吉林大学本科毕业设计（论文）外文文献翻译

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 学 院 | 软件学院 | | | | 专业 | 软件工程 |
| 学生姓名 | 杨树德 | | | | 学号 | 55160411 |
| 指导教师 | 柴胜 | | | | 职称 | 讲师 |
| 合作导师 | 无 | | 职称 |  | 单位 |  |
| 设计（论文）题目 | | 基于Android平台的今日头条app的设计与开发 | | | | |
| 一、外文文献翻译（翻译字数不少于3000字）  **基于本体的个性化财经新闻推荐算法**  任瑞a,b,c，张玲玲a,b,c\*，崔丽萌b,c,d，邓波d,e，雍世b,c  *中国科学院大学a管理学院，北京100190*  *b中国科学院大数据挖掘与知识管理重点实验室，北京100190c中国科学院虚拟经济与数据科学研究中心，北京100190d计算机与控制工程学院中国科学院大学，北京100190e中国科学院计算技术研究所，北京100190*  **摘 要**  为了应对信息过载的挑战，在本文中，我们提出了一种金融新闻推荐算法，可以帮助用户找到有趣的文章。为了解决歧义性问题，采用了一种新提出的OF-IDF方法，以关键概念，同义词和同义词集的形式表示非结构化文本数据，这些数据都存储在领域本体中。对于用户而言，推荐算法会根据他们的行为构建配置文件，以通过应用相关反馈的思想自动实时地检测出真正的兴趣并预测当前兴趣。最后，对金融新闻数据集进行的实验表明，该算法明显优于传统推荐器的性能。  *关*键字：新闻推荐算法；本体相关性反馈；国防军  **1.介绍**  随着万维网的飞速发展，人们倾向于在线阅读新闻，而不是通过订阅实际报纸阅读。同时，金融在世界上扮演着越来越重要的角色，如今人们比以往任何时候都对金融事件和知识更加好奇。人们面临的关键问题是信息过载，这意味着财经新闻的数量使用户不知所措。因此，在本文中，我们提出了一种金融新闻推荐方法，旨在向用户提供有趣的新闻文章，不仅可以帮助个人节省时间和精力，还可以企业提高用户忠诚度。  到目前为止，有三种不同的新闻推荐方式：基于内容的推荐，协作过滤推荐和混合推荐。基于内容的方法在推荐器系统中起着核心作用，因为它能够推荐以前未评级的信息，并能适应用户之间的个体差异[1,2]。因此，在本文中，我们将重点放在基于内容的推荐上。它的缺点之一是缺乏语义[3]。为了解决这个问题，我们使用本体来存储词汇化的金融领域概念，关系和同义词集。准确的用户配置文件是基于内容的推荐方法的关键部分。我们采用相关性反馈的思想来自动发现用户的真正兴趣，而不是手动创建和更新个人资料。  本文的贡献是三方面的。首先，我们构建一个特定的金融领域本体，该本体试图存储有关一定数量金融新闻的所有最重要的信息。其次，本文提出了一种称为OF-IDF的新方法来表示非结构化文本数据。第三，我们运用相关反馈的思想构造用户档案，自动检测用户的真实兴趣。  其余论文的结构如下。第二部分回顾了有关新闻推荐和本体的一些相关文献。第三部分介绍了构建金融领域本体的方法。第4节提出了用OF-IDF表示文本数据的方法。第5节详细介绍了个性化财经新闻推荐算法。第6节介绍了实验结果及其评估。第7节总结了论文并指出了我们的未来工作。 2.相关工作 新闻推荐器系统中通常使用三种不同的技术：协作过滤推荐，基于内容的推荐和混合推荐。  协同过滤推荐重点在于用户相似性或项目相似性。它已应用于个性化新闻阅读应用程序，例如GroupLens，这是第一个用于对网络新闻进行协同过滤的系统，可帮助人们在大量可用文章中找到他们可能喜欢的文章[4]和Google新闻推荐器的第一版。 [5]。尽管该方法可以检测到用户的潜在兴趣，但它并未考虑项目的内容，并且很难解决可伸缩性，稀疏性和冷启动问题。基于内容的推荐集中在用户配置文件上，该用户配置文件是通过分析用户过去访问和喜欢的项目的内容而构建的。它擅长处理文本数据，可以推荐不受欢迎的产品并列出产品的特征以说明原因。混合方法是前面提到的两个推荐系统的混合，试图结合两个方面的优点[6]，但是，它不能从根本上解决难题。  在本文中，我们将讨论限于基于内容的推荐，因为它不需要收集同行用户的意见，可以充分了解新闻的内容并生成具有良好解释的推荐。  术语频率倒文档频率-IDF）是Salton等人于1975年提出的一种众所周知的，最常见的基于内容的推荐方法[7]，通常用于信息检索和文本挖掘中。它是一种统计量度，用于评估单词对文档集合或语料库中文档的重要性。重要性与单词在文档中出现的次数成正比，但被语料库中单词的出现频率所抵消。结合矢量空间模型和相似度计算，TF-IDF可用于向特定用户推荐新闻项目[8]。TF-IDF是基于术语的，需要计算文本中所有术语的TF-IDF权重，因此，它不适合处理大数据。此外，基于术语的方法可能会导致语言歧义。  为了解决歧义问题，本文尝试利用本体对TF-IDF方法进行优化。本体不仅可以呈现关键概念和关系，而且还可以处理信息的内容[9]。术语“本体论”来自与研究存在有关的哲学领域。在计算机和信息科学中，本体是一个技术术语，表示为某个目的而设计的工件，这是为了使有关某个领域（真实的或想象的）的知识能够建模[10]。根据Thomas R. Gruber（1993）的观点，本体是对概念化的明确说明[11]。Borst（1997）认为本体论是领域内共享知识的形式化描述[12]。Rudi Studer等人（1998年）将本体定义为对共享概念化的正式，明确的规范[13]，这表明四个因素对于本体必不可少：概念化，显式，形式化和共享。本体之所以如此流行的原因很大程度上是由于它们的承诺：对可以跨人和计算机进行交流的某些领域的共享和共识[13]。在自然语言应用中，本体可以用于自然语言处理[14]，或者自动或半自动从文本中提取知识。Wordnet是英语中最大的词汇本体之一。名词，动词，形容词和副词被分组为认知同义词（同义词集），每组表达不同的概念。同义词集通过概念语义和词汇关系相互关联[15]。 3.金融领域本体 本文基于Hogenboom F，Vandic D，Frasincar F等[3]提出的2010年新闻数据库中的1823个金融新闻，构造了一个特定的金融领域本体，该本体试图存储有关以下内容的所有重要信息财经新闻。本体是可以跨人和计算机进行交流的某些领域的共享和共识[10]。领域本体包括捕获文章语义上下文以及它们之间的关系的关键概念，例如hasCompetitor和isCEOof。例如，如果用户对Google的新闻感兴趣，那么他很可能喜欢Yahoo的信息。此外，对于本体中的每个概念，都定义了同义词属性，以消除概念的词汇表示形式的歧义。还定义了WordNet感知属性以检索概念的同义词集。本体以OWL格式存储，在多种软件环境中都很好地支持该本体。与基于术语的推荐者相比，专注于本体的推荐者可能会产生更快，更准确的推荐，因为他们不需要考虑所有单词。与语言不同，概念不是模棱两可的。此外，概念之间的关系可以从多个维度检测用户满足其需求的潜在兴趣。 4.金融新闻的结构化表示 要推荐的财经新闻文章用表示，N是文章总数。本体中的概念表示为，M是在给定的文章中予以推荐的概念的总数。是第j条文章中第i个关键字的权重。是第条中的第个。因此，文章中的本体频率计算如下：  m是合计条中概念的数量    本体频率衡量一个概念在文章中出现的频率，而逆文章频率则通过权衡频繁的概念而扩大稀有的概念来衡量概念的重要性。的逆频率为IDF，显示文章中发生了多少次。    *ki的OF-IDF为：*    因此，文章d*j*表示为包含未读新闻文档中不同单词的所有OF-IDF值的向量   5.个性化财经新闻推荐算法   假设用户i阅读了M篇文章，并且这些文章已经通过Sector4中的方法表示为本体权重向量，N是M篇文章中出现的概念总数，a,b,c...r表示OF-IDF权重。因此，用户配置文件直观地表示为图1所示的用户概念矩阵。  假设用户i单击了ArticleI，用户i不仅阅读了ArticleII，而且还喜欢它，用户i单击了Articlel但不喜欢它，在这种情况下，我们使用了相关反馈的思想来修改兴趣爱好。修改后的用户概念矩阵如图2所示，D，E，J为修改后的参数，显然，D，E为正，而J为负。修改后的用户个人资料的计算公式如下，其中使用了相关反馈的思想，其中SD显示用户i单击了m篇文章，SE显示用户i阅读了n篇文章并且也喜欢它们，SJ显示用户i单击了l篇文章，但不喜欢它们，SD，SE，SJ都是向量。因此，用户个人资料表示为矢量与用户个人资料的新闻文档中的单词不同。    最后一步是计算新闻内容（dj）与用户个人资料之间的相似度，并将相似度描述为，其中是向量的范数。    所有结果均按降序排序，并且向用户推荐相似度值高于临界值的那些新闻文档。  流程图中显示了实现算法的过程。    图3.算法流程图 6.实验 *6.1实验设计*  在本节中，我们将对提出的算法进行实验研究。新闻数据来自Hogenboom F，Vandic D，Frasincar F等[3]提出的2010年新闻数据库，其中包括1823个金融新闻，涵盖了不同方面的2010年国际金融新闻。  我们建立了一个临时网站来收集用户的浏览历史记录。对于每篇文章，用户可以选择指出是否有趣。我们选择了33个注册用户的浏览数据，其中包含3600条记录。  删除停用词，阻止并计算OF-IDF之后，对每个新闻进行矢量化处理。数据集的处理基于监督学习。每个用户的浏览历史记录的五分之四数据被视为训练集，而五分之一的数据被视为测试集。训练集用于创建用户个人资料，并通过实施我们上面提出的算法和基线方法（经典的TF-IDF加权和余弦相似度）向用户推荐有趣的新闻。测试集用于通过比较推荐结果与包含真实用户行为的测试集之间的相似性来评估算法。  *6.2评估指标*  精度和召回率是信息检索领域中的常见指标。F1分数同时考虑了准确性和召回率。  精度：正确推荐的网页数除以建议的总页数。回忆：推荐的网页数除以用户访问的网页数即可。F1得分：精确度和查全率的谐波平均值。  运行时间：运行算法的总时间成本。  *6.3实验实施*  我们采用五重交叉验证方法来处理每个用户的浏览数据，并将其随机分为5部分，其中4部分被视为训练数据，其余部分作为测试数据。然后我们在相同的参数设置下执行推荐算法。另外，根据某些算法的推荐结果，计算了精度和召回率。最后，我们计算了五种不同训练集的准确性和召回率的平均值。  在本实验中，我们将我们的方法与基线方法进行了比较。根据前面的实验结果，我们提出的算法在精度，查全率和F1得分方面都优于基线方法。这表明金融本体可以更准确地表征用户的阅读习惯。此外，我们提出的方法可以更好地捕捉用户的真实兴趣。我们总结了33个用户的推荐结果，并在下面对两种算法进行了比较。    图4显示使用基准方法的所有用户的精度，召回率和F1分数。从图中我们可以看到，对于某些用户而言，精度相当高。但是对于某些用户而言，准确性，召回率和F1得分都非常低。随后，我们可以得出结论，基准线方法无法推断用户的兴趣，并且无法生成用户可能感兴趣的新概念。    图5使用我们的方法演示所有用户的准确性，召回率和F1得分。请注意，某些用户数据下的准确性已下降，我们可以推断出，尽管本体为用户提供了对不同新闻门户的广泛视野，但这可能会导致准确性下降。    图6.提出算法的改进  图6体现了我们的方法与基准方法之间的改进。透明条是我们方法的改进。显然，我们可以看到准确性，召回率和F1得分得到了显着提高。  表1.两种方法的运行时    从表1中可以看出，基线方法的运行时间随着新闻数据量的增加而迅速增加。随着新闻数量的增加，单词词典的大小将成倍增长。并且计算成本将是无法忍受的。但是，在我们的方法中，字典取决于本体数据库中的概念。因此，对于大型数据集，我们的方法取得了更好的效果。 7.结论 在本文中，我们提出了一种金融新闻推荐方法，旨在向用户传递有趣的新闻文章，不仅可以帮助个人节省时间和精力，还可以帮助企业提高用户忠诚度。为了解决歧义性问题，我们首先构建了一个特定的金融领域本体，该本体存储了关键概念及其之间的关系。其次，使用一种新的提出的称为OFIDF的方法来表示非结构化文本数据。此外，提出了一种推荐算法，通过运用相关性反馈的思想对用户的真实兴趣进行建模并自动预测当前兴趣。最后，我们进行了一项实验，将新方法与采用经典TF-IDF加权和余弦相似度的基线方法进行比较。实验结果表明，本文提出的算法明显提高了新闻推荐的质量。  在未来的工作中，我们希望将我们的算法与[19，20]中提出的其他语义方法结合起来。此外，位置数据可用于扩展我们的模型，以提高新闻推荐的效果。最终，最好将我们的算法与其他一些传统方法进行比较 致谢 作者非常感谢中国国家自然科学基金（授权号71471161、71071151、71331005、71110107026）的研究支持。 参考文献  1. R. Carreira，R.，Crato，JM，Gonçalves，D.，Jorge，JA评估新闻分类的自适应用户配置文件，第9届智能用户界面国际会议论文集，2004年 2. Good，N.，Schafer，JB，Konstan，JA，Borchers，A.，Sarwar，B.，Herlocker，J.，Riedl，J.将协作过滤与个人代理相结合以获得更好的建议，第16届全国人工会议论文集人工智能与人工智能的第十一届创新应用大会，人工智能的创新应用，1999年。 3. Hogenboom F，Vandic D，Frasincar F等。Hermes新闻处理框架中新闻项的查询语言和排名算法。计算机编程科学，2014，94：32-52。 4. Resnick P，Iacovou N，Suchak M等。GroupLens：一种用于协作过滤netnews的开放架构。1994年ACM计算机支持合作会议会议记录。ACM, 1994: 175-186. 5. Liu J，Dolan P，Pedersen ER。基于点击行为的个性化新闻推荐。第十五届智能用户界面国际会议论文集。ACM, 2010: 31-40. 6. Adomavicius，G.，Tuzhilin，A.迈向下一代推荐系统：最新技术和可能扩展的调查。IEEE知识和数据工程学报，2005，17（6）：734–749。 7. Salton G，Wong A，Yang CS。用于自动索引的向量空间模型。ACM通讯，1975，18（11）：613-620。 8. Goossen F，IJntema W，Frasincar F等。使用CF-IDF语义推荐器的新闻个性化。Web智能，挖掘和语义学国际会议论文集。ACM，2011年：10。 9. McGuinness DL，Van Harmelen F. OWL网络本体语言概述。W3C建议，2004，10（10）：2004。 10. 汤姆·格鲁伯。数据库系统百科全书，Ling Liu和M. TamerÖzsu（编辑），Springer-Verlag，2009年。   <http://tomgruber.org/writing/ontology-definition-2007.htm>   1. Gruber TR。一种可移植本体规范的翻译方法。知识获取，1993，5（2）：199-220。 2. Borst WN。构建用于知识共享和重用的工程本体。特温特大学，1997年。 3. Studer R, Benjamins VR, Fense ID.知识工程，原理和方法。数据与知识工程，1998,25（12）：161〜197 4. ] Bateman J A.关于本体论建构与自然语言之间的关系：一种社会符号学观点。国际人类计算机研究杂志，1995，43（5）：929-944。 5. <http://wordnet.princeton.edu/> 6. Lavrenko V，Schmill M，Lawrie D等。财经新闻推荐的语言模型。第九届信息和知识管理国际会议论文集。ACM, 2000: 389-396. 7. 李青，王健，陈玉平，等。用户对基于论坛的社交媒体中的新闻推荐的评论。情报科学，2010，180（24）：4929-4939。 8. Das AS，Datar M，Garg A等。Google新闻个性化：可扩展的在线协作过滤。第16届万维网国际会议论文集。ACM, 2007: 271-280. 9. 商业智能应用程序和Web：模型，系统和技术。商业科学参考，2012年。 10. IJntema W，Goossen F，Frasincar F等。基于本体的新闻推荐。2010年EDBT / ICDT研讨会论文集。ACM，2010年：16。 11. 崔莉，史莹。一种基于一类支持向量机的新闻推荐方法。Procedia计算机科学，2014，31：281-290。 12. 崔丽，孟芳，史Y，等。一种基于LDA和SVM的新闻分类方法。数据挖掘研讨会（ICDMW），2014 IEEE国际会议。IEEE, 2014: 60-64.   二、文献出处：  **[1]Rui Ren,Lingling Zhang,Limeng Cui,Bo Deng,Yong Shi. Personalized Financial News Recommendation Algorithm Based on Ontology[J]. Procedia Computer Science,2015,55.** | | | | | | |

**注：**1.专业一定要填写全称。[专业名称：软件工程、软件工程（国家卓越工程师教育培养计划）]

2.附外文文献。