M. Roach, 和 L. Jiang. 在线论坛中的学习. 欧洲管理杂志, 21(5):565 – 577, 2003.

[5] B. Fetahu、S. Dietze、BP Nunes、D. Taibi 和 MA Casanova。生成链接数据图的结构化配置文件。收录于 E. Blomqvist 和 T. Groza 编辑的《国际语义网会议》,CEUR 研讨会论文集第 1035 卷,第 113-116 页。CEUR-WS.org,2013 年。

[6] N. Li 和 DD Wu。使用文本挖掘和情感分析进行在线论坛热点检测和预测。决策支持系统, 48(2):354 – 368,2010 年。

[7] M.彭德加斯特。用于评估在线讨论论坛中学生参与和实施动态的分析工具。 SIGITE Newsl.,3(2):10–17,2006 年 6 月。

[8] C. Romero、M.-I. Lopez、J.-M. Luna 和 S. Ventura。通过参与在线讨论论坛预测学生的最终表现。计算机教育,68:458–472,2013 年 10 月。

[9] C. Scaffidi、A. Dahotre 和 Y. 张。在线论坛如何促进新手动画程序员之间的讨论和协作?载于 LAS King、DR Musicant、T. Camp 和 PT Tymann,编辑,SIGCSE,第 191-196 页。

美国CM,2012。

[10] AL Veerman、JEB Andriessen 和 G. Kanselaar.通过计算机辅助论证进行协作学习。

1999 年计算机支持协作学习会议记录,CSCL 99。国际学习科学学会,1999 年。