基于概念向量的文本语义相似度方法探索*

郭红梅[1]，袁国华[1]，胡正银[2]
（1. 中国科学院文献情报中心，北京 100190；2. 中国科学院成都文献情报中心，成都 610041）

摘要：在对概念语义相似度方法调研的基础上，本文提出基于概念向量的文本语义相似度测度方法，借助MetaMap工具抽取文本中的概念术语，将概念术语通过词表层级结构转化为概念向量，通过计算两文本中概念向量的语义相似度来测度两文本的语义相似度。为验证基于概念向量文本语义相似度方法的准确性，选取TREC-05 genomics track数据进行实验，实验结果表明，本文提出的方法较常用的余弦方法更优，与专家评估方法更接近，在测度文本语义相似度上具有一定的可行性和有效性。

关键词：概念向量；语义相似度；文本相似度
中图分类号：G254
DOI：10.3772/j.issn.1673-2286.2017.06.006

1 引言

随着网络技术的发展，出版商将科技文献加工成可供用户查阅的PDF或HTML格式，并发布在Web上，这种电子化形式极大地提高了文本时效性[1]，但同时增加了用户从海量资源中快速准确查找所需知识内容的难度。科技文献间除书目信息的关联外，还存在丰富的语义知识关联[2]，但目前由于缺乏对科技文献完整的语义标注及文本内容相似度的准确测度，读者很难在短时间内把握科技文献发展脉络及知识内容关联[3]。如何测度文本间语义相似度，辅助用户对科技文献间内容关联的挖掘，同时提高检索系统效率，一直是文本挖掘研究中的重要问题。

目前衡量两篇文本相似度大多基于概念空间向量模型，将文本转换为词汇包，却未考虑概念的语境信息和语义层级关联[4-5]。不少学者基于网页查询结果测度概念间语义相似度，如Li等提出非线性测度模型，融合了结构语义信息和信息概念[6]；Cilibrasi等利用搜索引擎检索页面数量测度两个概念的距离，但未考虑同音异义情况，因此，对于不依赖层级分类词表的概念，实施效果不佳[7]；Sahami等通过搜索引擎返回的词片段测度两个查询术语间的语义相似度，特征向量是利用词片段中2 000个句法模式频次形成的，并考虑到4个指标（dice相关系数、重叠相关系数、jaccard系数和逐点相互信息）[8]；Bollegala等通过网页中两个概念的关联页面数，测度两个概念或实体的语义相似度[9-10]；Pilehvar等将文本表示为图结构，从文献、词、段落三个层级分析文本间语义相似度[11]。但这些方法仅是基于定量指标来测度概念的距离相似度，并未考虑概念在词表中的语义相关性及领域信息。也有学者提出依照词表中概念层级结构测度概念间的语义相似度，Zhou等开发MeSHSim R语言包，具体包括5种基于路径的测度方法和5种基于信息内容的测度方法[12]；Yang等基于WordNet中概念的层级位置来测度概念间语义相似度[13]；Lin等基于MeSH词表概念间的层级语义关系，提出文本主题相似度测度定量指标[14]；Bhattacharjee等提出基于概念层级的概念语义相似度测度方法[15]。以上研究仅是探索基于概念层级来测度概念语义相似度，并没有将概念语义相似度方法扩展应用在文本内容的语义测度中。

在借鉴已有学者研究的基础上，本文提出一种基于领域词表的概念向量语义相似度方法，并将基于概念层级的语义相似度方法应用在文本语义相似度测度中。本文首先基于领域词表将概念间层级关系表示为

* 本研究得到ISTIC-EBSCO文献大数据发现服务联合实验室基金项目“基于clique子团聚类的文本主题识别方法研究”资助。

概念向量，然后基于概念向量算法计算概念间的语义相似度，进一步依据两文本中所抽取术语概念的语义相似度来测度两文本的语义相似度。医学领域有较成熟的MeSH词表，词表中疾病、药物、基因序列、蛋白质等概念间存在丰富的语义关联，同时该领域已有完善的术语和语义关系抽取工具和算法，因此，本文选取医学领域数据进行实验，以客观准确测度基于概念向量的文本语义相似度方法的有效性和可行性。

2 相关概念

2.1 概念层级的向量表示

概念向量由概念层级关系得到，概念层级由领域中概念间的隶属关系形成。本文提出利用向量表示概念间层级关系，具体过程如图1和图2所示。对于概念层级中的概念C，其对应的概念向量为$\vec{C}$，表示为“$\vec{C}$：$<N, S, \vec{V}>$”，其中N为概念的名称，S为概念的同义词集，$\vec{V}$为概念的层级向量，这样每个概念都可以表示为“$<N, (S_1, S_2, \cdots, S_m), (\overrightarrow{V_l \rightarrow V_{l-1} \rightarrow \cdots V_1})>$”的形式，其中$m$为同义词的个数，$l$为层级顶端到概念的个数。通过将概念层级中的每个概念转为概念向量，这样概念层级也转化为概念向量层级。概念向量不仅可以清晰反映概念间的层级隶属关系，还可通过概念向量中相同和相异层级的个数直观揭示两个概念的语义相似性。

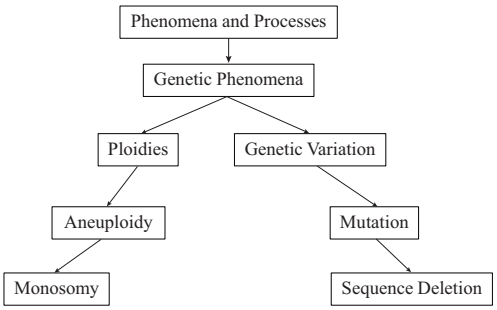

图 1 概念的向量表示——概念层级

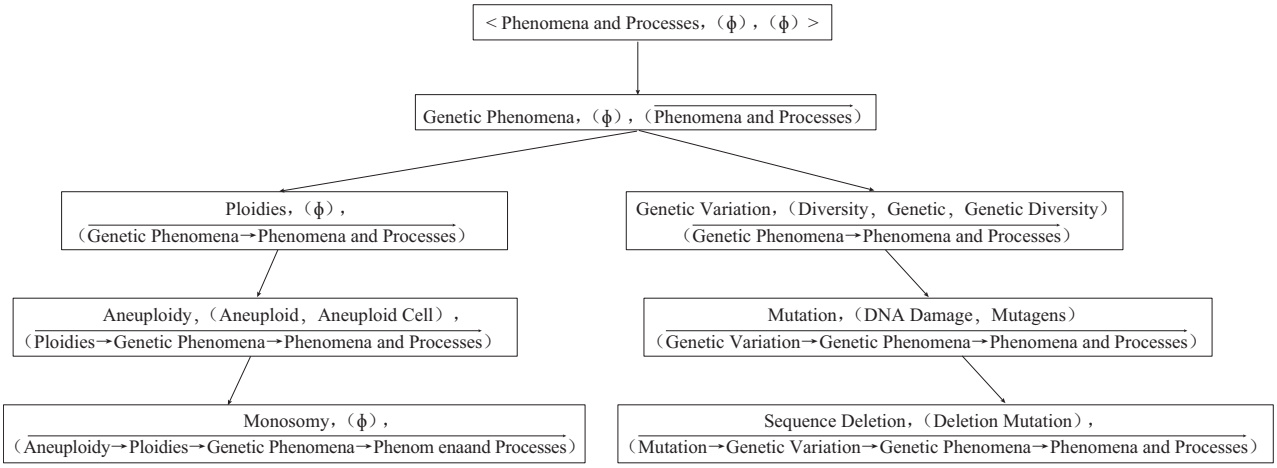

图 2 概念的向量表示——概念向量

2.2 基于概念向量的语义相似度测度方法

2.2.1 概念语义相似度

概念向量中明确给出概念间的层级语义关系，基于概念向量的表示结构可知，两个概念的语义相似度可通过其概念向量层级中相同和相异概念数来测度，将其称为概念向量的语义同质性（semantic homogeneity）和语义异质性（semantic heterogeneity）。语义同质性通过两个概念向量层级中相同的概念来表征，语义异质性通过两个概念向量层级中不相同的概念来表征，文中将概念向量语义同质性和语义异质性的差值作为测度概念语义相似度（concept similarity）的标准，如对于概念C_1和C_2，其语义同质性、语义异质性、概念语义相似度计算方法具体如下。

$$\text{语义同质性}=\frac{\frac{j}{n_1}+\frac{j}{n_2}}{2} \quad (1)$$

$$\text{语义异质性}=\frac{\frac{n_1-j}{n_1}+\frac{n_2-j}{n_2}}{2} \quad (2)$$

$$概念语义相似度=\frac{\frac{j}{n_1}+\frac{j}{n_2}}{2}-\frac{\frac{n_1-j}{n_1}+\frac{n_2-j}{n_2}}{2} \quad (3)$$

其中，n_1为C_1概念向量中包含的概念数，n_2为C_2概念向量中包含的概念数，j为C_1和C_2概念向量中相匹配的概念数。当C_1和C_2为近义词时，二者的语义同质性为1，语义异质性为0，从而进一步计算出C_1和C_2间的概念语义相似度为1。

2.2.2 文本语义相似度

文本由一系列概念术语通过一定的逻辑关系组成，本文将基于向量的概念语义相似度测度方法扩展至文本语义相似度测度，一般认为两篇文章含有的术语概念向量语义相似度越大，这两篇文章的内容相关性越强，语义关联也越强。因此，本文提出通过构建文本的概念向量语义相似度来测度文本语义相似度的方法。为减少低频术语和文本长度对文本语义相似度测量结果的影响，按照布拉德福分布定律选取前半部分的高频术语作为文本内容的表征，进而分别计算两篇文章中高频术语间的语义相似度，所有高频术语语义相似度的均值即为两个文本的语义相似度，具体计算见公式（4）。

$$文本语义相似度=\frac{\sum_{i=1}^{i=n}\sum_{i=1}^{i=m}\text{concept similarity}(C_i, C_j)}{m\times n} \quad (4)$$

其中，m为文本1包含的概念术语数，n为文本2包含的概念术语数，c_i为文本1中的概念术语，c_j为文本2中的概念术语。下文将以具体实验数据来验证该指标用于文本语义相似度测度的有效性和科学性。

3 实验

本文将重点对提出的基于概念向量的概念语义相似度方法和文本语义相似度方法进行实验论证。对基于概念向量的概念语义相似度测度方法，选取WordNet中的28个概念对作为实验数据，对基于概念向量的文本语义相似度测度方法，选取TREC-05 genomics track数据进行实验，验证方法的有效性和可行性。

3.1 数据集的构建

以往学者基于领域词典进行概念语义相似度测度研究大多选取WordNet中的28个概念对进行实验[16]，为更好地与以往研究进行对比，仍选取这28个概念对，分别计算概念对中两个概念间的语义相似度，以验证概念语义相似度测度方法的有效性。

为验证本文提出的基于概念向量的文本语义相似度计算方法的有效性，选取TREC-05 genomics track数据进行实验，其共包含PubMed数据库的34 633篇文献，这些文献被分为5个研究领域，分别为进行某项实验或过程的标准方法或协议、在某种疾病中基因的作用、在特定生物过程中基因的作用、在某种疾病或器官功能中两个或更多基因间的交互作用、特定基因变异和其生物效应和作用，此外，每个研究领域又分为10个主题。领域专家或评估人员分别对50个主题中的每篇文章与该子主题的相关性进行打分（0—2分），0分表示不相关，1分表示部分相关，2分表示非常相关，共有4 232篇文献相关性分值为1分或2分。选取相关性论文数大于100篇的11个主题的文献构建实验数据集，具体如表1所示。

表 1 11个主题数据基本情况

主题编号	主题描述	不相关论文数/篇	相关论文数/篇
117	基因ApoE在疾病Alzheimer的作用	385	653
146	下视丘分泌素受体2变异和其在嗜睡中的作用	388	421
114	APC基因在结肠癌中的作用	375	346
120	NM23基因在肿瘤发展过程中的作用	182	331
126	P53基因在细胞凋亡中的作用	1 013	307
142	音猬因子变异和其在疾病发展中的作用	257	263
108	识别在活体中蛋白与蛋白交互的过程或方法	889	191
107	用于微阵列数据标准化的过程或方法	294	189
111	PRNP基因在疯牛病中的作用	473	185
109	5-核酸序列荧光实验过程或方法	210	175
106	染色体IP到分离蛋白的过程或方法	1 061	158

3.2 实验过程

MetaMap是美国国立医学图书馆基于一体化医学语言系统（Unified Medical Language System，UMLS）开发的句法解析工具，可根据语义将句子拆分成若干具有意义的短语片段，并进一步将短语中的词或词组与UMLS词表进行映射，获取各术语的概念向量。实验首先利用医学领域术语抽取工具MetaMap对4 232篇实验数据进行术语识别，并将抽取到的前半部分高

频术语与MeSH词表中的概念和层级结构进行映射；其次，按照上述概念层级的向量表示将每篇文章中抽取出的高频概念术语转化为概念向量表示，这样每篇文章即可利用高频的术语概念向量表示；最后，按照概念语义相似度和文本语义相似度的计算方法分别得出概念语义相似度和文本语义相似度值。

在方法的有效性验证方面，针对概念语义相似度方法，选取以往研究的28个概念对，对比文中方法和以往较具有代表性的概念相似度测度方法，通过Person和Spearman相关分析来验证文中方法的有效性；针对文本语义相似度测度方法，将TREC-05 genomics track人工标注结果和余弦相似度测度方法进行对比，分别验证概念语义相似度方法和文本语义相似度方法的有效性和可行性。具体实验过程如图3所示。

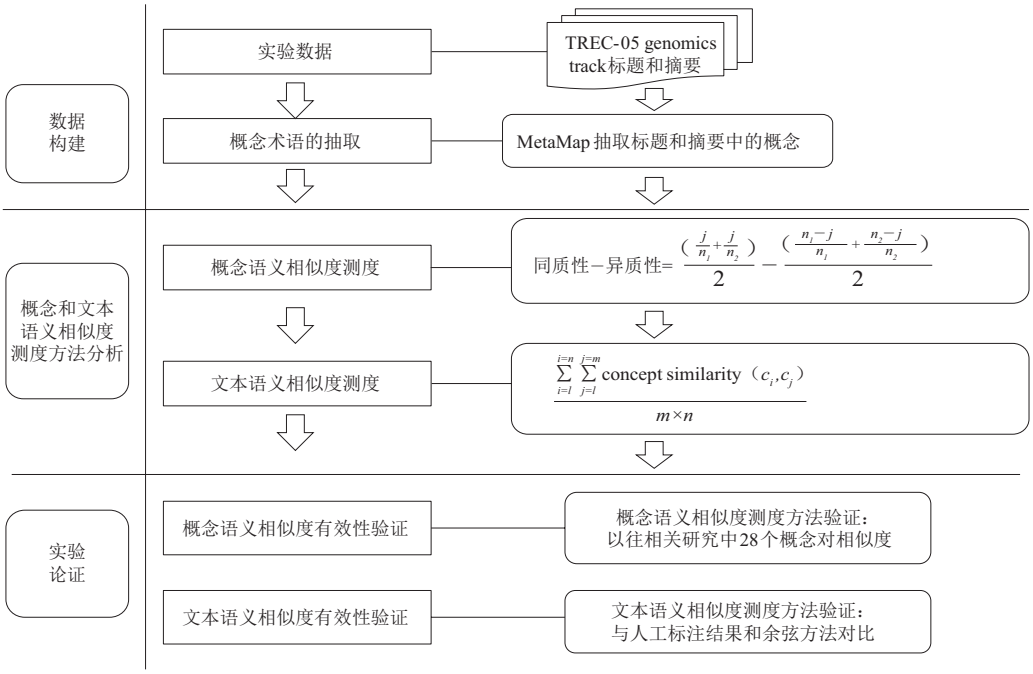

图 3 实验过程

3.3 实验结果分析

表2中28个概念对通过本文方法和以往具有代表性的概念相似度计算方法PMI[8]、Support Vector Machine-Based Approach（SVM）[8]、Relational Model Based Similarity Measurement Approach（RMSS）[9]、Co-occurrence Double Checking Model（CODC）[10]的相似度分值，表3给出5种方法的Person和Spearman相关性检验，从不同测度方法在28个概念对相似度的Person和Spearman相关系数可以看出，基于概念向量的文本语义相似度方法高于其他指标，且兼顾概念语境和领域信息，能更好地测度概念间的语义相似度。

利用MetaMap术语抽取工具分别对实验集中的4 232篇文章进行标题和摘要中概念术语抽取。由于抽取概念术语的个数与文本长度有关，一般认为文本越长抽取到的术语概念越多。由公式（2）可知，基于概念向量的文本语义相似度方法与抽取概念术语个数相关，为减少低频术语和文本长度对文本语义相似度测量结果的影响，本文按照布拉德福分布定律对每篇文章前半部分的高频术语进行语义相似度分析。

结合公式（3）和每篇文章抽取的前半部分高频术语分别计算每个主题下两个文本间的语义相似度。语义相似度越大认为两篇文章的研究内容越相似，一般可认为在同一个主题中语义相似度越大的文本集与该主题越相关，因此将每个主题中语义相似度大于均值的文本集等同于TREC-05 genomics track系统中相关度分值为1分或2分的文本。从11个主题中抽取的高频术语数、平均语义相似度和相似文本数，由表4中数据可知，有8个主题通过基于概念向量方法得到的相关论文数大于TREC-05 genomics track系统中专家标注相关论文数。

由表5可见，本文方法与余弦方法或TREC-05

表 2 本文方法与其他四种方法概念对语义相似度分值

概念对	PMI	SVM	RMSS	CODC	本文方法
car-automobile	0.427	0.980	0.918	0.008	1.000
gem-jewel	0.468	0.996	1.000	0.005	0.846
journey-voyage	0.688	0.686	0.817	0.012	1.000
boy-lad	0.632	0.974	0.958	0.000	0.850
coast-shore	0.561	0.945	0.975	0.006	0.750
asylum-madhouse	0.813	0.773	0.794	0.000	0.864
magician-wizard	0.863	1.000	0.997	0.008	1.000
midday-noon	0.586	0.819	0.987	0.010	1.000
furnace-stove	1.000	0.889	0.878	0.011	0.600
food-fruit	0.449	0.998	0.940	0.004	0.571
bird-cock	0.428	0.593	0.867	0.006	0.864
bird-crane	0.516	0.879	0.846	0.000	0.654
tool-implement	0.297	0.684	0.496	0.005	0.813
brother-monk	0.623	0.377	0.265	0.007	0.414
crane-implement	0.194	0.133	0.056	0.000	0.524
lad-brother	0.645	0.344	0.132	0.005	0.414
journey-car	0.205	0.286	0.165	0.004	0.000
monk-oracle	0.000	0.328	0.798	0.000	0.505
food-rooster	0.207	0.060	0.018	0.000	0.181
coast-hill	0.350	0.874	0.356	0.000	0.333
forest-graveyard	0.495	0.574	0.442	0.000	0.000
monk-slave	0.611	0.375	0.243	0.000	0.678
coast-forest	0.417	0.405	0.150	0.000	0.100
lad-wizard	0.426	0.220	0.231	0.000	0.478
chord-smile	0.208	0.000	0.006	0.000	0.000
glass-magician	0.598	0.180	0.050	0.000	0.000
noon-string	0.228	0.017	0.052	0.000	0.000
rooster-voyage	0.102	0.018	0.000	0.000	0.000

表 3 本文方法与其他语义相似度测度方法的相关性检验

	PMI	SVM	RMSS	CODC	本文方法
Person相关系数	0.549	0.834	0.867	0.694	0.872
Spearman相关系数	0.518	0.818	0.849	0.702	0.869

genomics track系统人工标注结果的对比信息可知，在11个主题中，基于概念向量语义相似度方法识别出的相关论文数有9个主题小于余弦方法，但是与专家匹配论文数均高于余弦方法。

由表6中数据可知，基于概念方法的准确率和召回率均高于余弦方法。同一主题的论文具有相同关键词，余弦方法只把术语表示为特征向量，基于文档分布分析文本的相似度，没有考虑术语自身的语义关联，因此很容易错误地将更多具有相同关键词的文本判断为一个主题。基于概念向量的方法兼顾术语的语义层级关联，在语义相似度对比上更合理。层级结构中包含一定的语义和语境，明确语境信息的重要性，通过实验证明该方法的合理性和有效性。

表 4 基于概念向量语义相似度方法实验结果

主题编号	专家判断相关论文数/篇	相关论文数/篇	平均语义相似度	本文方法判断的相关论文数/篇
117	653	4 573	4.65	524
146	421	3 724	5.65	477
114	346	3 589	6.65	407
120	331	2 045	7.65	329
126	307	6 007	5.45	387
142	263	3 021	6.31	304
108	191	4 231	10.65	183
107	189	2 864	9.87	226
111	185	3 252	8.65	221
109	175	1 769	9.29	208
106	158	5 013	6.24	189

表 5 基于概念向量语义相似度方法与余弦方法相关论文集判断结果

篇

主题编号	专家判断相关论文数	基于概念向量相似文本数	余弦方法相似文本数	专家-概念向量匹配数	专家-余弦方法匹配数
117	653	524	626	449	408
146	421	477	428	327	289
114	346	407	443	284	276
120	331	329	338	252	245
126	307	387	405	273	267
142	263	289	297	185	169
108	191	183	205	131	124
107	189	226	242	152	137
111	185	221	207	146	115
109	175	208	227	139	124
106	158	189	219	132	117

4 讨论

已有学者尝试基于词表中的概念层级测度概念间的语义相似度，但大多局限于对词表概念间测度方法的理论研究，并没有将基于概念层级的测度方法应用在文本相似度分析或文献检索系统中。本文尝试将概念层级转化为概念向量，并将概念向量语义相似度方法扩展到文本间语义相似度测度研究中，通过实验验证该思路和方法的合理性和有效性，具体体现在以下两点。

表 6 基于概念向量语义相似度方法与余弦方法实验结果的对比

%

主题编号	概念向量准确率	余弦方法准确率	概念向量召回率	余弦方法召回率
117	85.69	65.18	68.76	62.48
146	68.55	67.52	77.67	68.65
114	69.78	62.30	82.08	79.77
120	76.60	72.49	76.13	74.02
126	70.54	65.93	88.93	86.97
142	65.13	56.90	75.29	64.26
108	71.58	60.49	68.59	64.92
107	67.26	56.61	80.42	72.49
111	66.06	55.56	78.92	62.16
109	66.83	54.63	79.43	70.86
106	69.84	53.42	83.54	74.05

（1）文本由许多具有语义信息的概念术语按照一定的逻辑关系构成。基于概念向量的文本语义相似度测度方法在计算两个文本的相似度时，除考虑相同或相似概念术语数外，还兼顾概念术语在词表层级结构上存在的逻辑和语义关系，符合文本构成规律，利用该方法测度文本间的语义相似度具有一定的科学性和合理性。

（2）文中方法具有一定的可行性。目前很多词表已提出适合自身的概念语义相似度测度方法，并且提供相应算法，这为基于概念语义相似性测度文本语义相似度提供了理论和底层数据支持。此外，文中实验数据也证明该方法较以余弦为代表的特征向量方法效果更优。

基于概念向量的文本语义相似度方法在概念术语集的构建和语义相似度阈值的选取上仍有待进一步优化。目前，按照布拉德福分布定律选取前半部分的高频术语进行语义相似度分析，造成一些低频概念术语信息的丢失。在筛选每个主题下的相关文本时，仅选取集合内的均值作为阈值，未来研究将考虑通过一定量的训练集来设定相似度阈值。此外，由于语义相似度越大的两个文本在内容上越相关，研究主题越相似，因此未来也可进一步探索将基于概念向量的文本语义相似度方法应用在聚类或文本语义网络分析中，以实现对内容相似文本的聚类分析或重要主题识别。

参考文献

[1] CASTRO L J G,BERLANGA R,GARCIA A.In the pursuit of a semantic similarity metric based on UMLS annotations for articles in PubMed central open access[J].Journal of Biomedical Informatics, 2015,57(C):204-218.

[2] D' HONDT J,VERHAEGENP A,VERTOMMEN J,et al.Topic identification based on document coherence and spectral analysis[J].Information Sciences,2011,181(18):3783-3797.

[3] MEZA B A.Searching and ranking documents based on semantic relationships[C].International Conference on Data Engineering,2006.

[4] HLIAOUTAKIS A,VARELAS G,VOUTSAKIS E,et al.Information retrieval by semantic similarity[J].International Journal on Semantic Web and Information Systems,2006,2(3):55-73.

[5] RYANG W, BERNARDH R.Techniques to identify themes[J].Field Methods,2003,15(1):85-109.

[6] LI Y,BANDAR Z A,MCLEAN D.An approach for measuring semantic similarity between words using multiple information sources[J].IEEE Transactions on Knowledge and Data Engineering, 2003,15(4):871-882.

[7] CILIBRASI R L,VITANYI P M B.The Google similarity distance[J]. IEEE Educational Activities Department,2007,19(3):370-383.

[8] SAHAMI M,HEILMAN T D.A web-based kernel function for measuring the similarity of shorttext snippets[C]//Proceedings of the 15th International Conference on World Wide Web.[S.1.]:[s.n], 2006:377-386.

[9] BOLLEGALA D,ISHIZUKA M,MATSUO Y.Measuring semantic similarity between words using web search engines[J].Computer Science,2015:757-766.

[10] CHEN H,LIN M,WEI Y,et al.Novel association measures using web search with double checking[C]//International Conference on Computational Linguistics.[S.1.]:[s.n],2006:1009-1016.

[11] PILEHVAR M T,NAVIGLI R.From senses to texts:an all-in-one graph-based approach for measuring semantic similarity[J]. Artificial Intelligence,2015,228:95-128.

[12] ZHOU J,SHUI Y,PENG S,et al.MeSHSim:an R/Bioconductor package for measuring semantic similarity over MeSH headings and MEDLINE documents[J].Journal of Bioinformatics and Computational Biology,2015,13(6):1542002.

[13] YANG D,POWERS D M W.Measuring semantic similarity in the taxonomy of WordNet[J].Journal of Structural Biology, 2007, 159(1):36-45.

[14] LIN J,WILBUR W J.PubMed related articles: a probabilistic topic-based model for content similarity[J].BMC Bioinformatics, 2007,8(1):1-14.

[15] BHATTACHARJEE S,GHOSH S K.Measurement of semantic similarity: a concept hierarchy based approach[C]//Proceedings of 3rd International Conferenceon Advanced Computing, Networking and Informatics,Smart Innovation,Systems and Technologies.[S.1.]: Springer India,2016:407-418.

[16] MILLER G A.WordNet:a lexical database for English[J]. Communications of the ACM,1995,38(11):39-41.

作者简介

郭红梅，女，1985年生，博士，馆员，研究方向：文本挖掘、科学计量分析，E-mail：guohm@mail.las.ac.cn。
袁国华，男，1983年生，博士研究生，工程师，研究方向：文本挖掘、网络与信息安全。
胡正银，男，1979年生，博士，副研究员，研究方向：文本挖掘、语义分析。

Measurement of Text Semantic Similarity on the Basis of Concept Vector

GUO HongMei[1], YUAN GuoHua[1], HU ZhengYin[2]
(1. National Science Library, Chinese Academy of Sciences, Beijing 100190, China;
2. Chengdu Documentation and Information Center, Chinese Academy of Sciences, Chengdu 610041, China)

Abstract: Based on the previous studies on the concept semantic similarity, this paper proposed measurement of text semantic similarity on the basis of concept vector. First, mining the concepts or terms from the texts. Second, transforming concepts or terms into concept vector followed by hierarchical structure of vocabulary. At last, measuring the sematic similarity of concepts or terms and further measuring the text semantic similarity. The paper used TREC-05 genomics track data to experiment. The results showed that the method of text semantic similarity on the basis of concept vector was better than cosine, which was more closely to expert evaluation result.

Keywords: Concept Vector; Semantic Similarity; Text Similarity

（收稿日期：2017-05-08）