中文文本中两词语关联规律分析*

□李学文　周子璇　熊熊　陈瑜

摘要　分析文本中词语关联规律对于解决信息检索与文本语义研究中许多问题具有重要的价值和意义。首先建立测试平台，对词对语义与构成词对的两词在文本中语义的相符性进行人工判断，然后依据两词语在文本中所处位置差异、文本类型、分句长度、词频等标准，对数据进行分组统计和聚类分析得出两词在文本中的关联规律，最后指出不足及下一步研究的方向。

关键词　中文文本　词语关联　规律

分类号　G354

DOI　10.16603/j.issn1002－1027.2018.01.007

1　引言

研究词语之间的语义关系对解决自然语言理解、人工智能以及机器翻译等方面的问题，都具有重要的价值和意义[1]。信息检索与文本语义研究时，为了提高信息过滤、关联度计算和语义索引建立等各项技术处理结果的完整性、准确性和可靠性都有必要对文本中词语之间的语义关联规律进行分析。信息过滤中，当输入线索是多个检索词（或可转化为多个检索词）时，字段检索、关键词索引检索和全文检索技术会通过布尔检索连接这些词语并判断出所需信息，布尔检索将文本中是否含有检索词作为信息命中与否的依据，用户检索时不管这些检索词有多么生疏，只要在文本中出现了一次就可以检索到[2－3]。例如，对包含文字“宁夏枸杞、兰州百合”的信息进行“兰州 and 枸杞”的检索时，这段信息将符合检索条件，导致获取错误检索结果。排序技术将信息检索结果按照与输入线索的关联度排列，关联度主要是通过关键词在文本中出现的位置和频率进行计算[4]。许多研究利用词的关联关系按照主题凝聚的原则提取出反映主题信息的关键词词典，从而发掘文章主题并进行文本内容分析[5]。语义检索提出增加对文本内容语意的理解，借助语义索引定位符合输入线索的信息，语义索引就是在概念空间的基础上构造具有网状结构的索引，从文档中抽取出概念，同一文档可由具有相关语义的多个概念进行索引[6]。

信息检索与文本语义研究领域有许多关于词语关系的研究，杨梁彬探讨了潜在语义索引解决文本检索中存在的同义和多义问题[7]；国内外有关词语在文本中的语义角色标注的研究比较丰富[8－9]，目前已有成熟的语义角色标注软件；张建娥利用复杂网络中节点的度与聚集特征表示词语之间的关联度[10]；孙曰昕等分析了文本中词语的内联关系和外联关系并指出词语间互信息表征两个词在同一篇文档中的相关性大小[11]；赵冬晓等从词、句子和篇章粒度概括了现有文本语义挖掘方法、算法[12]。这些研究可分为两类：一是基于规则，主要利用语言的词法、句法、词性等知识以及上下文信息来识别词语关系；一是基于统计，主要根据词语在文本中出现的频率、位置等信息，应用不同的统计参数分析词语关系，本文采用第二类方法。

2　测试数据获取

两个词语组合时将两词称为词对，这两个词会限制出比它们各自更具体的语义，称之为词对语义，本文中两词语关联性是通过它们所组成的词对语义与它们所在文本中的语义的相符性来体现，语义相符表示这两个词在文本中关联，不相符则表示不

* 国家社科基金西部项目“丝绸之路宁夏段档案文献遗产目录数据库建设研究”（编号：17XTQ014）的研究成果之一。
通讯作者：李学文，ORCID：0000-0002-7494-1730，23396597@qq.com。

关联。

2.1 约定条件

为了便于分析，特做以下约定：

(1)文本中两词关系设定为：同义词、可搭配、不可搭配。此处不可搭配指两词在语义或语法上矛盾，不可能组合在一起或组合在一起不包含任何语义信息。当可搭配时，两个词所在文本中的语义与词对语义关系分为相符和不相符两种。本文约定：两个词所在文本中的语义与词对语义都匹配时，表示两词在此文本中的语义与词对语义相符；当其中有一个(或两个)所在文本中的语义与词对语义不匹配时，表示两词在此文本中的语义与词对语义不相符，例如：文本“枸杞病虫害可持续调控技术”中包含病虫害调控的含义，但与枸杞调控无关，因此该文本中“病虫害”“调控”两词组成的词对与两词在文本中的语义相符，而“枸杞”“调控”两词组成的词对与两词在文本中的语义不相符。本文主要通过分析语义相符词对数与可搭配词对数的比率特征来发现文本中两词语关联性规律。

(2)提取文本中的词语，并以标点符号为分隔号标记它们所在段、句、分句，同时对段、句、分句按顺序进行编号。其中段分隔号有：“回车符”“换行符”，测试中多段落文本取的是同一标题下相连的段落，且限制在三个自然段以内；句分隔号有：问号、惊叹号、分号、句号；分句不包含任何标点符号，其分隔号有：逗号、顿号、冒号、破折号、引号、书名号、括号等。

2.2 测试过程

2.2.1 建立测试平台并录入信息

首先根据需求建立测试平台，然后选取与“枸杞”相关的网页、期刊、图书等目前常见类型的信息，录入标题、摘要及正文文摘等文本，最后将文本按照段、句、分句等层次进行分割，自动加手动提取文本中的词语，并标记它们所在段、句、分句及分句中的位置。

测试选取的文本样本共 30 个，其中网页 6 个、论文 10 个、图书 13 个、实体介绍 1 个，涉及摘要 5 个、标题 11 个、正文文摘 14 个，多段落文本 2 个。提取词语共 936 种，称每个文本中提取的词语字数与该文本字数(不含标点)比率为词语覆盖率，本测试平均词语覆盖率为 0.80，所有文本样本中最大词语覆盖率为 1.02，最小词语覆盖率为 0.57。

2.2.2 人工判断词对关系并获取测试数据

将每个文本中提取的词语两两组合成词对，并由人工确认词对关系，可选择关系有：默认、相符、不相符、不可搭配、同义词，其中相符与不相符均为可搭配关系。为排除人为因素，本测试选择不同专业不同职称多个人员对词对关系进行判断，以此获取测试数据，共组合词对 26133 组，其中可搭配词对 25872 组。

3 测试数据分析

根据测试需求对人工确认的词对关系以多种因素作为标准分组统计出相符数、不相符数并进行聚类分析，定义相符率为：相符数/(相符数＋不相符数)或相符数/可搭配数，相符率可反映两词语在文本中的关联概率。下面从以下几个方面对两词语关联规律进行分析。

3.1 两词语在文本中所处位置差异

根据词语所在段、句、分句及分句中的位置可确定词语在文本中的位置，称文本中两词语中间所夹文本长度(分句数)为词间距(分句间距)。

测试位于同一分句的两词组成的可搭配词对样本数 5283 组，相符率 43.6%，表示同分句中的两个词语在文本中的语义有 43.6%的可能与这两个词语组合成的词对语义是相符的，即两个词出现在同分句中有 43.6%的可能是关联的，信息检索或语义分析时如果同分句中出现需要检索或分析的两个词，那么这条信息有 43.6%的可能满足或符合要求，43.6%这个值可用作相关度排序依据。所有文本样本中最大相符率 76.3%，最小相符率 24%。图 1 中实线是位于同一分句的两词组成的词对相符率随两词间距变化的折线图，图中仅取了词间距对应可搭配词对数不小于 50 的数据。

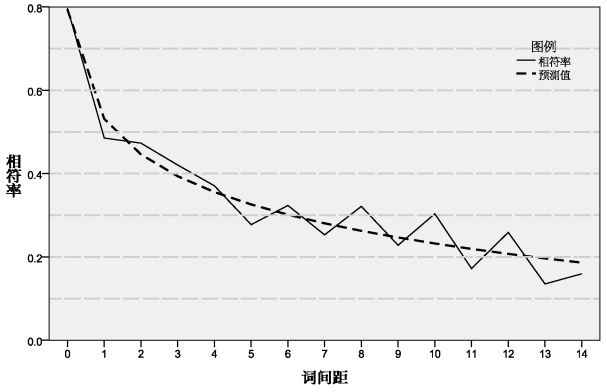

图 1 位于同分句的词对相符率随词间距变化折线图

表 1　位于同分句的词对相符率随词间距变化的函数拟合数据表

函数	参数估计值					ANOVA[a]			
对数函数 a+b*LG(x+c) (b<=0)	参数	估计	标准误	95% 置信区间		R 方	0.934		
				下限	上限	源	平方和	df	均方
	a	0.554	0.045	0.456	0.652	回归	2.017	3	0.672
	b	−0.319	0.049	−0.426	−0.213	残差	0.026	12	0.002
	c	0.178	0.126	−0.098	0.453	未更正的总计	2.042	15	
						已更正的总计	0.390	14	
对数函数 a+b*LN(x+c) (b <= 0)	参数	估计	标准误	95% 置信区间		R 方	0.934		
				下限	上限	源	平方和	df	均方
	a	0.554	0.045	0.456	0.652	回归	2.017	3	0.672
	b	−0.139	0.021	−0.185	−0.092	残差	0.026	12	0.002
	c	0.178	0.126	−0.098	0.453	未更正的总计	2.042	15	
						已更正的总计	0.39	14	

当两词间互相包含(如:abcd、bc)或首末位有交集(如:abc、bcd)时两词间距会小于0,此种情况的样本数252组,相符率26.2%,本文不做过多分析。由图1可见,当两词间距大于等于0时,随着两词间距增大相符率减小。通过SPSS软件对该数据集进行非线行回归分析,依据曲线图型选择适当函数进行拟合,表1是R方值最大的两个函数回归分析结果。

图1中虚线为函数 y=0.554−0.319*LG(x+0.178)的分布曲线。更多函数回归分析结果如下:

幂函数:d+a*(x+c)^{b},(a >= 0; b <= 0; c >= 0)。参数值:a=3.724,b=−0.040,c=0.225,d=−3.160,R方=0.933。

双曲线函数:1/(a+b/(x+c))+d,(b <= 0)。参数值:a=16.174,b=−373.031,c=25.226,d=0.052,R方=0.915。

指数函数:a*e$^{(b*(x+c))}$+d,(a>=0;b<=0)。参数值:a=0.568,b=−0.333,c=0.206,d=0.208,R方=0.890。

a*e$^{(b/(x+c))}$+d,(a >= 0;b >= 0)。参数值:a=0.566,b=2.478,c=3.221,d=−0.452,R方=0.915。

以上函数回归分析R方均接近或大于0.9,说明这些拟合模型能解释因变量90%左右的变异,拟合效果较好。

位于同句不同分句的两词组成的可搭配词对样本数5441组,相符率20.3%。图2是该情况下词对相符率随两词所在分句间距变化的折线图,该图只取了分句间距对应可搭配词对数不小于100的数据。当两词位于同句不同分句时相符率在20%附近徘徊,最大24.9%,最小16.7%。

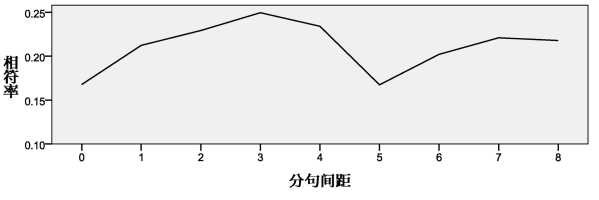

图2　位于同句不同分句的两词相符率随所在分句间距变化折线图

位于同段不同句的可搭配词对样本13541组,相符率9.5%。

位于同一文本不同段的可搭配词对样本1607组,相符率为1.9%。

3.2　分句长度

将位于同分句的词对相符数据以所在分句长度(不含标点符号)进行分组,相符率随分句长度变化如图3所示,其中仅取了分句长度对应可搭配词对数不小于48的数据。由图可知,当分句长度小于等于25时曲线两头低中间高:小于8时相符率在33%附近;在8−20区间内相符率基本在40%到50%之间;大于20时平均相符率为35%。当分句长度大于25时相符率随分句长度变化波动较大。

3.3　文本类型

表2、表3分别从文本出处(标题、摘要、正文文

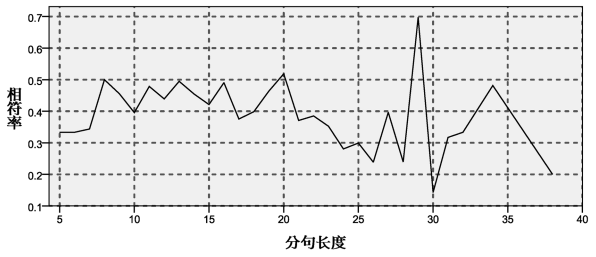

图 3　位于同分句的两词相符率随所在分句长度变化折线图

摘)和信息类型(图书、网页、论文)两个方面对文本中位于同分句的两词语相符率进行分类统计。表 2 显示来自标题、摘要、正文文摘等不同出处的词对相符率差别较大,摘要比正文文摘文本中词对相符率高出 18.7%;表 3 显示三种信息类型文本中词对相符率相差不大。

表 2　位于同分句的两词语相符率按照出处分类统计表

文本出处	相符	不相符	相符率
标题	55	62	0.470
摘要	353	248	0.587
正文文摘	1826	2674	0.406

表 3　位于同分句的两词语相符率按照信息类型分类统计表

信息类型	相符	不相符	相符率
图书	579	852	0.405
论文	776	1002	0.436
网页	824	1068	0.436

3.4　词频

词频和位置对于分析词语和文献主题的关系有重要作用,那么词语在文本中的词频对于它在该文本中与其他词的关联性是否有影响?对词语在每个文本中的频次分别统计,以词频进行分组分析,结果如图 4 所示,其中相符率 1 是先对每个文本以词频分组计算相符率,再计算全部样本中各词频相符率的平均值;相符率 2 是统计每个文本以词频分组后的相符数与不相符数,再合计全部样本中各词频的总相符数与总不相符数,最后计算得出相符率,这两组值有所不同,曲线变化却基本吻合,相符率随着词频的增加在 33%与 62%之间波动。

词频随文本长度增加而增加,对于某文本中的词语,称词频与文本字数之商为词现率,即词现率=词频/文本字数,对相符率与词现率关系统计分析,结果如图 5 所示。其中相符率是先对每个文本以词现率分组计算相符率,再计算全部样本中各词现率对应相符率的平均值,可见,相符率与词现率没有明显函数关系。

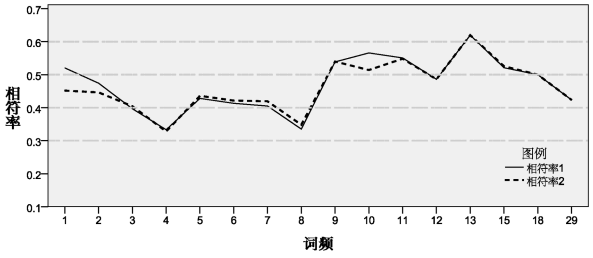

图 4　位于同分句的两词相符率随词语在文本中的词频变化折线图

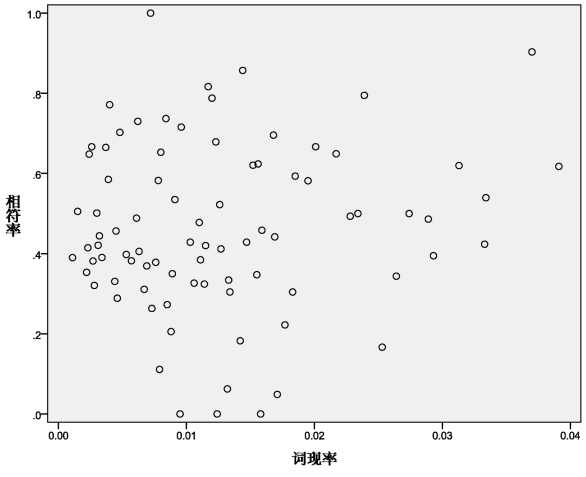

图 5　位于同分句的两词相符率随词现率变化散点图

4　结论及下一步研究方向

4.1　结论

词对所限制语义与构成词对的两词在文本中语义的相符率反映了两词语在文本中的关联概率。由以上测试数据可以得出如下结论:

(1)同分句中两个词关联概率为 43.6%,当两词语相连(词间距为 0)时关联概率为 79.7%,关联概率随着词间距的增加而减小,与对数函数 $y=0.554-0.319*LG(x+0.178)$ 及 $y=0.554-0.139*LN(x+0.178)$ 的拟合度较高。位于同句不同分句时两词语关联概率降低为 20.3%,不到同分句时的一半,没有发现两词语关联概率随着它们所在分句间距的增加而减小或增大的趋势,只是在 20%附近上下波动。位于文本不同段或者不同句时两词语关联概率低于 10%,相对于前面的情况此时两词语关联规律的研究意义相对较小。

(2)位于同分句的两词语关联概率随分句长度

变化而波动。分句长度在8－20时，两词关联概率稳定在45%附近，相对较高；分句长度小于8时关联概率在33%附近，相对较低且稳定；分句长度大于20时关联概率有减小趋势，平均关联概率为35%；分句长度大于24时关联概率波动较大。

(3)标题、摘要和正文等词语出处影响到词语关联概率，概括和总结性文本中词语关联概率较高。图书、论文和网页等不同信息类型文本中词语关联概率差别不大，都在平均值43.6%附近，即没有发现信息类型对词语关联概率的明显影响。

(4)位于同分句的两词关联概率随词语在文本中词频的增加而波动，但维持在40%附近，没有发现关联概率随词频的增加而有减小或增大的趋势。

4.2 不足及下一步研究方向

(1)测试样本不足。因每一条样本数据都来自人工标注，局限于样本数，本测试将文本样本限制到单一领域，选择了枸杞相关文本，分析结果可在该领域内应用，对于其他领域或更大领域内是否具有同样的结论需要进一步研究。

(2)只对中文词语关联规律进行分析，没有对外文进行分析。从语义角度来看，无论什么语种词语关联规律都会存在，但在分词技术及语法上中外文有所不同，因此外文词语关联规律也会表现出不同结果，尤其是英文用空格分割词语，分词更加准确，词语关联规律也将表现得更加明显。

(3)仅对标点符号进行了分类，没有分析不同标点符号对词语关联的影响。本文结论中位于同句不同分句的两词关联概率为20.3%，不到同分句时的一半，由于位于不同分句的两词语词间距比同分句的大，且两词语关联概率随词间距增加而减少，同时不同分句的两词由标点符号分割，各标点符号的作用和意义不同，因此针对标点符号对所分割的词语关联性是否有影响、影响大小等问题的分析具有实际意义，需进一步研究。

(4)未考虑词法、句法、词性、专指度、与文本主题相关性等特性对词语关联概率的影响。表4是统计位于同分句的词对中以词进行分组且词对样本数不低于30的数据。其中“信息数”指包含该词语的文本数，为了避免单一文本对统计结果的影响，表4只取了信息数大于1的样本，从中可以看出不同词语相符率相差很大，这其中是否有规律可循尚需研究。

表4 枸杞相关文本中以词进行分组的同分句词对相符性数据

排序	词语	相符数	可搭配数	相符率	信息数
1	国内	22	31	0.710	3
2	食品	37	54	0.685	3
3	技术	37	56	0.661	6
4	研究	67	108	0.620	8
5	中国	34	55	0.618	4
6	价值	28	46	0.609	4
7	中宁枸杞	44	74	0.595	4
8	栽培	24	43	0.558	6
9	发展	25	45	0.556	4
10	活性	20	36	0.556	2
11	植物	28	51	0.549	5
12	分离	47	86	0.547	3
13	产区	21	39	0.538	3
14	宁夏枸杞	22	41	0.537	2
15	枸杞	383	743	0.515	18
16	产业	19	38	0.500	4
17	安全	15	30	0.500	2
18	中宁	59	122	0.484	5
19	全国	20	42	0.476	4
20	产品	53	112	0.473	6
21	生产	53	115	0.461	6
22	科技	16	35	0.457	3
23	质量	51	114	0.447	6
24	要求	15	34	0.441	4
25	条件	24	55	0.436	4
26	宁夏	30	71	0.423	6
27	企业	34	81	0.420	3
28	追溯	13	31	0.419	2
29	生物	13	31	0.419	3
30	方法	17	43	0.395	3
31	存在	19	50	0.380	5
32	品质	13	35	0.371	4
33	利用	18	50	0.360	3
34	进行	21	59	0.356	7
35	市场	15	44	0.341	6
36	主要	19	57	0.333	7
37	加工	11	36	0.306	3
38	生长	10	34	0.294	5
39	是否	9	31	0.290	2

40	标准	11	38	0.289	3
41	种植	16	57	0.281	6
42	方面	10	43	0.233	4
43	种子	7	31	0.226	2
44	不同	11	54	0.204	4
45	养生	6	30	0.200	3

(5)没有对两个词以上的词对关联规律进行分析。检索线索往往不只包含两个词语,多个词语对语义范围的限制更加具体,分析多词语在文本中的关联规律不但可以满足用户检索需求而且可以提高信息检索、语义分析等操作结果的准确性。

5 结语

本文结论不足以支撑文本中两词语是否关联的确定,测试首先是通过人工确认两词语是否关联,然后以不同标准通过分组的形式对关联与不关联的数据进行聚类分析以发现其中存在的规律性,当将这些规律应用于词语间关联关系的计算时,计算结果与人工确认的关系能达到一定程度匹配(按照二八定律,须达到80%的匹配率)时,文本中词语关联规律才能支撑词语关联关系的确定。

虽然已发现的词语关联规律不足以支撑文本中两词语关联关系的确定,但其中计算文本中两词语关联概率的结论可以用于许多领域。搜索系统利用倒排索引进行预搜索实现数据过滤,获取尽量小的满足用户需求的结果集[13],其中索引技术是当前主流检索系统的主要技术之一,记录有关键词在文本中出现的次数和位置,在现有索引技术的基础上利用文本中两词语关联规律可提升搜索系统的质量。结论应用于信息过滤可排除更多不符合需求的信息,提高数据过滤的准确性,为关键词检索、排序技术提供一种科学的信息相关度排序依据,为文本内容、语义分析以及建立语义索引梳理出新的可行方法。

参考文献

1 常敬宇.语义在词语搭配中的作用——兼谈词语搭配中的语义关系[J].汉语学习,1990,(6):4-8.
2 向桂林,刘锦华.全文检索系统中动态索引技术的研究与实现[J].现代图书情报技术,2003,(3):51-54.
3 方志,夏立新,刘启强.中外全文检索研究的现状及趋势[J].图书情报知识,2006,(5):71-75.
4 杨思洛.搜索引擎的排序技术研究[J].现代图书情报技术,2005,(1):43-47.
5 林宇航.基于词关联关系的文本内容分析[D].北京:北京邮电大学,2013.
6 钟翠娇.网络信息语义组织及检索研究[J].图书馆学研究,2010,(17):68-75.
7 杨梁彬.文本检索的潜在语义索引法初探[J].大学图书馆学报,2003,(6):68-84.
8 Hacioglu K. Semantic role labeling using dependency trees[C]. Proceedings of the 20th International Conference on Computational Linguistics. Association for Computational Linguistics,2004.
9 宋毅君等.基于条件随机场的汉语框架语义角色自动标注[J].中文信息学报,2014,28(3):36-47.
10 张建娥.基于TFIDF和词语关联度的中文关键词提取方法[J].情报科学,2012,30(10):1542-1555.
11 孙曰昕等.融合词语关联关系的自适应微博热点话题追踪算法[J].计算机应用,2014,(12):3497-3501.
12 赵冬晓等.面向情报研究的文本语义挖掘方法述评[J].现代图书情报技术,2016,(10):13-24.
13 吴晓等.基于综合倒排索引的个性化搜索引擎研究[J].微计算机信息,2008,24(27):201-203.

作者单位:北方民族大学图书馆,银川,750021
收稿日期:2017年4月1日

Study on Two Words Correlation Rules in Chinese Text

Li Xuewen Zhou Zixuan Xiong Xiong Chen Yu

Abstract: Analyzing the word correlation rules in a text is of great value and significance to solve many problems in the research of information retrieval and text semantics. This study firstly sets up a test platform and manually confirms the semantic conformity between the word pair and the two words that form the word pair in the text, and then makes the grouping statistics and cluster analysis on the basis of the standards, such as position difference of the two words in the text, text type, clause length, word frequency and gets the correlation rules of the two words in the text. Finally the study points out the deficiency and the following research direction.

Keywords: Chinese Text; Word Correlation; Rule