中国科普网站的特征向量研究

吴晨生[1,2]　郭金忠[2]　罗　植[3]　廖　涛[1]

（北京市科学技术情报研究所，北京 100048）[1]
（北京师范大学系统科学学院，北京 100875）[2]
（北京市社会科学院，北京 100101）[3]

[摘　要] 在中国，识别科普网站的内容长期以来主要是依靠专家判断来进行。这种主观判断不仅费时费力，效果也并不好。这其中最主要的一个原因是网站内容比较丰富，人工浏览效率低下，在一定的时间只能处理有限的内容，对于整个网站的判断会存在不全面的地方，也具有主观性。对此问题的解决需要提出一个基于人工智能的可以进行快速定量计算的方法。本文提出的科普网站特征向量就是讲网站内容通过计算机进行处理抽象出来的一个向量空间模型，它能比较好的表现网站的文字内容和意思，可以最终实现机器自动判断网站内容是否含有科普成分以及什么性质的科普内容。

[关键词] 科普网站　特征词频　向量空间

[中图分类号] N4，TP39，O29　　**[文献标识码]** A　　**[文章编号]** 1673-8357 (2013) 05-0043-05

The Characteristic Word Vectors of Chinese Science Websites

Wu Chensheng[1,2]　Guo Jinzhong[2]　Luo Zhi[3]　Liao Tao[1]

(*Beijing Institute of Scientific and Technical Information*, *Beijing 100048*) [1]
(*School of Systems Science*, *Beijing Normal University*, *Beijing 100875*) [2]
(*Beijing Academy of Social Sciences*, *Beijing 100101*) [3]

Abstract: In China, the recognizing whether a website belongs to science websites relies mainly on expert judgment to proceed. This kind of subjective judgment is not only time-consuming, and the results are not reliable. Browsing and judging by experts have low efficient because the rich website content. They only can process very limited part of any website under certain time and energy. Besides this, different people may make different judgments. It is necessary to propose a quantitative method based on machine intelligence. This paper will discuss the feature word vectors of Chinese popular science websites what is processed by computer abstracted from real content based on vector space model. We think it can better the performance of the site's textual content and meaning. Based on this method, people may make a system to automatically

收稿日期：2013-07-05
作者简介：吴晨生，北京市科学技术情报研究所副所长，研究员，Email：wu1082@163.com；
郭金忠，北京师范大学系统科学学院博士研究生，Email：guojinzhong123@163.com；
罗　植，北京市社科院与清华大学联合培养博士后，Email：establown@gmail.com；
廖　涛，北京市科学技术情报研究所声像部主任，Email：liaotao@163.com。

determine the ultimate realization of website content if it contains science ingredients as well as what kind of science content.

Keywords: popular science website; characteristic word frequency; vector space

CLC Numbers: N4, TP39, O29 **Document Code**: A **Article ID**: 1673-8357 (2013) 05-0043-05

1 研究背景

在中国，识别科普网站的内容长期以来主要是依靠专家判断来进行。这种主观判断不仅费时费力，效果也并不好。这其中最主要的一个原因是网站内容比较丰富，人工浏览会存在不全面的地方。此外，人工判断的主观性也不能避免。解决此问题的唯一方法是发展合适的智能系统，基于语义分析等先进方法和手段，建立适当的数学模型进行定量客观的分析[1]。

目前，对于中文文字的自动智能分析主要基于词频分析和向量空间模型来进行[2-4]。

汉语分词(Chinese word segmentation)是一个基本而重要的环节，指的是将一个汉语句子切分成一个一个单独的词。这在汉语的处理上具有特殊的困难性。我们知道，在英文的行文中，单词之间是以空格作为自然分界符的，而中文只是字、句和段能通过明显的分界符来简单划界，唯独词没有一个形式上的分界符，句子的分解是要通过对意义的理解切分的，人工的切分最为准确但难以处理大量的文本，而机器的切分需要复杂的算法。目前，不同学者们提出了许多不同的汉语分词方法[5-9]。

向量空间模型（vector space model）是由Salton 等人于 20 世纪 60 年代末提出，是一种简便、高效的文本表示模型，其理论基础是代数学[10]。向量空间模型把用户的查询要求和数据库文档信息表示成由检索项构成的向量空间中的点，通过计算向量之间的距离来判定文档和查询之间的相似程度。然后，根据相似程度排列查询结果。向量空间模型的关键在于特征向量的选取特征向量的权值计算两个部分，对其的改进和一直在进行并在不同领域得到应用[2, 11-13]。

本文中，我们将词频分词结果化为向量空间模型，利用这个方法对 55 个中国科普网站进行分析研究。首先，我们采用中科院计算所的 ICTCLAS 模块（http://ictclas.org）进行直接分词，进一步我们利用门户网站的词频结果作为基底对初始的科普网站词频分析结果进行进一步缩减，得到更加简洁但更具明显效果的特征词频。这个可以为科普网站的自动判别系统打好基础。

2 词频分析

我们采用中科院计算所提供的 ICTCLAS 分词系统。据其官方网站报道，该系统经过国内和国际权威的公开评测、五万客户的认可（ICTCLAS 在国内 973 专家组组织的评测活动中获得了第一名，在第一届国际中文处理研究机构 SigHan 组织的评测中都获得了多项第一名。其综合性能卓越）。我们选择了 55 个网站（包含科普和一些门户网站）。如表 1 所示，是其中 6 个网站的一些高频词分析结果，它们是：生物科普网（www.103edu.cn）、谈天天文网（www.2-sky.com）、中国动物信息网（www.animal.net.cn）、中国法医学会（www.fyxh.org）、中国环保网（www.ep.net.cn）、低碳网（www.eedu.org.cn）。

可以看出，分析出来的高频词明显地表现出网站在内容上的差异。通过这些词就可以大致地判断这个网站比较多地讨论了哪一方面的知识。比如看到“生物”、“科普”、“研究”等词语，我们就能判断出这是一个关于生物知识的网站；看到“观测”、“天文”、“望远镜”、“火星”、“太空”等词，我们就知道这是一个关于天文的网站；看到“动物”、“夜蛾”、“翅”、“蝶”、“物种”等词，我们就知道这是一个关于动物的网站；看到“鉴定”、“司法”、“法医学”、“质量”、“标准”、“检测”等，我们就知道这是一个有关法医的网站；看到“环保”、“环境”、“保护”、“处理”、“污染”、“水”等词语，我们就知道这是一个有关环保的网站；而看到“碳”、“经济”、“发展”、“节能”等词，我们就知道这是一个有关低碳的科普网站。

表 1　科普网站的部分词频列表

名称	生物科普网			谈天天文网			中国动物信息网		
排名	词	个数	频率	词	个数	频率	词	个数	频率
1	的	169 139	0.057	的	35 102	0.055	属	944 985	0.056
2	是	33 375	0.011	在	7 846	0.012	数据库	653 249	0.039
3	在	33 050	0.011	是	6 800	0.011	中国	624 744	0.037
4	生物	31 888	0.011	一	5 036	0.008	动物	585 760	0.035
5	一	26 780	0.009	了	4 631	0.007	夜蛾	578 848	0.034
6	了	22 900	0.008	页	4 148	0.007	中国科学院	471 993	0.028
7	和	21 849	0.007	观测	4 066	0.006	目	415 031	0.025
8	评论	19 232	0.007	和	3 962	0.006	点	388 242	0.023
9	中	15 388	0.005	天文	3 834	0.006	研究所	366 010	0.022
10	科普	14 704	0.005	望远镜	3 825	0.006	源	361 111	0.021
11	有	14 305	0.005	有	3 358	0.005	数据	361 079	0.021
12	与	13 114	0.004	文章	3 358	0.005	科	343 678	0.02
13	不	13 091	0.004	年	3 183	0.005	生物	290 461	0.017

名称	中国法医学会			中国环保网			低碳网		
排名	词	个数	频率	词	个数	频率	词	个数	频率
1	的	24 399	0.058	的	86998	0.042	的	149 402	0.05
2	和	8 500	0.02	环保	25116	0.012	碳	124 471	0.041
3	中国	3 934	0.009	和	17154	0.008	低	89 584	0.03
4	或	3 286	0.008	环境	17126	0.008	经济	33 474	0.011
5	鉴定	3 255	0.008	是	16142	0.008	是	30 896	0.01
6	在	3 228	0.008	在	14099	0.007	发展	30 535	0.01
7	应	3 035	0.007	一	12268	0.006	在	29 601	0.01
8	法医学	2 918	0.007	了	9490	0.005	和	29 370	0.01
9	对	2 835	0.007	信息	8844	0.004	一	24 923	0.008
10	机构	2 781	0.007	不	8487	0.004	能源	21 682	0.007
11	检查	2 637	0.006	有	8477	0.004	中国	19 952	0.007
12	会	2 176	0.005	等	8042	0.004	技术	19 139	0.006
13	司法	2 021	0.005	为	8014	0.004	了	19 097	0.006

3 特征词频向量

但是，不管是什么专业网站，实际上都含有普通语言的成分，这部分词实际上会对科普网站本身的内容有一个掩盖和干扰作用。基于此，我们尝试从初步获得的词频中去掉普通网站的常用词部分然后进行分析，然后关注剩下的高频词部分。具体做法是，我们先对两个综合门户网站（网易和新浪网）进行词频分析，然后将这两个网站词频排名的前 3 000 的共同部分拿出来作为基底词库，如“的”，“中国”，“我们”等，显然这个词库不具有任意专业特征。然后分别对科普网站进行词频统计，取出排名前 1 000 的词，排除存在于基底词库的词，显然剩下的词会具有鲜明的专业特征。我们将其作为该科普网站的特征词库，具体的词频分布作为网站的特征向量。

这样的处理相当于对科普网站的词语进行提纯，就是把所有网站共有的部分词语（通用词语）去除掉，剩余词语完全可以更好地表征这一个网站的信息。这样处理完后可以很明显地看出，以上分别代表生物、天文、动物、法医学、环保和低碳六个领域的科普网站。这些保留下来的词成为该网站的特征词汇。

此外，我们知道一个科普网站越具有科技方面的专业性，就会大量使用专业词汇，那么按上面的方法被去除的词就较少，保留的特征词就越多；反之，如果没有太多的专业内容，保留的词就较少。我们定义一个网站的最后特征词语数量为该网站的特征向量长度，如表 3 所示。科普网站特征向量长度实际上反映了科普网站的专业化程度，如作为极端情况，第 13 号网站测试时实际上已改变为专业网站，因此特征向量长度很长。如表中 3 号昆明大学数字博物馆网站和 13 号中国

表 2 一些科普网站的特征词语和频数（部分）

名称	生物科普网			谈天天文网			中国动物信息网		
排名	词	个数	频率	词	个数	频率	词	个数	频率
1	科普	14 704	0.005	观测	4 066	0.0064	属	944 985	0.056
2	粤	7 083	0.0024	天文	3 834	0.006	数据库	653 249	0.0387
3	站长	7 077	0.0024	望远镜	3 825	0.006	夜蛾	578 848	0.0343
4	文章中	6 299	0.0021	火星	2 692	0.0042	中国科学院	471 993	0.0279
5	破坏	5 614	0.0019	科学家	2 292	0.0036	研究所	366 010	0.0217
6	生态	5 531	0.0019	地球	2 019	0.0032	源	361 111	0.0214
7	植物	3 828	0.0013	太空	1 993	0.0031	多样性	253 396	0.015
8	灵	3 683	0.0013	月球	1 984	0.0031	蛾	239 231	0.0142
9	奥秘	3 623	0.0012	颗	1 667	0.0026	尺	115 314	0.0068
10	留言簿	3 578	0.0012	行星	1 601	0.0025	鱼	114 926	0.0068
11	模板	3 564	0.0012	太阳	1 561	0.0025	形	109 362	0.0065
12	影音	3 560	0.0012	站长	1 551	0.0024	翅	97 864	0.0058
13	仙子	3 549	0.0012	彗星	1 442	0.0023	物种	97 667	0.0058

名称	中国法医学会			中国环保网			低碳网		
排名	词	个数	频率	词	个数	频率	词	个数	频率
1	鉴定	3 255	0.0077	供求	4 484	0.0022	生态	13 518	0.0045
2	法医学	2 918	0.0069	剂	3 273	0.0016	排放	13 020	0.0043
3	司法	2 021	0.0048	生态	3 262	0.0016	打印	5 942	0.002
4	法医	1 922	0.0045	仪	3 184	0.0015	二氧化碳	5 919	0.002
5	评审	1 793	0.0042	时空	2 970	0.0014	窗口	4 430	0.0015
6	认可	1 661	0.0039	污水	2 918	0.0014	转变	3 352	0.0011
7	实验室	1 595	0.0038	清洁	2 558	0.0012	太阳能	2 760	0.0009
8	校准	1 539	0.0036	废	2 454	0.0012	气体	2 702	0.0009
9	检测	1 479	0.0035	大全	2 445	0.0012	实践	2 690	0.0009
10	会员	1 333	0.0031	废水	2 337	0.0011	温室	2 493	0.0008
11	认证	1 147	0.0027	材	2 312	0.0011	清洁	2 454	0.0008
12	科协	1 069	0.0025	投诉	2 206	0.0011	排放量	2 238	0.0007
13	王京海	1 012	0.0024	排放	2 159	0.001	循环	2 173	0.0007

动物信息网，说明他们具有非常明显的专业特征。同时，向量长度较短的说明和门户网站差异不大，如 34 号光明网科技频道和 53 号网易网科技频道。

4 结论和展望

科普网站的特征词频向量的提出很好地体现出了科普网站内容，其长度也可以很好地说明该科普网站和普通门户网站的差异，可以用来描述了科普网站的专业化程度。进一步，我们将讨论这个特征词频的差异以表示这两个网站的相同点与不同点。这些概念为网络科普的层次性的定量描述和可以用来进行层次性定量分析的智能系统设计打下了基础。

表 3 科普网站特征向量长度

网站编号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
长度	198	580	266	164	265	188	309	281	238	329	232	277	684	239	315
网站编号	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
长度	226	140	230	237	262	210	181	320	254	191	307	212	295	427	329
网站编号	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
长度	345	261	181	128	232	294	403	297	403	237	298	247	377	400	281
网站编号	46	47	48	49	50	51	52	53	54	55					
长度	290	288	193	281	356	194	236	120	302	0					

（下转第 88 页）

总而言之，《中学生》在推动我国科普事业方面做出了巨大贡献，使中国从落后走向先进迈出了巨大一步，成为推动中国科学事业发展史上不可磨灭的一笔。而在媒介融合趋势不断加强的今天，随着媒体竞争日益激烈，科普期刊要站稳自己的地盘，可从《中学生》杂志中得到以下启示：一是定位要清晰。针对不同的受众群体，应对其教育程度、兴趣爱好等进行分析，在符合社会主流价值观、注重科学性的前提下，“投其所好”，选对题材，选准角度。二是内容要出色。由于科技内容与其他相比，更具科学性和专业性，要写出特色并不容易。科普期刊可结合时事热点，利用图文结合形式，用贴近生活的语言解读事件背后的科学原理，传递科学知识、科学方法、科学态度和科学精神。三是要注重互动。随着科学技术的进步、媒介技术的不断发展，受众已不再是简单地“被动接受”。因此，科普期刊要进一步加强和树立读者意识，“纸上纸下”增强互动，不断扩大影响力。具体来说，“纸上”可利用有奖问答、有奖征文、趣味益智游戏等形式增强与受众的互动；而“纸下”则可邀请科普作家、相关领域的专家开展形式多样的讲座，组织户外科学考察、科学实验等活动，加强读者的忠诚度和信任感，做一个有社会责任感的权威科学媒体。

致谢 本文在成稿过程中得到了湖南大学新闻传播与影视艺术学院李浩鸣教授和陶贤都副教授的指导和帮助，在此深表感谢！

参考文献

[1] 编辑部. 发刊辞[J]. 中学生，1930年创刊号：1.
[2] 本志同人. 谈谈本志的旨趣[J]. 中学生，1947(7)：4.
[3] 编辑部. 本志复刊四周年[J]. 中学生，1943(4)：4.
[4] 佚名. 怎样解决饭厅风潮[J]. 中学生，1931(5)：22.
[5] 学人译. 以眼还眼[J]. 中学生，1944(6)：45.
[6] 罗家琅. 今日的钨[J]. 中学生，1940(2)：10.
[7] 编辑部. 编辑室[J]. 中学生，1946(3)：13.
[8] 小有. 电石是什么[J]. 中学生，1930(8)：68.
[9] 桑洛卿. 研究地理之兴趣[J]. 中学生，1930(5)：29.
[10] 陈原. 书林漫步[M]. 北京：三联书店，1998：45.

（责任编辑 谢小军）

（上接第46页）

参考文献

[1] Russell.S and Norvig.P. Norvig. Artifical Intelligence：A Modern Approach (3rd Edition)，Englewood Cliffs，NJ：Prentice-Hall，2009.
[2] 庞剑锋，卜东波，白硕. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究，2001，18(9)：23-26.
[3] 张海燕. 基于分词的中文文本自动分类研究与实现[D]. 湖南大学硕士学位论文，2002.
[4] 赵彦斌，李庆华. 汉字关联性量化方法及其在文本相似性分析中的应用[J]. 计算机应用，2006，26(6)：1396-1397.
[5] 丁丰，董娜. 自然语言处理系统中自动分词的研究[J]. 北方交通大学学报，1999，23(6)：31-33.
[6] 王瑞雷，栾静，潘晓花，卢修配. 一种改进的中文分词正向最大匹配算法[J]. 计算机应用与软件，2011，28(3)：195-197.
[7] 郑晓刚，韩立新，白书奎，曾晓勤. 一种组合型中文分词方法[J]. 计算机应用与软件，2012(7)：26-28.
[8] 唐籍涛，李飞，郭昌松. 网络舆情监控中新词识别问题的研究[J]. 计算机技术与发展，2012(1)：119-121.
[9] 黄昌宁，赵海. 中文分词十年回顾[J]. 中文信息学报，2007，21(3)：8-19.
[10] 吴扬扬，陈怀南. 基于关联规则的通信网络告警相关性分析模型[J]. 通讯和计算机，2004，1(1)：57-69.
[11] 宋丹，王卫东，陈英. 基于改进向量空间模型的话题识别与跟踪[J]. 计算机技术与发展，2006，16(9)：62-67.
[12] 常娥，张长秀，侯汉清，惠富平. 基于向量空间模型的古汉语词义自动消歧研究[J]. 图书情报工作，2013，57(2)：114-118.
[13] 徐明子，吕立，李喜旺. 改进空间向量模型主题网络爬虫系统[J]. 计算机系统应用，2013，22(7)：36-39.

（责任编辑 谢小军）