doi:10.3772/j.issn.2095-915x.2019.02.003

基于学术论文的学者研究兴趣标签发现研究

池雪花 刘丽帆 章成志

南京理工大学信息管理系 南京 210094

开放科学（资源服务）标识码（OSID）

摘要：标签构建对信息检索和个性化推荐有重要的辅助作用，学者的研究兴趣标签体现了一定时期内学者和某一个领域的研究热点与发展方向。以学者为研究对象，对学者的研究兴趣标签进行发现研究，有助于学者兴趣标签自动构建与推荐，对加强学术交流合作有重要作用。本文基于学术论文信息，采用 LDA 与 Doc2Vec 两种文本表示方法，对学者和兴趣标签分别进行表示，然后计算两种方法得到的学者与研究兴趣标签的余弦相似度，最终采用集成方法对兴趣标签进行融合，得到学者的研究兴趣标签。结果证明，集成方法能够获得更好地标注效果。

关键词：兴趣标签；LDA；Doc2Vec；余弦相似度；集成方法

中图分类号：G35

Analysis of Scholars Research Interest Tag Discovery Based on Academic Papers

CHI Xuehua LIU Lifan ZHANG Chengzhi

Department of Information Management, Nanjing University of Science and Technology, Nanjing 210094, China

Abstract: Tag construction plays an important role in information retrieval and personalized recommendation. Scholars' research interest tags reflect the research hotspots and development directions of scholars and a certain field in a certain period. Taking

基金项目：富媒体数字出版内容组织与知识服务重点实验室开放基金项目（ZD2018-07/01）：“富媒体数字出版内容的知识挖掘及发现技术研究”。

作者简介：池雪花（1993-），硕士生研究生，研究方向：文本挖掘与科学计量，E-mail：joyce.chi@qq.com；刘丽帆（1995-），硕士研究生，研究方向为文本挖掘与科学计量；章成志（1977-），博士，教授，博士生导师，研究方向：信息组织、信息检索、数据挖掘及自然语言处理。

scholars as the research object, research on the scholars' research interest tags is helpful to automatic construction and recommendation of scholars' interest tags, which plays an important role in strengthening academic exchanges and cooperation. Based on the academic paper information, this paper used LDA and Doc2Vec to represent the scholars and interest tags respectively. Then, we calculated the cosine similarity between the scholars and the research interest tags obtained by these two methods. Finally, we use the ensemble method to fuse interest tags and get the research interest tags of scholars. The results turns out that the ensemble method can achieve better tagging results.

Keywords: Interest tag; LDA; Doc2Vec; cosine similarity; ensemble method

引言

大数据时代，技术的支持和互联网的快速普及让网络上的信息呈现指数增长，导致人们从大量数据中获取所需信息变得困难。标签具有很强的代表性，通常能够帮助用户在海量的信息中有效定位，兴趣标签能够让用户方便且快捷地找到所需资源或者找到相同兴趣的用户[1]。当下用户的相关数据不断数字化、丰富化，为用户兴趣标签的自动构建提供了资源支持，自然语言处理技术的发展也为用户兴趣标签发现提供了技术支持。

在学术界，科研学者和学术研究成果数量的猛增，推动了科学研究发展的同时，管理与分析工作也面临着巨大挑战。学者研究兴趣不仅能够体现学者本身的学术研究内容与方向，也体现学者对某一个或多个科研领域的关注程度。学者的研究兴趣发现是学者用户画像技术中的基本任务之一[2]，本文所提出的学者研究兴趣标签是指一个学者自身的科研内容信息，是对一个学者的整体性描述[3]。学者兴趣标签的目的就是对学者研究兴趣进行准确定位，为所需用户提供一个简洁、直观的认识。

学者研究兴趣标签的自定义设置出自学者的自我意愿，部分学者并没有为自己的研究兴趣设置相应的标签；而对于设有标签的学者，其研究内容会随着时间产生变化，学者自定义标签会出现更新不及时、匹配不准确、内容不全面等问题。这些问题不利于推动机构与学者、学者与学者之间的交流合作。学术相关资源数量巨大且分布散乱，这些现象让学术信息的检索变得繁琐，加大了学者兴趣标签自动构建的难度，降低了学者画像的准确度[2]。因此，学者的研究兴趣标签发现具有重要的现实意义。

学者研究兴趣的标签生成主要依据学者的科研数据，本文从学者的研究兴趣维度出发，基于对学术论文的挖掘，有效利用学者的科研数据，对学者的研究兴趣标签进行发现研究。

1 相关研究

本研究以学者为研究对象，基于学者已发表的学术论文信息，从中挖掘学者的研究兴趣，构建学者的研究兴趣标签，对学者的研究兴趣进行标签推荐。根据本文的主要研究内容，本

doi:10.3772/j.issn.2095-915x.2019.02.003

节从用户兴趣挖掘研究和标签推荐研究两个角度对现有工作进行梳理。

1.1 用户兴趣挖掘研究概述

用户兴趣主要是通过用户的行为体现出来的，对用户的兴趣挖掘研究主要体现在对用户的行为记录进行挖掘。石光莲等人在对用户的兴趣研究中指出，用户兴趣主要从基于资源和基于用户两种思路展开，传统的用户兴趣研究主要基于统计和聚类等方法[4]。例如黄镇圣对高校图书馆用户进行研究，采用用户在图书馆网站的浏览记录作为研究数据，只考虑了资源方面[5]。石豪等人考虑到了用户之间的相似性，但这些研究主要基于简单的统计方法[6]。易明等人从用户、资源、标签的聚类进行用户兴趣挖掘[7]。这些都属于传统的研究方法，缺乏语义层面的研究。

社交网络的兴起使得用户在网络上的发文内容、传播信息以及社交关系等资源增多，这些信息从用户角度出发，包含了用户对自身的兴趣认知。Weng 等人从社交软件中提取文本内容，采用 LDA 算法挖掘主题分布，从而获取用户的兴趣特征[8]。Macskassy 对用户微博平台的关键词进行挖掘，构建用户兴趣属性向量[9]。Hung 等人通过用户自定义标签、用户收藏以及社交关系，构建用户兴趣模型[10]。

此外，随着当下社交平台的多样性发展，用户的兴趣挖掘研究从单一平台研究发展到多平台、跨平台关联分析，基于资源与基于用户的研究变得更加全面。Bruce 等通过数据库等大数据信息，以用户为中心，从多平台、跨平台的多源数据中寻找用户的兴趣特征[11,12]。针对当下主流网络品台“内容化”的特征，屠守中等人从用户角度，实现了在大规模异构社交网络中挖掘用户兴趣[13]。

1.2 标签推荐研究概述

标签作为主体特征的一种代表，可以由用户来定义，但随着标签的广泛应用，质量问题影响了标签作用的发挥[14]。为了提高标签的质量，有效发挥标签的作用，标签推荐研究逐渐增多[15]。目前，国内外对标签推荐已经有许多研究，主要分为通过统计方法进行标签推荐以及通过文本处理方法进行标签推荐。基于统计方法的推荐是传统的推荐方法，一般通过对用户的标签以及关键字等信息进行统计，整合后排序，例如 Mishne 和 Sood 等人的用户推荐研究[16,17]。基于对文本处理的推荐方法是对文本内容展开的，例如 Fujimura 与 Tatu 等人基于文档内容进行标签推荐[18,19]。但许多基于内容的标签推荐过程存在语义模糊的问题，忽略了标签之间的语义联系[20]。因此，王晓耘等在进行个性化推荐研究中进行了潜在语义主题挖掘[21]，吴超等关注到了标签之间的语义相关性[22]，吴小兰等构建了标签语义网进行标签推荐[23]。

目前有关学者标签推荐的研究相对较少。传统的学者标签推荐多通过统计方法，目前的标签推荐多采用分类模型，并且逐渐出现组合推荐算法[24]。学者标签推荐主要针对学者的学术信息，涉及学者信息标注、学者兴趣挖掘、标签生成及推荐等方面[2]。而孙赛美等从学者的社交网络出发，利用 LDA 信任度进行学者推荐[25]。

综上所述，由于数字化资源数量的增加和

多元化，以及数据挖掘技术的推动，用户兴趣发现研究对象从资源到用户，研究范围从单平台到跨平台，研究方法从传统的统计到文本信息处理，各方面逐渐由单一变得全面。用户兴趣挖掘研究能够更好地为用户提供个性化服务。标签推荐能够为用户提供一个简洁且具有综合性的认识，在精准定位所需信息过程中具有重要作用。当下的标签推荐主要是基于内容的推荐，多采用分类方法。而结合自身语义特征进行推荐更加有效，能够改进用户标签的质量。本文以学者为研究对象，基于学者发表的论文空间和研究兴趣标签空间，通过 LDA 与 Doc2Vec 两种方法的集成化结果，进行学者研究兴趣标签推荐。

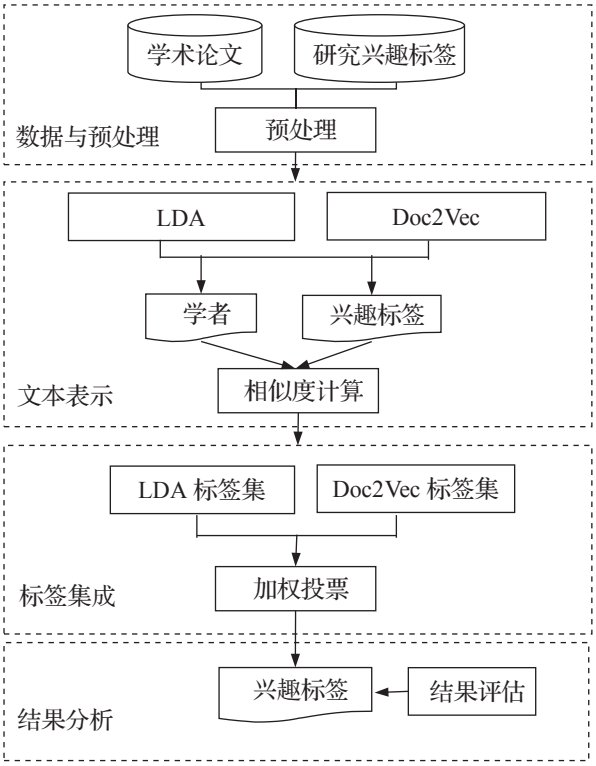

图 1　学者研究兴趣标签发现研究思路图

2　研究思路及实现方法

2.1　研究思路

学者研究兴趣标签是通过学者发表的学术论文体现的。本研究的核心思想是利用学者已发表的学术论文空间，和研究兴趣标签空间，计算学者和兴趣标签的相似度，将相似度作为学者与该研究兴趣标签的匹配度。本文的研究框架如图 1 所示。

本研究首先将获取得到的每位学者的所有论文文本内容连接成文档，对文本文档进行预处理，构成学者的学术论文空间；同时根据每个兴趣标签对应的学术论文集合进行文本表示与预处理，构成研究兴趣标签空间。然后分别采用 LDA 和 Doc2Vec 两种模型得到对应的学者研究兴趣标签集，进行余弦相似度计算。最后，通过集成法选取最终的学者研究兴趣标签。

2.2　实现方法

2.2.1　数据预处理

学术论文中包含章节、段落、句子、图片、表格等信息，本文针对获取的学术论文的摘要文本数据进行预处理。预处理过程主要包括文字检查，统一大小写，词干提取，词形还原，去停用词等步骤。

数据预处理过程提取了学者学术论文的文本信息，将所有论文内容连接成文本文档；然后，对每个兴趣标签对应的学者所发表的论文进行汇总，生成另一个文本文档，用来表示兴趣标签。因此首先对论文中的图片和表格等信息进行过滤。文字检查是对文章中可能出现的拼写错误进行检查，保障数据的准确性。学者的学术论文采用英文的文章，需要统一改为小写模式，

doi:10.3772/j.issn.2095-915x.2019.02.003

方便后期对文本进行统一处理。

词干提取和去停用词是数据预处理过程中的主要部分。词干提取也叫做词性规范化，是将一个词出现的不同形态进行归并。例如一个英文单词“have”，可能出现的形式有现在进行时“having”，第三人称单数形式“has”，过去时“had”。通过词干提取，可以将属于同一个单词的不同形式进行统一，也能够去除单词中出现的不同词缀，获取词根，提高文本处理效率。本文的词干提取过程主要采用了 nltk.stem 模块。

去停用词是删除文本中包含的语气词、助词、标点符号等没有实际意义的词，这些停用词都综合在停用词表内。中文停用词表中还包括一些单个的汉字、英文字符等形式。本研究所需的语料都是英文，根据研究所需的语料还可以向停用词表内添加停用词。

2.2.2 学者和研究兴趣标签表示

本研究主要采用了自然语言处理的方法，文本挖掘需要对文本信息进行处理。首先将每位学者的学术论文文本进行整合，表示成计算机能够处理的表示形式。然后采用 LDA 和 Doc2Vec 两种模型进行文本表示。

（1）LDA 主题模型

在对学者的研究兴趣标签进行表示时采用了 LDA 和 Doc2Vec 两种不同的表示方法，分别属于基于主题模型的方法和基于神经网络的方法。LDA 是一种文档主题生成模型，全称 Latent Dirichlet Allocation，由 Blei 在 2003 年提出 [26]。LDA 是一个三层贝叶斯概率模型，包含单词层、主题层和文档层。LDA 的模型结构如图 2 所示。

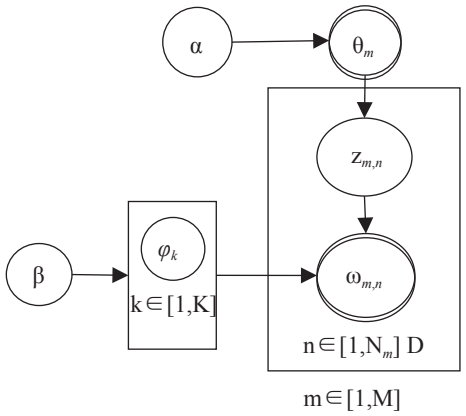

图 2 LDA 模型结构 [26]

在图 2 中，M 代表某一个学者发表的论文总数，N_m 是第 m 篇论文的总词数，K 为主题个数。α 和 β 分别为每篇论文的主题和单词的多项分布的狄利克雷先验参数，每个论文都有一个主题向量 θ，第 m 篇论文的主题向量的概率分布表示为 θ_m。φ_k 代表第 k 个主题下的单词分布变量，θ_m 和 φ_k 都服从 Dirichlet 分布。对第 m 篇论文中的第 n 个单词 $\omega_{m,n}$，生成它对应的主题 $z_{m,n}$，该主题也对应 Multinomial 分布。

LDA 作为一种非监督机器学习技术，能够根据文档生成文档的主题，以及对应的主题词 [29]。本文在对学者和研究兴趣标签对应的文本文档进行预处理之后，使用 LDA 模型，对学者和兴趣标签都进行了表示。主题数分别选取了 100，200，300，400，500，600 六种情况。然后计算学者和兴趣标签的余弦相似度，取出相似度高的前 5 个兴趣标签作为学者的兴趣标签。

（2）Doc2Vec 模型

Doc2Vec 模型由 Quoc Le 和 Tomas Mikolov 在 2014 年提出 [27]。与 LDA 不同，Doc2Vec 是一个双层的神经网络模型，是 Word2vec 模型的

改进模型。Word2Vec 的核心思想就是根据提供的上下文来预测上下文的其他单词。Word2Vec 能够提供高质量的词汇向量。Doc2Vec 的核心思想与 Word2Vec 相似，是通过训练词向量的方法训练句向量。本研究将某一个学者发表的所有论文进行整理合并为一个文本，因此可以将每篇论文作为该文档的一个段落，赋予该段落唯一的 id 标识，每个段落或句子都被映射到向量空间中，每个单词同样映射到向量空间。因此 Doc2Vec 也叫做 paragraph2vec, sentence embeddings。Doc2Vec 模型也采用了两种不同隐藏层技术，分别为：Distributed Memory（DM）和 Distributed Bag of Words（DBOW），如图 3 和图 4 所示。

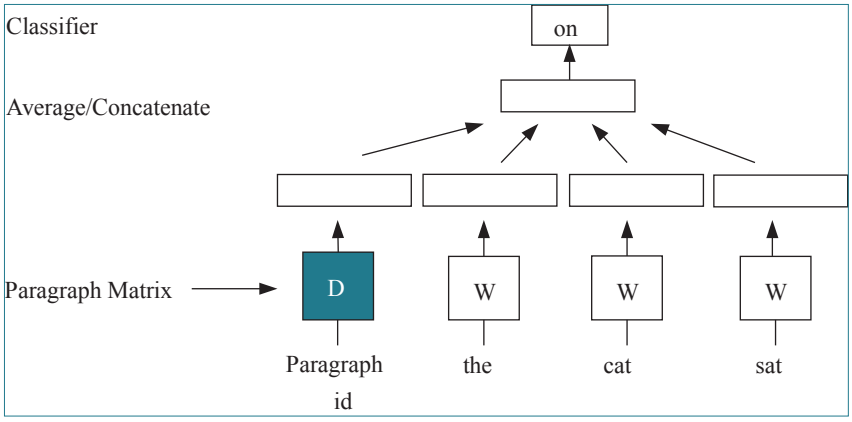

图 3　**DM** 模型结构 [27]

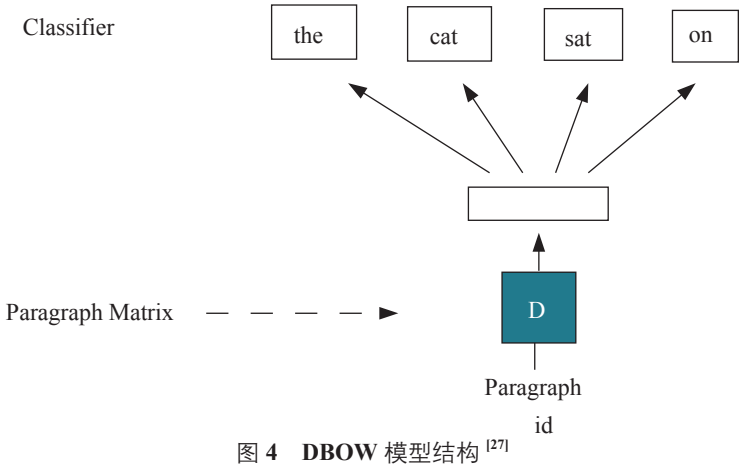

图 4　**DBOW** 模型结构 [27]

用 Doc2Vec 模型训练词向量，方法简单，且考虑了上下文，在大量语料预先训练下，提取出的词向量会越准确。由于大多数任务中，DM 的方法表现很好，因此本文使用了 Doc2Vec 中的 DM 模型对学者和研究兴趣标签文档进行了句子级别的表示。

首先汇总每位学者发表的论文，将论文题目进行连接生成一个文本文档，用来表示学者。然后，将每位学者的兴趣标签看成一个整体标签，汇总每个整体标签对用的学者，并将这些学者所发表的论文题目汇总为一个文本文档用来表示这个整体标签。对整理好的语料采用

doi:10.3772/j.issn.2095-915x.2019.02.003

Doc2Vec 中的 DM 模型进行训练，进行了 50，100，200，300，400，500 六种向量维度尝试，得到每位学者和每个整体标签的向量表示。同样计算对每位学者和整体标签的余弦相似度，取前 3 个相似度最高的结果。

2.2.3 学者和研究兴趣标签相似度计算

本文采用余弦相似度来计算学者和研究兴趣标签之间的相似度。余弦相似度的公式如公式 1 所示。本研究将学者和研究兴趣表示成词向量形式，$X=(x_1, x_2,\ldots, x_n)$，$Y=(y_1, y_2,\ldots, y_n)$[28]。

$$\cos(X,Y)=\frac{\sum_{i=1}^{n}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\times\sqrt{\sum_{i=1}^{n}(y_i)^2}} \quad （公式 1）$$

其中，X 表示学者，Y 表示研究兴趣标签，cos(X,Y) 表示二者之间的相似度。x_i 表示学者词向量中第 i 维的值，同理 y_i 表示研究兴趣标签词向量第 i 维的值（i=1,2,3,…n），n 为词向量的维度。求得的余弦相似度越高，表明学者与兴趣标签之间的相似度越高，标签对学者研究兴趣的表示更好。

本文在研究中对 LDA 和 Doc2Vec 两种方法的结果进行了余弦相似度计算。在 LDA 方法得到的结果中，计算学者和兴趣标签的余弦相似度，取出相似度高的前 5 个兴趣标签作为学者的兴趣标签。最后计算不同主题数下的兴趣标签标注得分。而在 Doc2Vec 方法中，计算对每位学者和整体标签的余弦相似度，取前 3 个相似度最高的结果，并分别赋予权重 0.5,0.3,0.2。每个整体标签中包含三个兴趣标签，因此相当于兴趣标签共有 9 个，其中包含了重复的兴趣标签，将每个兴趣标签出现的频次乘以对应的权重，取得分最高的前 5 个兴趣标签作为学者的兴趣标签。最后，计算学者兴趣标签标注的准确率。

2.2.4 标签集成

对 LDA 和 Doc2Vec 两种方法的结果通过余弦相似度进行计算之后，各得到 5 组学者和最有可能对应的研究兴趣标签。两种方法的结果准确率不同，本文采用集成方法以求得最终的结果。集成方法 (Ensemble Method) 就是综合考虑多个方法得到的结果，以获取更好的效果[29]，本文采用加权投票法对两种结果进行集成，具体方法如图 5 所示。

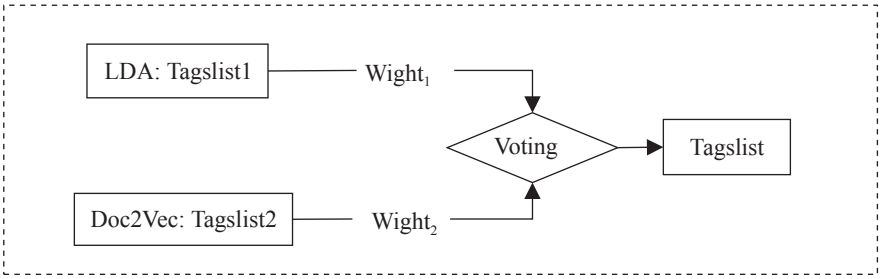

图 5 加权投票法

Tagslist1 和 Tagslist2 分别为两种方法得到的兴趣标签集合，通过代表投票法得到所有标签的得分列表 Tagslist。Tagslist 中每个标签对应得分的计算公式如公式 2 所示，其中，$Weight_n$ 表示每种方法对应的权重，计算方法如公式 3 所示。

$$Rank_{nk} = \sum_{i=1}^{n} T_{nk} * Weight_n \quad （公式 2）$$

$$Weight_n = \frac{P_n}{\sum_{n=1}^{2} P_n} \quad （公式 3）$$

公式 2 中 $Rank_{nk}$ 代表每个兴趣标签的加权得分，用 T 代表兴趣标签，Tagslist={T_{11}, T_{12}, …, T_{1k}; …;T_{n1}, T_{n2}, …, T_{nk}}（k=1,2,…,5，n=1,2），T_{nk}=1，$Weight_n$ 表示第 n 种方法对应的权重。在公式 3 中，权重的确定方法参考 Klein[30]，P_n 代表第 n 个方法的准确度，指标的准确性越高，则权值越高，其作用更明显。

本文通过调整不同方法对应的权重，计算得到每个兴趣标签的最终分值，按照分值大小进行排序，最终选取分值高的前 5 个兴趣标签作为学者的兴趣标签，结果示例如图 6 所示。

	Author	Tag_1	Tag_2	Tag_3	Tag_4	Tag_5
2	Sung-Pil Choi	natural language processi	semantic web	data mining	information retrieval	human computer interaction
3	Uri J. Schild	semantic web	information system	network analysis	information retrieval	knowledge representation
4	Paul Bonsma	algorithms	computational geometr	data structure	combinatorics	planar graph
5	Stephen S. Intille	human computer interacti	ubiquitous computing	natural language process	social interaction	virtual reality
6	Shai Shalev-Shwartz	machine learning	optimization problem	optimization	algorithms	online algorithm
7	Liu Ting	security	data mining	privacy	mobile device	data quality
8	Songyun Duan	data mining	database	query processing	sensor network	semantic web
9	Nabil Aouf	computer vision	image processing	robotics	computer graphics	artificial intelligence
10	María S. Pérez	parallel processing	cloud computing	parallel computing	grid computing	distributed systems
11	Tomasz M. Rutkowski	signal processing	artificial intelligence	brain computer interface	feature extraction	virtual reality
12	Monika Kaczmarek	semantic web	data mining	knowledge managemen	text mining	linked data
13	Sebastien Goasguen	grid computing	cloud computing	distributed systems	operating system	distributed computing
14	Michel dos Santos So	embedded system	system on chip	software process	software engineering	data warehouse
15	Rasmus Larsen	computer vision	image registration	medical image analysis	machine learning	medical imaging

图 6　学者兴趣标签结果样例

2.2.5　结果评估

实验结果采用 SMPCUP2017 学者精准画像构建竞赛中的评估方法，实验结果是为每位学者标注 5 个最适合的兴趣标签，其线上的得分是计算标注的学者兴趣与给定的学者兴趣完全相同的比例，得分计算公式如公式。

$$\text{score} = \frac{1}{N} \sum_{i=1}^{N} \frac{|T_i \cap T_i^*|}{|T_i^*|} \quad （公式 4）$$

其中，N 为需标注的学者个数，T_i 为标注的学者 i 的兴趣标签集合，T_i^* 为给定的学者 i 的兴趣标签集合，这里的 $|T_i^*|$=3。

3　结果分析

3.1　实验数据概述

本文采用 SMPCUP2017 开放学术精准画像大赛提供的数据集。数据集中包含了 10000 名学者以及他们所发表的论文信息，这些学者大部分来自计算机领域。数据集中还包含 789 个标签空间，每个学者都有 3 个对应的研究兴趣标签。本研究将学者和论文信息划分为测试集和训练集，其中训练集包含 6000 名学者，测试集包含 4000 名学者。竞赛要求根据提供的数据信息为学者标注出 5 个最合适的兴趣标签。本文对所有的研究兴趣标签进行频次统计，得到了每个研究兴趣标签的标签比率，即每个标签出现的次数在所有标签出现总次数的占比。统计结果如图 7 所示。

从图 7 中可以发现，训练集的标签空间中，machine learning 的标签比率为 54%，在所有标签出现的总次数中出现超过一半。表明了 machine learning 是计算机学科领域学者关注的热点方向，也是该领域的研究热点。排在前五位的标签还有 data mining，artificial intelligence，computer vision 和 computer

doi:10.3772/j.issn.2095-915x.2019.02.003

science，这表明了计算机领域的学者对数据挖掘和人工智能方向很感兴趣，这两个方向在计算机学科中也是研究的热点。

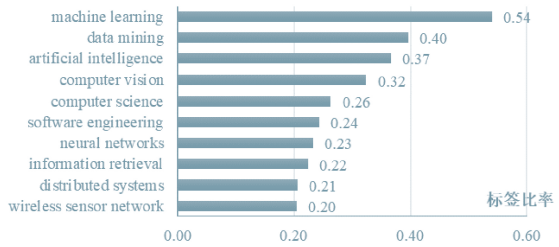

图 7　训练集兴趣标签分布

此外，本文对训练集中的标签进行了共现分析。标签的共现是指在同一个学者的兴趣标签中同时出现的标签。对学者标签的共现分析有助于进行标签之间的相关性分析。图 8 中给出了最常共同出现的共现标签对 Top10。

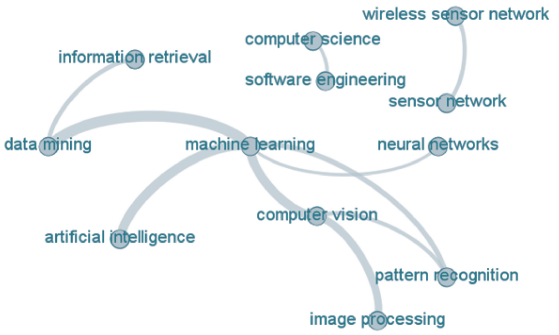

图 8　共现标签对 **Top10**

从图 8 中可以根据节点之间连线的粗细来表示标签共现的频次。其中，machine learning 和 data mining，machine learning 和 computer vision，machine learning 和 artificial intelligence，computer vision 和 image processing 这四组共现标签对的共现情况最明显。前三组共现标签对中都出现了 machine learning，这与 machine learning 在标签中出现的频次最多有关，同时也表明了 machine learning 在计算机学科中属于研究重点和学者的兴趣热点。最后一组共现标签对表明了在计算机学科领域，计算机视觉与图像处理之间具有较强的相关性。

3.2　实验结果分析

3.2.1　训练集的兴趣标签发现实验结果

（1）基于 LDA 方法的兴趣标签标注实验结果

本文基于 LDA 方法对学者和兴趣标签进行表示，通过计算学者和兴趣标签之间的余弦相似度得到学者适合的兴趣标签。研究中尝试了 10，50，100，200，300，400，500 六种不同主题数，针对不同主题数的 LDA 模型，采用十折交叉验证法，分别得到不同主题数下兴趣标签标注效果，如图 9 所示。

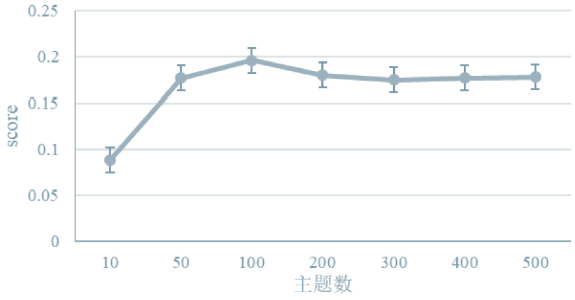

图 9　不同主题数下的兴趣标签标注效果

从图 9 可以看出，主题数为 10 时，标注效果较差，标注得分为 0.088，随着主题数增加，效果不断提升。在主题数为 100 时，得分为 0.196，达到最高，当主题数继续增加时，效果有所下降并趋于稳定。从结果可以得出本文在基于 LDA 建模方法的兴趣标签训练学习过程中，当主题数为 100 时，取得的标注效果最好。因此在下一步构建测试集的兴趣标签标注模型时，采用主题数为 100 的 LDA 模型对学者和兴趣标签分别进行表示。

（2）基于 Doc2Vec 方法的兴趣标签发现实验结果

基于 Doc2Vec 方法对学者和兴趣标签进行表示，通过计算学者和兴趣标签之间的余弦相似度得到学者适合的兴趣标签。在对学者和兴趣标签进行 Doc2Vec 方法建模训练过程中，进行了六种词向量特征维度的尝试，包含 50，100，200，300，400，500 维度。针对不同维度的 Doc2Vec 模型，采用十折交叉验证法，得到了不同词向量特征维度下的兴趣标签标注效果，如图 10 所示。

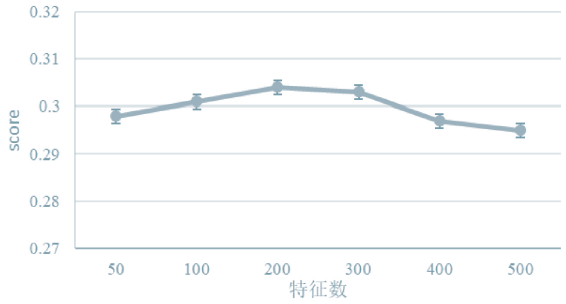

图 **10**　不同特征维度下的兴趣标签标注效果

从图 10 可以看出，标注效果在词向量特征维度为 200 时达到最好，得分为 0.304。维度从 50 到 200 时，标注效果逐渐提高，但当维度从 200 继续增加时效果逐渐下降。分析其原因，则应是随着词向量维度的增多，维度带入的特征噪音也会增加，由此导致了得分下降的结果。根据这一结果，在下一步构建测试集的兴趣标签标注模型时，将采用词向量维度数量为 200 时的 Doc2Vec 模型对学者和兴趣标签分别进行表示。

（3）标签集成

本文采用加权投票法对两种方法进行集成，试图获得更好的效果。本文基于 LDA 和 Doc2Vec 方法得到的标注效果存在差异，它们的标注得分分别是 0.196，0.304，根据标注得分采用公式 3 确定每种方法的权重，分别为 0.392 和 0.608。确定投票权重后，根据公式 2 计算两种方法结果中出现的每个标签分值，得到兴趣标签有序列表，最后取出分值高的前 5 个兴趣标签作为学者兴趣标签。不同模型下的结果对比如图 11 所示。

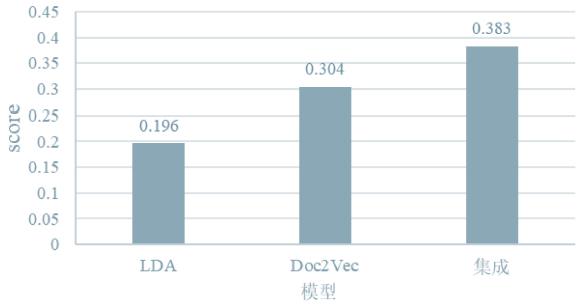

图 **11**　不同模型在训练集中的效果

图 11 中横坐标表示不同模型表示方法，纵坐标表示在训练集上学者兴趣标签标注的得分。从图 11 可以看出，训练集下采用集成方法得到的兴趣标签标注得分为 0.383，获得了更好的标注效果。因此，和 LDA 方法、Doc2Vec 方法相比，使用集成方法的效果有所提升。

3.2.2　测试集的兴趣标签标注实验结果

根据 3.2.1 训练交叉验证的结果，在基于 LDA 建模方法的兴趣标签标注训练过程中，主题数为 100 时，取得的标注效果最好；在基于 Doc2Vec 建模方法的兴趣标签标注训练过程中，词向量维度数量为 200 时，取得的标注训练效果最好。因此对于测试集，将使用主题数为 100 的 LDA 模型、词向量维度为 200 的 Doc2Vec 模型进行学者的兴趣标签发现，并采用集成方法结合两种模型。最终不同模型在测试集上学者兴趣标签标注的得分如图 12 所示。

doi:10.3772/j.issn.2095-915x.2019.02.003

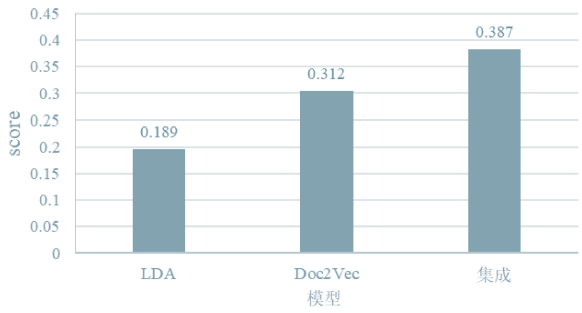

图 **12** 不同模型在训练集中的效果

从图 12 中可以看出 Doc2Vec 模型的效果要好于 LDA 模型，达到了 0.312，但二者都低于集成模型方法效果 0.387。即在本研究的学者兴趣标签发现过程中，结合 LDA 和 Doc2Vec 两种模型的集成方法能够取得最好的效果。

本文的语料中，学者大都来自于计算机领域，学者发表的论文以及其研究兴趣标签主题类别也为计算机，没有特别大的区别，因此在使用 LDA 方法时，学者和兴趣标签文本表示时差异性较小，学者和研究兴趣标签之间的相似性也无较大区别，最终的兴趣标签标注效果比较差。而 Doc2Vec 方法是基于句子维度的语义表达，与 LDA 主题方法相比，提高了向量语义上的准确度，因此在学者的兴趣标签标注过程中能取得较好的效果。基于前两种方法的结果，再采用加权投票进行集成，很好的集合了 LDA 方法和 Doc2Vec 的优点，取得了更优的兴趣标签发现效果。

4 结语

本文利用学术论文信息进行学者研究兴趣标签的标注研究。采用 LDA 和 Doc2Vec 两种不同的文本表示方法，对学者和兴趣标签进行表示；然后利用学者和兴趣标签之间的相似性，即计算学者和兴趣标签之间的余弦相似值，为学者选择出相似值最高的 5 个兴趣标签；接着使用加权投票法融合以上两种方法得到的兴趣标签结果，由此完成学者的兴趣标签标注。通过对比不同建模方法和集成方法下的结果，发现使用集成的方法能获得更好的标注效果。

学者研究兴趣标签发现研究对学者的自我定位和标签自动构建与推荐有重要作用，也是构建用户精准画像的重要环节，有助于加强学术交流合作。本文在测试集中考虑了 LDA 方法表示过程中不同主题数的影响，Doc2Vec 方法表示过程中不同词向量维度的影响，均采取了最优表示。同时采用了集成方法对两种方法优势进行融合，优化了学者研究兴趣标签的标注结果。但本文在研究过程中还存在一些不足之处。在对论文信息处理与使用过程中，本文发现有部分学者存在同名问题，但由于本文的论文数据单一，无法进行同名学者的判断。因此，在未来工作中，可以通过扩展学者信息来解决同名歧义问题。

参考文献

[1] Krestel R, Fankhauser P. Tag recommendation using probabilistic topic models[J]. ECML PKDD Discovery Challenge, 2009: 131.

[2] 袁莎 , 唐杰 , 顾晓韬 . 开放互联网中的学者画像技术综述 [J]. 计算机研究与发展 , 2018, 55(9): 1903-1919.

[3] 黄宣 . 基于开放存取的科研数据获取与专家相似度研究 [D]. 杭州 : 浙江大学 , 2016.

[4] 石光莲 , 杨敏 . 基于 FCA 的 Folksonomy 用户兴趣研究述评 [J]. 现代情报 , 2017, 37(5): 172-177.

[5] 黄镇圣 . 基于 Web 浏览的高校图书馆用户个性化

研究 [J]. 科技信息 , 2009(12): 183-183.

[6] 石豪 , 李红娟 , 赖雯 , 等 . 基于 folksonomy 标签的用户分类研究 [J]. 图书情报工作 , 2011, 55(2): 117-120.

[7] 易明 , 邓卫华 . 基于标签的个性化信息推荐研究综述 [J]. 情报理论与实践 , 2011, 34(3): 126-128.

[8] Weng J, Lim E P, Jiang J, et al. Twitterrank: finding topic-sensitive influential twitterers[C]. Proceedings of the 3rd ACM International Conference on Web Search and Data mining. New York, USA, 2010: 261-270.

[9] Macskassy S A, Michelson M. Why Do People Retweet? Anti-Homophily Wins the Day![C]. Proceedings of the 6th International Conference on Weblogs & Social Media. Catalonia, Spain, 2012: 209-216.

[10] Hung C C, Huang Y C, Hsu J Y, et al. Tag-based user profiling for social media recommendation[C]. Proceedings of the 2008 Workshop on Intelligent Techniques for Web Personalization & Recommender Systems at AAAI. Hawaii, USA, 2008: 49-55.

[11] 项连城 , 桑基韬 , 徐常胜 . 跨社交媒体网络大数据下的用户建模 [J]. 大数据 , 2016, 2(5): 32-42.

[12] Krulwich B. Lifestyle Finder: Intelligent User Profiling Using Large-scale Demographic Data[J]. AI Magazine, 1997, 18(2): 37-45.

[13] 屠守中 , 闫洲 , 卫玲蔚 , 等 . 异构社交网络用户兴趣挖掘方法 [J/OL]. 西安电子科技大学学报 :1-6. [2019-03-27]. http://kns.cnki.net/kcms/detail/61.1076.TN.20181217.1102.002.html.

[14] 范永全 , 刘艳 , 陆园 . 社会化推荐系统的研究进展综述 [J]. 现代计算机 (专业版), 2014(29):29-33.

[15] 张引 . 社会标注系统中标签推荐方法研究 [D]. 沈阳 : 东北大学 , 2012.

[16] Mishne G. Autotag: a collaborative approach to automated tag assignment for weblog posts[C]. Proceedings of the 15th international conference on World Wide Web. ACM, 2006: 953-954.

[17] Sood S, Owsley S, Hammond K J, et al. TagAssist: Automatic Tag Suggestion for Blog Posts[C]. ICWSM. 2007.

[18] Fujimurs S, Fujimurs K,Okuda H.Blogosonomy: Autotagging any text using bioggers'knowledge [C]. IEEE/WIC/ACM International Conference on Web Intelligence, 2007: 205-212.

[19] Tatu M, Srikanth M, D'Silva T. Rsdc'08: Tag recommendations using bookmark content[J]. ECML PKDD discovery challenge, 2008: 96-107.

[20] 陆子龙 . 社交网络中的用户标签推荐 [D]. 哈尔滨 : 哈尔滨工业大学 , 2013.

[21] 王晓耘 , 赵菁 , 徐作宁 . 基于社会化标注的用户兴趣发现及个性化推荐研究 [J]. 现代情报 , 2018, 38(7): 67-73+80.

[22] 吴超 . 在线社会化网络的语义分析和语义社会网的构建 [D]. 杭州 : 浙江大学 , 2010.

[23] 吴小兰 , 章成志 . 结合用户关系网和标签共现网的微博用户标签推荐研究 [J]. 情报学报 , 2015, 34(5):459-465.

[24] 朱雨晗 . 基于用户兴趣标签的混合推荐方法 [J]. 电子制作 , 2018(22):42-44.

[25] 孙赛美 , 林雪琴 , 彭博 , 等 . 一种基于信任度和研究兴趣的学者推荐方法 [J]. 计算机与数字工程 , 2019(3):608-615.

[26] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2012(3): 993-1022.

[27] Le Q, Mikolov T. Distributed representations of sentences and documents[C]. Proceedings of the 31st International Conference on Machine Learning. Beijing, China, 2014:1188-1196.

[28] Salton G, McGill M J. Introduction to modern information retrieval[M]. New York: McGraw Hill Book Company, 1983: 402-403.

[29] Tsoumakas G, Vlahavas I. Random k-labelsets: An ensemble method for multilabel classification[C]. Proceedings of the 18th European conference on machine learning. Berlin, Germany, 2007: 406-417.

[30] Klein D, Toutanova K, Ilhan H T, et al. Combining heterogeneous classifiers for word-sense disambiguation[C]. Proceedings of the ACL-02 workshop on word sense disambiguation: recent successes and future directions-Volume 8. Association for Computational Linguistics. Charlotte, USA, 2002: 74-80.