

中文消费者健康词的发现方法研究^{*}

张兴厅 文 栋

(1 北京大学第三医院 北京 100191 2 北京大学医学信息学中心 北京 100191)

雷健波

(1 北京大学医学信息学中心 北京 100191 2 西南医科大学医学信息与工程学院 泸州 646000)

〔摘要〕 探索一种中文消费者健康词的发现方法, 阐述其构建过程, 验证利用 Word2vec 进行消费者健康词表构建的可行性和合理性, 为开发完整的中文消费者健康词表奠定基础。

〔关键词〕 深度学习; 机器学习; 文本挖掘; Word2vec; 消费者健康词表

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2017.05.001

Research on Discovery Method of Chinese Consumer Health Words ZHANG Xing-ting, WEN Dong, 1Peking University Third Hospital, Beijing 100191, 2Center for Medical Informatics, Peking University, Beijing 100191, China; LEI Jian-bo, 1Center for Medical Informatics, Peking University, Beijing 100191, 2School of Medical Informatics and Engineering, Southwest Medical University, Liuzhou 646000, China

〔Abstract〕 The paper explores a discovery method of Chinese consumer health words, elaborates its establishment process, verifies the feasibility and reasonability of establishing the consumer health vocabulary with Word2vec, and lays a foundation for developing the complete Chinese consumer health vocabulary.

〔Keywords〕 Deep learning; Machine learning; Text mining; Word2vec; Consumer health vocabulary

1 引言

随着当代医疗保健环境的形成以及以疾病为中心的传统医疗模式向以健康为中心、以人为中心的

健康管理模式的转变, 非医学专业人士逐渐成为医学信息服务的消费者, 消费者健康信息学作为医学信息学的分支得以产生和发展。根据“2015 年中国网民科普搜索需求行为分析报告”^[1], 网民最关注的科普主题为健康与医疗, 在所有搜索主题中占比 55.15%。然而由于医学知识的复杂性, 以及消费者健康用语与医生专业词汇存在的差异, 消费者对医学健康信息的理解和利用问题成为消费者健康信息学面临的重大挑战^[2-3]。对健康信息的获取、分析和理解水平的高低会影响消费者做出合适的健康

〔收稿日期〕 2017-03-31

〔作者简介〕 张兴厅, 硕士研究生; 文栋, 硕士研究生; 通讯作者: 雷健波, 副教授, 硕士生导师。

〔基金项目〕 国家自然科学基金项目(项目编号: 81471756)。

决策, 低的健康信息素养会导致不良的健康结局。在日常的医患面对面沟通中, 患者遇到不理解的术语, 医生会及时给予解释和说明, 对于重要的信息会反复强调; 但是在其他健康信息环境中, 如论坛、报纸、新媒体等, 信息是静态的, 此时信息能否准确传达就要看消费者自身的知识储备以及不同作者的表达方式了。基于这样的现状, 研究者开始关注如何构建面向消费者的健康知识库, 知识库组织系统包括多种形式, 如术语集、字典、分类、本体等。这些形式的知识组织一般由许多清晰表达语义概念的对象构成, 标准的医学术语集如 SNOMED-CT、ICD 都是用来在不同信息系统间交换的标准术语; 但是由于这些类型的术语集是专业术语的集合, 并未包括消费者的习惯表达方式, 所以构建消费者健康词表(Consumer Health Vocabulary, CHV)显得尤为重要^[4]。消费者健康词表是指能提供消费者乐于使用、识别、理解的非专业词汇知识库或者词典, 其目的就是解决消费者理解专业术语困难这一问题。本研究旨在探索中文消费者健康词表的构建方法, 以期缩小专业术语和消费者用词之间的鸿沟, 帮助消费者更好地阅读和理解健康相关知识。

2 构建方法

运用 Scrapy 爬虫工具, 抓取网络消费者健康论坛——天涯医院的信息作为语料, 使用结巴分词对原始语料进行中文分词处理, 利用 Word2vec 构造词向量模型, 以 ICD-10 的中文版疾病名称作为领域的种子词汇, 生成候选词列表及其对应的余弦距离值, 通过人工审核选出候选词表中的目标词汇纳入消费者健康词表。最后, 使用预测准确率、准确率和召回率 3 个评价指标对词向量的模型效果进行评估, 引入 TF-IDF 方法对消费者健康词表进行扩充。整体技术路线, 见图 1。

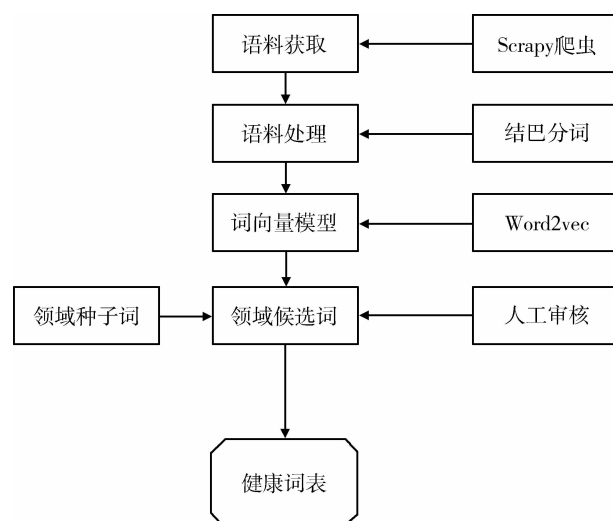


图 1 技术路线

3 构建过程

3.1 数据来源

天涯社区创建于 1999 年 2 月 14 日, 是一个具有全球影响力的网络社区, 自创办以来以其开放、包容、充满人文关怀的特色受到全球华人网民的推崇, 经过 10 多年的发展已成为最具影响力的全球华人网上家园, 2015 年中国 Alex 排名第 11 位。如今天涯社区覆盖的品质用户超过 2 亿, 注册用户超过 8 500 万, 单个帖子访问量甚至达到了 1.25 亿人次, 回复量超过 120 万^[5], 其积累的数据量级是其他传统和网络媒体无法企及的, 而且该社区数据开放、抓取方便, 也为文本挖掘提供了不可多得的宝贵语料。本研究选取天涯社区的医疗健康版块——天涯医院为语料源, 该版块经过长达 18 年的积累, 拥有大量中国消费者关于医疗健康话题发起的各种各样的讨论帖, 按种类不同可分为内科、外科、妇产科、儿科、检验科等 21 个类别。

3.2 数据处理

由于抓取到的语料是中文语料, 中文不像英文那样有空格作为词与词之间的标志, 这就给自然语

言处理带来了困难，所以在生成词向量模型之前必须进行分词。当前已经开发出了多种用于中文分词的工具，如中国科学院计算所汉语词法分析系统 ICTCLAS、哈尔滨工业大学 LTP、东北大学的 NIU Parser、盘古分词、结巴分词^[6-8]等，鉴于本研究的技术实现大多基于 Python，且结巴分词代码清晰、扩展性较好，因此选择用 Python 编写的结巴分词处理语料。

3.3 构建词向量模型

Word2vec 是 Goolge 在 2013 年开源的一款自然语言处理工具，将处理过的大量文本语料输入 Word2vec，其使用深度学习的连续词袋（Continuous Bag-of-words，CBOW）模型和 Skip-gram 模型计算词向量，输出的是每个词对应的高维空间上

的一个向量，训练好的词向量结果可以用于自然语言处理的多个方面，如词性分析、同类词聚类、同义词分析等。经过分词处理后的语料可以作为 Word2vec 的输入文本进行模型训练，从而得到一个词向量模型。Word2vec 不仅有 CBOW 和 Skip-gram 两种模型来实现词语的向量化表示，而且还有层次（Hierarchy SoftMax，HS）和负采样（Negative Sampling，NS）两种算法来进行效率优化，4 者两两组合就可以得到 4 种词向量训练模式^[9]，见表 1。除了训练模式，Word2vec 的词向量训练需要设置很多参数，参数解释^[10]总结，见表 2。

表 1 Word2vec 的 4 种训练模式

训练模式	CBOW	Skip-gram
Hierarchy SoftMax	CBOW&HS	Skip-gram&HS
Negative Sampling	CBOW&NS	Skip-gram&NS

表 2 Word2vec 训练参数及各参数的说明

参数		说明	参数		说明
Train	输入训练文件		Sample	采样的阈值	
Output	训练结果文件		Binary	是否用二进制存储	
Window	窗口大小		Alpha	学习速率	
CBOW	0 表示使用 Skip-gram 模型，1 表示使用 CBOW 模型		Min-count	设置最低频率	
Size	表示输出的向量维度		Class	表示词聚类的个数	
HS	0 表示使用 Negative Sampling 算法，1 表示使用 Hierarchy Softmax 算法		Workers	并行线程数	
Negative	Negative Sampling 算法中负采样样本的数量		-	-	

本研究使用 Word2vec 软件训练词向量模型，基于 Chiu 等^[11]以及 Pyysalo 等^[12]的研究结果设置 Word2vec 的参数，最终选择 Skip-gram 模型，优化算法使用 Hierarchy Softmax，窗口值为 5，输出维度为 200，阈值采用 1e-4，用二进制形式存储。训练结束后，得到一个词向量模型 tianya. bin，该语料共包含 101 964 个词语，34 159 800 个字。

3.4 种子词选择

对于疾病名称的种子词，本研究以 ICD-10 的术语集为来源，将术语集列表中的疾病名词与语料库中的文本进行匹配，然后根据疾病名词在语料中的出现频次排序，共匹配到 516 个词语，舍弃掉出现频次小于 50 的名词，最终剩余 303 个疾病名词。

3.5 获取消费者表达方式

将所有种子词在 Word2vec 模型中运行 distance. exe 程序，种子词的候选词将按照余弦距离进行排列，如输入“腹泻”，则输出结果见表 3，可以看出，“拉肚子”恰恰就是消费者在生活中的惯用表达，“腹泄”是消费者输入的同义错别词，而“呕吐”、“腹胀”则是种子词“腹泻”的同类词，不是消费者用词，说明词向量将同类词和“同义词”全都识别出来而并没有加以区分。所以本研究需要通过人工审核方式将种子词的消费者用词全部筛选出来。取候选词列表的前 10 个词，将 303 个疾病名种子词按科室分成几类，如消化科、皮肤性病科等。表 4、表 5 是经过人工审核后消化科和皮肤

性别科疾病名词的健康词表，频率表示种子词在语料中出现的次数，Cosine distance 表示健康词和种子词之间的余弦距离。

表 3 输入“腹泻”后模型生成的
同义词列表及其余弦值距离

词语	余弦距离
呕吐	0.770 379
腹胀	0.759 555
拉肚子	0.742 035
腹痛	0.722 890
便秘	0.722 803
消化不良	0.721 490
腹泻	0.711 614
厌食	0.707 516
肠鸣	0.701 677
返酸	0.700 117

表 4 消化科疾病名词的健康词表

种子词	频率	健康词（同义词）	余弦距离
腹泻	2 519	拉肚子	0.726 909
		腹泄	0.683 475
腹痛	2 102	腹胀痛	0.750 704
		肚痛	0.681 580
胃溃疡	2 006	溃疡病	0.617 287
胃痛	1 031	胃疼	0.719 431
消化不良	668	食欲不振	0.861 641
食管炎	283	食道炎	0.740 845
脾大	147	脾肿大	0.756 725
胃酸过多	134	泛酸	0.759 513
		反酸	0.714 074

表 5 皮肤科疾病名词的健康词表

种子词	频率	健康词（同义词）	余弦距离
早泄	5 431	早泻	0.695 738
脱发	2 216	脂脱	0.596 243
		掉头发	0.574 252
白癜风	1 648	白斑	0.676 770
疣	1 265	疣体	0.723 342
黄疸	1 209	黄胆	0.661 823
银屑病	965	牛皮癣	0.772 418
		鱼鳞病	0.634 020

续表 5

痤疮	910	Psoriasis	0.632 331
		粉刺	0.747 065
		青春痘	0.726 131
艾滋病	768	HIV	0.667 679
		艾滋	0.645 393
		爱滋病	0.635 955
皮疹	617	疹子	0.650 239
		疤痕	0.718 023
		泛红	0.737 449
白发	415	少白头	0.662 090
		白头发	0.634 320
粉刺	387	青春痘	0.765 382
		痤疮	0.747 065
腋臭	363	狐臭	0.650 369
变态反应	163	超敏反应	0.721 075
脚癣	162	足癣	0.793 234
		香港脚	0.765 194
斑秃	134	鬼剃头	0.653 145
花斑癣	52	汗斑	0.657 865

3.6 结果对比

为了验证本方法的有效性，本研究从百度百科里找到的种子词同义词作为对比的金标准。根据百度百科内容中的标志词，如“简称”、“别称”、“俗称”、“又称为”、“又称”、“也称”等，将上面的消化科和皮肤性别科的种子词在百度百科里寻找它们的同义词，两种方法的效果对比，见表 6。

表 6 本研究和通过百度百科找到的健康词结果对比

种子词	本研究	百度百科
腹泻	拉肚子、腹泄	拉肚子
腹痛	腹胀痛、肚痛	—
胃溃疡	溃疡病	—
胃痛	胃疼	胃脘痛
消化不良	食欲不振	功能性消化不良
食管炎	食道炎	食道炎
脾大	脾肿大	—
胃酸过多	泛酸、反酸	胃酸过高
早泄	早泻	—
白癜风	白斑	白斑风、皮肤白斑

续表 6

脱发	掉头发、秃顶	秃头、鬼剃头
疣	疣体	-
黄疸	黄胆	显性黄疸
银屑病	牛皮癣、鱼鳞病、Psoriasis	牛皮癣
痤疮	粉刺、青春痘	粉刺、青春痘、壮疙瘩
艾滋病	HIV、艾滋、爱滋病	获得性免疫缺陷综合征
皮疹	疹子	-
瘢痕	疤痕	疤痕
发红	泛红	-
白发	少白头、白头发	-
粉刺	青春痘、痤疮、	青春痘、痤疮、壮疙瘩
腋臭	狐臭	狐臭、臭汗症
变态反应	超敏反应	超敏反应
脚癣	足癣、香港脚、脚气病	香港脚、脚气
斑秃	鬼剃头	圆形脱发、圆秃、鬼剃头
花斑癣	汗斑	花斑糠疹、汗斑

4 讨论

如何减少消费者用语和专业术语之间的差距是一个重要的研究课题。本研究选择 Word2vec 作为生成词向量模型的工具，选择天涯社区中的天涯医院版块作为语料来源，以 ICD - 10 疾病名词为种子词，探索 Word2vec 在发现疾病同义词方面的巨大潜力，对候选词的识别效果进行评估。通过观察和归纳，消费者用词的特点可以分为以下几类：（1）口语化表达，如“腹泻”表达为“拉肚子”。（2）别字，如“黄疸”被写成“黄胆”。（3）用英文表达，如“银屑病”的消费者用词有“Psoriasis”的形式。总而言之，消费者健康用词随意性大、形式变化多样，消费者健康信息服务系统在内容组织、功能设计时必须考虑到这一特点，才能更好地为消费者提供健康信息服务。与百度百科里找到的健康词对比发现，两种方法各有优势，本研究基于文本挖掘的方法可以发现非常具有消费者特点的词汇，如“腹泄”、“掉头发”，而百度百科里也可以找出不少的消费者用词，将两种方法相结合构建完整的中文消费者健康词表是未来的研究方向。

5 结语

本研究验证了 Word2vec 在中文词向量表示方面的有效性以及在识别消费者用词方面的强大能力，利用 Word2vec 技术可以开发中文消费者健康词表，从而改善消费者的健康信息查找体验。研究发现，Word2vec 自动生成的候选消费者用词列表质量不足以自动生成消费者用健康词表，但是能作为人工审核构建的重要参考。本研究得出的词向量模型在领域的健康词同义词预测方面具有很高的使用价值，为开发完整的中文消费者健康词表奠定了坚实基础。

参考文献

1 2015 中国网民科普搜索需求行为分析报告 [J]. 科技导报, 2016, (12): 29 - 33.

2 Redlich FC. The Patient's Language: an investigation into the use of medical terms [J]. Yale Journal of Biology & Medicine, 1945, 17 (3): 427.

3 Zeng QT, Tse T. Exploring and Developing Consumer Health Vocabularies [J]. Journal of the American Medical Informatics Association Jamia, 2006, 13 (1): 24 - 29.

4 Li H. Extended Abstract: study on the design for consumer health knowledge organization system in China [M]. Smart Health, 2014: 127 - 129.

5 孙爽. 基于 iOS 的天涯 BBS 客户端的设计与实现 [D]. 成都: 电子科技大学, 2014.

6 刘晓婉, 胡燕祝, 艾新波. 开源中文分词器在 Web 搜索引擎中的应用 [J]. 软件, 2013, 34 (3): 80 - 83.

7 赵涛, 张太红. 农业搜索引擎中文分词工具对比 [J]. 计算机系统应用, 2016, 25 (4): 226 - 231.

8 修驰. 适应于不同领域的中文分词方法研究与实现 [D]. 北京: 北京工业大学, 2013.

9 周练. Word2vec 的工作原理及应用探究 [J]. 科技情报开发与经济, 2015, (2): 145 - 148.

10 罗文馨, 陈翀, 邓思艺. 基于 Word2Vec 及大众健康信息源的疾病关联探测 [J]. 现代图书情报技术, 2016, 32 (9): 78 - 87.

11 Chiu B, Crichton G, Korhonen A, et al. How to Train Good Wrod Embedding for Biomedical NLP [J]. ACL 2016, (2016): 166.

12 Pyysalo S, Ginter F, Moen H, et al. Distrivtional Semantics Resources for Bimedical Text Processing [C]. Proceedings of LBM, 2013, 2013: 39 - 44.