英文学术资源主题判定策略

黄莉[1,2]，胡青[1,2]，华丽君[1]

（1.武汉科技大学计算机科学与技术学院，湖北 武汉市 430080；2.智能信息处理与实时工业系统湖北省重点实验室，湖北 武汉市 430080）

摘 要：为了方便专业领域研究人员在信息爆炸时代快速、准确查找到与该领域相关的学术资源，提高工作效率，本着简单快捷处理的原则，提出了一个英文学术资源主题判定策略。该策略将不同形式的学术资源转换为统一格式进行处理，采用化整为零的方法计算资源与领域进行相关度匹配。利用样本分析得出合适的判定阈值。并通过判定结果对资源进行自动标签。通过在计算机领域资源的测试，结果表明，该策略能有效的处理英文学术资源的主题判定。

关键词：主题判定；信息抽取；主题相关度；自动标签

Topic Identification Strategy for English Academic Resources

HUANG Li[1,2], HU Qin[1,2] , HUA Lijun[1]

（1. College of Computer Science and Technology, Wuhan University of Science and Technology , Wuhan 430080, China; 2. Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan 430080, China）

Abstract: The network resource grows up exponentially with the rapid development of Internet. It is very important to identify whether the resources belong to the professional field or not. However, resources in professional field are sorted up mostly manually at present, there exist some drawbacks in the process. To tackle with these problems, the Topic Identification Strategy of Academic Resources proposes topic identification method based on the main text of resources. The strategy first transfers different forms of academic resources into a unified format. And then break up the whole text into parts to calculate the topic degree of relevance. Samples are used to decide the value of the thresholds. The proposed method could also add labels for resources automatically. Experiments show that, this strategy could effectively deal with english academic resources topic determination.

Key words: topic identification; information extraction; topic degree of relevance; Automatic labels

1 引言

随着互联网络的迅速发展，人类产生的网络资源呈爆炸式增长。

据估计，70 年代以来全世界每年出版期刊 10 万种以上，发表科技报告约 90 万件，刊登会议文献 10 多万篇，每年发表的科技论文总数近 500 万篇[1]，并呈逐年增长趋势。虽然现在已经出现大量高效且有效的信息检索系统[2]来帮助用户查找信息资源，如何从这些浩如烟海的网络信息资源中找到所需的专业领域学术资源，就成为专业领域文献检索所面临的一个巨大问题。用户迫切需要一个数据

收稿日期：

基金项目：本课题得到国家自然科学基金(61100055)、国家社科基金重大项目(No.11&ZD189)、湖北省自然科学基金(500104)支持

分类细致、精确、全面、更新及时的面向主题的检索系统来获取主题资源信息。

从文本分类的角度研究专业领域的学术资源主题判定技术是目前主题判定系统通常所采用的基本路线。传统的文本分类方法有 kNN、Naïve Bayes、神经网络、SVM、最大嫡模型、回归模型、遗传算法等方法[3]。但是些方法也存在着一些问题，Naïve Bayes 算法[4]需要属性之间完全独立性的理想状态。KNN 算法[5]比较适用于样本容量比较大的类域的自动分类，而那些样本容量较小的类域采用这种算法比较容易产生误分。SVM 算法[6]在大数据集上的训练收敛速度比较慢，需要大量的存储资源和高的计算能力。

上述的方法在判定上可以达到较好的效果，但操作处理比较复杂。为了简化处理流程，提出一种新的英文学术资源判定的策略。该策略目的为网络上多样的学术资源（包括科技文献，个人主页，实验室网站等等）进行统一的转换处理。提供统一处理方法。在相关度计算上提高处理效率，化整为零的进行处理。并为学术文献加注自动标签，方便研究者进行参考。

2 预处理及判定原理

由于网络上学术资源的多样性，为了方便进行统一的判定，先要将资源进行统一的格式转化。同时，不同专业领域特性的一般由该领域的特征词来表征。因此需要在判定前对专业领域进行特征词库的建立。在介绍专业领域主题判定策略之前，需要进行一系列的准备工作，这里称为预处理操作。本节将对这些预处理操作进行简单说明。并简要介绍英文学术资源主题判定策略依据的主要原理。

2.1 统一学术资源格式

学术资源是一个泛指的概念，并不单指科技文献一种。它应该包括网络上所有的与学术相关的资源形式，包括：学术网站，作者个人主页等等一些形式。由于网络上信息庞杂而多样，各种学术资源形式多样，组织形式丰富多彩，这给机器的判定带来了困难。因此在判定之前，需要将所有的学术资源形式进行格式的统一转换。

统一格式的选择应该遵循以下几个原则

1）所有格式都可以等内容进行转换，转换后的形式要与原形式在学术资源内容上一致，不能因为转换而漏掉重要专业内容。

2）所有格式都可以快速转换。因为需要处理的信息量的庞大，需要快速的进行格式的转换，以进行后续判定工作。

最通用的格式是文本格式，所以的其他格式都是在其基础之上，加入相关的标签而得到的，因此，可以将所有的学术资源转换成文本格式。现在有一些工具可以辅助将文件转化为文本格式，例如 Lucene[7]只能解决文本格式的文件，但是如何将各种不同文件格式的文献资源转化为文本文件格式，需要借用 Lius[8]。Lius 是一个强大的解析器，但它不只是一套解析器。它的主要目的是用于更新和搜索 Lucene 的索引。学习 Lius，不仅要学习它的文档解析方法，也要学习它对 Lucene 的封装。Lius 构造了许多新类，在一定程度上简化了编程。借用 Lius 将一般常见的一些文件格式内容转化为了 String 类型，存放在内存缓冲区中。

2.2 基于维基百科的领域特征词库建立

要对专业领域的学术文献进行主题判定，需要建立表征专业领域的特征词词库，或是数字字典。由于各行各业专业领域性强，特征词量大，各专业领域的研究人员只对自己研究领域的特征词了解。现在网络上有很多专业领域的词典，专业性强，包罗广，可以进行采用，但由于其各式各样，没有统一版本，有的属一家之言，没有说服力，这些也给研究人员带来了困难。同时，各种词典只是简单罗列了本领域的专业词条，并没有一个强弱排列之分，这也给判定带来了不稳定的因素。

维基百科[9]是一个基于 wiki 技术的多语言百科全书协作计划，也是一部用不同语言写成的网路百科全书，也被称作“人民的百科全书”。截至 2011 年 11 月，已经有超过 3172 万的注册用户以及为数众多的未注册用户贡献了 282 种语言超过 2024 万篇的条目，其编辑次数已经超过 12 亿 3192 万次。由于是大量人工参与的编辑词条，其具有专业和详尽的特性。并且维基百科的词条并不是一个个孤立存在的，在人为进行编辑的同时，加入了词条与词条之间的联系。

可以利用这些词条和关系，通过转化将这种联系形成领域知识(概念)之间涵盖关系的层次表达，构成某专业领域概念图，从而形成该领域的特征词库的建立。同时，根据不同的层次的词条分配不同的权值。高层次的词条，是该领域的较为核心的关键词，较能够表征该领域的特征。因此处在较高层次上的概念词条，所具有的权重比处在较低层次上

的词条要大。给每个领域词条赋权值为 w。

2.3 学术资源主题判定原理

至此，已经将网络中不同的形式的学术资源转换成统一的文本格式，并利用维基百科建立领域特征词库。学术资源主题判定的预处理的工作已经完成。

学术资源主题判定策略的判定基本原理，是通过计算学术资源和专业主题的相关度进行判定。即是通过分析比较学术资源与专业领域的特征词之前的关联度来判定该学术资源与该专业领域之前的关系。

为了方便阐述，在后续的章节中，采用统一符号化表示方法，如不特别说明，用字母 P 来表示学术资源，特征词用字母 L 来表示，一般词条用 l 表示，特征词权重用字母 w 表示。

3 学术资源主题判定策略

依据上一节所介绍的判定原理，可以得到学术资源主题判定策略处理流程。即在预处理操作完成的基础之上，用特征词提取工具将统一格式的学术资源进行特征词的提取并进行统计，再同过相关度的计算算法将资源特征词与专业领域的特征词进行比较和相关度的计算。将计算得出的相关度值与抽样统计的样本值进行对比，从而进行判定和标签的工作。本节将对里面的关键问题进行详细说明。

3.1 化整为零的相关度计算

学术资源已经转换成统一的文本格式，因此可以采用统一的方法来对学术资源进行特征词的提取。目前已经有很成熟的工具可以将英文的词汇从文本中提取出来并统计词频，在此采用已经有的成熟的工具进行提取。提取出来的词条是组成该学术资源的所有词条，要对词条进行筛选去噪的操作。从所有的词条中，筛选出与主题特征词相似度在一定范围内的词条。

对于学术资源 P_i，通过工具将其中所有的词汇提取出来放入集合 $R_{Pi}=\{l_k\}$。该专业领域的词条集合为 $Q=\{L_t\}$。词条匹配即求 R_{Pi} 与 Q 之间的交集。当然，在文本中，随着语境等情况的不同，词条会以不同的形式出现，因此，在词条匹配时，允许词条相似度在一定较小的范围内浮动。匹配公式如式 1 所示。

$$|1-Simword(l_k,L_t)|\le\delta \tag{1}$$

因为现有的工具还不能将形式不一样的同一词的词频归并，因此，在对特征词统计词频时需要按公式 1 的计算，将相似词条进行词频归并。

通过如上的处理，可以得到学术资源 P_i 中与该专业领域相关的词条的集合 $R_{Pi}\cap Q=\{L_u\}$，以及各词条的词频 f_u，以及各特征词条在该专业领域中的权值 w_u，依据判定原理，可以计算出该学术资源与专业领域之间的相关度，如式 2 所示。

$$Sim(P_i,Q)=\frac{\sum_{L_u\in R_{Pi}\cap Q} f_u * w_u}{\sum_{L_t\in Q} w_t} \tag{2}$$

然而，在很多情况下学术资源所包含的词汇量较多，例如一篇普通的学术论文单词量在 5000 已上，词汇量的增大对于计算的耗时成指数级增长。如果公式 2 方法进行计算，效率不高。因此需要对上述的方法进行改进，提出一种化整为零的办法来提高处理效率。

该方法的思想，不是将学术资源做为一个整体进行计算相关度，而是将其划分成若干小块进行分布式处理和计算，以提高计算的效率，同时小量的计算也提高了计算的精确度。对于每一块的划分，还是采用词条数做为主要的依据。设每一块的词条数为 seg。然后，对这一块中的词条 Seg_{Piv} 进行统计相关度计算 $Simseg$，即计算这一块的专业领域相关度，计算公式将式（2）改为式（3）。

$$Simseg(Seg_{P_{iv}},Q)=\frac{\sum_{L_v\in Seg_{Piv}} f_v * w_v}{\sum_{L_t\in Q} w_t} \tag{3}$$

然后所以块的相关度取平均值，做为整篇学术资源的相关度值，如式 4 所示。

$$Sim(P_i,Q)=\sum_{seg\_num} Simseg(Seg_{piv},Q) \tag{4}$$

通过式 4，可以计算出学术资源与专业领域的相关度的值。为后续的判定提供支撑。

3.2 基于样本学习的阈值设定策略

上一节通过化整为零的方法求出了学术资源与专业领域的相关度，接下来就可以进行主题判定工作了。判定的原理同通常的判定问题一样，提前设定一个阈值 δ，然后将相关度与该值相比较。如果学术资源的相关度值大于阈值，则可以判定该学术资源属于该专业领域。

由判定原理可以分析出，判定准确与否的关键，在于阈值 δ 的选取。如果阈值设定的过大，则

判定结果的准确度会提高，但是会将一些本属于该领域的资源拒之门外。如果阈值设定过小，则更多的非本领域的资源也囊括进来，没有起到原有的目的。因此一个适当的阈值是方法性能好坏的关键。

不同的领域相关的材料，存在着很大的差异，特征词的涵盖量可能不在一个数量级别。因此，不可以对领域的判定进行一概而论。样本学习的方法，可以根据领域资源本身的特点来对判定进行调控。通过对不同的领域资源进行采样分析，得到适合该领域特点的阈值，可以增加判定的灵活性，提高判定效率。

选取某专业领域 Q 相关资源 sam_*num* 篇做为样本进行分析。对每篇样本进行上一节的化整为零的相关度计算方法计算其相关度，绘制相关度曲线，进行分析结果，其结果的平均值做为专业领域 Q 的学术资源相关度阈值 δ_Q，其理论计算公式如式 5 所示。

$$\delta_Q = \sum_{sam\_num} Sim(P_{sam_t}, Q) \Big/ sam\_num \tag{5}$$

式 5 中 P_{samt} 表示第 t 个学术资源样本，$Sim(P_{samt}, Q)$ 表示第 t 个样本与专业领域的相关度值。

通过式 5 求出的阈值，有较强的专业领域依赖性，较为合理的反应出该专业领域的学术资源的相关度平均值。它反应的是个平均水平，应该比临界值，也就是真正的阈值要高。所以，在进行判定时，应该在平均值下滑一个小区间做为真正判定的阈值，则阈值的计算公式应该对式 5 进行改进，得到式 6。

$$\delta_Q = \sum_{sam\_num} Sim(P_{sam_t}, Q) \Big/ sam\_num - \varepsilon \tag{6}$$

式 6 中的 ε 为一个下滑值。ε 的具体取值将由样本分析曲线来获得。具体的求解方法，将在第 4 节进行介绍。

3.3 自动标签

一篇学术资源的关键词或标签是表征该资源内容的核心，一般的研究人员，通过关键词或标签来对资源进行判断，要不要深入进行研读和学习[8]。但是遗憾的是，并不是所有的学术资源都有关键词和标签，并且现在所有的学术资源的标签为作者人工加入。对于大量的学术资源，如果手动加入标签，耗费大量的人力物力和财力。需要研究自动加注标签来辅助研究人员判断。

学术资源的标签，是资源内容的核心，因此在资源中出现的频率也较高。但它是必要条件，不是充分条件。并不是出现频率越高的词的就是该学术资源的标签。但是如果该词条同时又是专业领域的核心词，由于该学术资源属于专业领域，则基本可以认定其为该文献的核心词。

综上分析，对于学术资源的自动标签的选取，应该考虑词频与权重两个因素。由于这两个因素的取值范围不一样，将两个因素进行合并时，为了不忽略掉较小的一方的取值，采用取这两个值的几何平均数来求解该词的关键度 *degree*，如式 7 所示。

$$\deg ree_i = \sqrt{fre_i * w_i} \tag{6}$$

不需要对学术资源中所有的词条进行关键度的计算，只需要对词频高的前 m 项进行关键度计算并进行排序，选择其实中值高的 x（$m>x$）项做为学术资源的标签，供研究者参考。

4 测试与性能评价

本节以计算机领域相关学术资源判定为例，对英文学术资源主题判定策略进行测试及性能评价。

先取一篇计算机领域科技文献一篇进行整体计算和化整为零的策略计算进行比较测试，结果如图 1 所示。

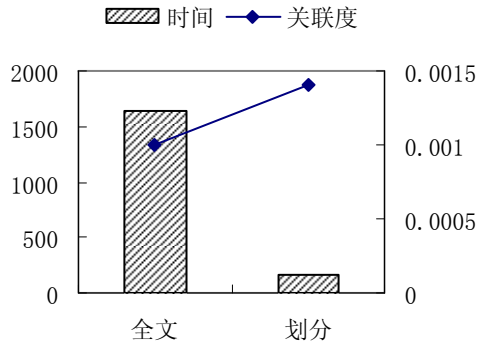

图 1 整体计算与化整为零计算比较

由图 1 结果可以看出，通过化整为零的策略可以大大减小计算的时间，计算精度要高一些。

选取 100 篇计算机相关科技文献进行样本分析，将每篇文献分为 200 个词一块进行计算。由于篇幅有限，图 2 中展示部分结果示意图。其横坐标表示切分块，纵坐标表示关联度的值。不同颜色的线表示不同的学术资源。

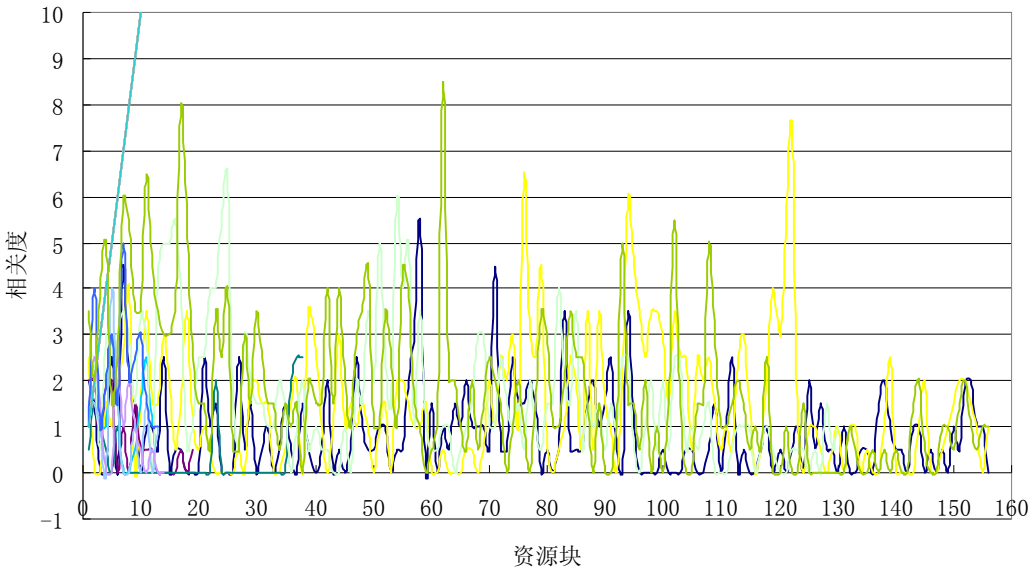

图 2 阈值经验初值的数据分析

通过对式 6 的计算计算机领域的阈值平均值为 δ =0.12。

主题判定的评价指标一般是准确率（Precision）、召回率（Recall）。其中，准确率是正确判定为主题相关的学术资源数占判定为主题相关的学术资源数的百分比，反映的是学术资源主题判定系统的准确性；召回率是正确判定为主题相关的学术资源数占实际上主题相关的学术资源数的百分比，反应了学术资源主题判定系统的完整性。选取学术资源（包括科技文献（PDF 格式、doc 格式），个人主页，实验室网页）200 篇。通过人工处理手段获取标准结果。其最后判定的结果如图 3 所示。

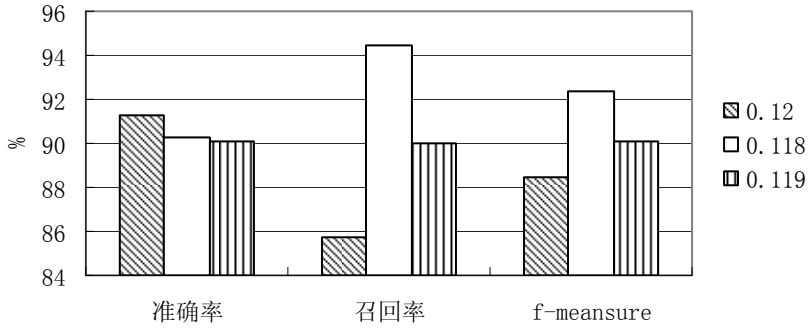

图 3 判定结果分析

图 3 中对不同阈值的取值进行了分析，可以看出当阈值为 0.118 时，性能最高，所以对于公式样 6 中 ε 取值为 0.002。

同时从图 3 中可以看出，提出的主题判定策略可以有效的实现主题判定，其准确率召回率均达到较高水平。

5 总结

随着计算机与网络的不断发展，信息爆炸时代的来临，专业领域的研究人员需要耗费大量的时间和人力在浩如烟海的信息中寻找和鉴别与领域相关的资源。这种人工进行整理工作量大、维护困难、适应性差。急需利用计算机辅助自动判定领域学术资源的方法。而大多数判定方法计算过于复杂，因此，提出了一种简单的学术资源主题判定的策略。

该策略主要针对多样的英文的学术资源进行统一处理。将各种学术资源进行统一格式转换，然后进行与领域特征词进行主题相关度计算从而进行领域判断。领域特征词的选取采用基于维基百科的领域词条建立分层次词典。为了减小工作量，相关度的计算采用化整为零的统计方法进行。并能够根据计算结果，给资源自动加注标签，以方法专业领域研究人员进行检索参考，提高研究效率。

通过对计算机领域的英文学术文献进行实验测试，结果表明该方法可以有效的对英文学术资源进行主题判定。

参考文献

[1] 曹宇容. 浅析科技文献资源状况与检索. 科技情报开发与经济 2011(33):120-123.

[2] Ricardo Baeza-Yates, Berthier Ribeiro-Neto. Modern Information Retrieval. New York: ACM Press, 1999. 1~20.

[3] Yuan Pingpeng, Chen Yuqin, Jin Hai, et al. MSVM-kNN: Combining SVM and k-NN for Multi-Class Text Classification. IEEE International Workshop on Semantic Computing and Systems (WSCS 2008). 2008. 133-140.

[4] Aleksander Kolcz, Wen-tau Yih. Raising the Baseline for High-Precision Text Classifiers. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07). San Jose: 2007. 400~409.

[5] Xiulan Hao, Xiaopeng Tao, Chenghong Zhang, et al. An Effective Method to Improve KNN Text Classifier. In: Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing. 2007. 1379~384.

[6] Xiaoou Li, Jair Cervantes, Wen Yu. Two-Stage SVM Classification for Large Data Sets via Randomly Reducing and Recovering Training Data. In: IEEE International Conference on Systems, Man, and Cybernetics (SMC'07). Montreal: 2007. 3633~3638.

[7] 朱雪莲. 基于Lucene全文检索引擎的应用研究[J]. 微型机与应用. 2010(22): 3-5.

[8] 闫晓妍. 基于语义 Wiki 的知识检索研究. 图书馆学研究. 2010(13): 75-80.

[9] 鞠彦辉,刘闯. 国外典型语义标注平台的比较研究[J]. 现代情报. 2009, 29(01): 215-217.

作者简介：

黄莉、1982、女、讲师/博士、主要研究方向：数据管理及知识发现；
通讯地址：湖北省武汉市青山区钢花 115 街 27 门 9 号 430080
13098882821
E-mail：huangli82@wust.edu.cn