情报工程 ISSN: 2095-915X
第1卷 第6期 102-108 2015 年 12 月

TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN: 2095-915X Vol.1 No.6 102-108 Dec 2015

doi:10.3772/j.issn.2095-915x.2015.06.014

科技政策语义分析方法研究

王小玉，董诚，曾文

（中国科学技术信息研究所 北京 100038）

摘要：本文对科技政策及语义分析方法研究现状的进行了调研，重点对句子相似度匹配算法和段落相似度匹配算法，以及倾向性分析算法等几种语义分析方法进行了对比分析，并总结其各自的适用性和优缺点，为下一步科技政策语义分析模型的构建研究提供研究基础。

关键词：科技政策，语义分析方法，相似度计算，倾向性分析

The Research for Semantic Analysis Method of Science and Technology Policy

WANG Xiaoyu, DONG Cheng, ZENG Wen

(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract: This paper is based on the semantic status of science and technology policy research and analysis.. It focuses on sentence similarity matching algorithm. Several paragraphs similarity matching algorithms and semantic analysis method tendentious analysis algorithms were compared and summarized their respective applicability, advantages and disadvantages. This paper sets the foundation of the construction for the semantic analysis model of science and technology policy in the future.

Key words: Science and technology policy, semantic analysis, similarity calculation, sentiment orientation analysis

基金支持：本研究得到国家科技支撑计划项目（2015BAH25F00）和国家社会科学基金项目（14BTQ038）的支持。

作者简介：王小玉（1992-），女，中国科学技术信息研究所信息资源管理专业硕士研究生，研究方向：信息资源管理、科技政策语义分析；董诚（1970-），男，中国科学技术信息研究所研究员，研究方向：科技管理与科技创新；曾文（1973-），女，中国科学技术信息研究所副研究员，研究方向：智能信息处理，数据分析和知识组织。

1 引言

科技政策是指政府为促进科技发展，利用科学技术为国家目标服务而采取的集中性和协调性措施，是科学技术与国家发展的有机整合[1]。科技政策作为国家科技发展战略的重要部分，对科技与社会、经济协调发展具有重要作用。近几年发达国家的政府和学界对于科技政策研究的规范和科学化异常重视，有学者[2]提出科技政策学，就是为了梳理科技政策研究领域的数据、方法、工具并将其作为新兴的交叉学科发展，可见科技政策研究的重要程度。科技政策纷繁复杂，不仅政策更新速度快而且政策范围广包含国家政策、地方政策，所以不少学者针对科技政策的具体问题进行研究。然而目前的这些研究大部分是针对政策的制定者提出政策建议与改进，可供政策受益者例如企业所做的政策服务鲜有研究。本文认为：通过对科技政策的深层语义分析，可以充分了解科技政策的影响方向如何、影响强度如何，不仅对政策制定者明确问题和调整政策提供一定的参考，也会对企业决策有一定的帮助。

2 相关研究现状与分析

对科技政策的语义分析，主要是科技政策的语义相似度分析和倾向性分析，语义相似度分析可以解析政策之间的语义相似性，倾向性分析可以解析政策的影响方向和强度。

2.1 科技政策研究现状与分析

由于科技政策的研究是一个颇具多样性的领域，对科技政策研究方法、工具和理论也多种多样，来自不同学科的研究都有其自身的研究基础和方法论，形成了各种各样的研究框架和研究方法，它们之间既存在共性，又有差异。目前针对科技政策的分析，从分析的主题上来看，包括科技评价、技术创新、高薪技术企业、科技成果转化等主要主题，研究的方法和工具涉及经济学、社会学、政治学、公共政策等多个学科，但是多以问题为导向进行研究。

黄萃等[3]提出一种政策工具视角下的政策文本量化研究方法，根据政策工具理论制定分析框架并进行频数统计，在量化分析的基础上提出政策建议，汪涛等[4]提出一种类定量化的科技政策文本分析框架，通过对一定年份北京市科技政策的演进分析来验证该框架的合理性并提出了政策实践的改进建议，仲伟俊等[5]同样是在政策工具的视角下构建政策分析框架，建立了基本政策工具纬度、科技活动类型纬度、科技活动领域纬度的三维分析框架，通过实证分析提出我国现有科技政策的不足和展望。以上文献均是为了政策制定者和政策主体制定政策、解决政策问题等提出的分析框架，为制定者明确问题和调整政策提供了参考。但是针对政策作用对象的针对性分析则较少，为了对科技政策进行针对性的分析，不仅要关注发布时间、作用范围等外部特征，还要关注内在的深层语义特征，及政策的应用对象是什么、影响方向如何、影响强度如何。与传统的依靠关键词检索的科技政策检索库相比，对科技政策的深层语义分析还可以帮助企业快速、准确找到其需求的政策。

2.2 语义分析研究现状与分析

2.2.1 语义相似度匹配算法研究

（1）句子相似度计算

句子相似度计算是对句子间的相似性给出一个度量，它在自然语言处理的许多领域中都发挥着重要的作用。从不同的句子分析形式来看，当前句义相似或相关度计算主要分为两大类：①基于句子词汇层面的句义相似度计算，主要包括基

于词的统计特征、词汇语义特征两方面；②基于句子结构层面的句义相似度计算。

基于词统计特征的方法主要是通过考虑词的特点来简介衡量句子的相似度，例如词频、词性等信息，如向量空间模型法，它将语料库中的句子表示为特征词向量，再根据向量的空间位置简介得到相似度，一般是用向量夹角余弦表示，但是该方法未考虑句子的结构特征，而且只有当语料库有一定的规模时这种统计的效果才会体现出来。基于词汇语义特征的方法主要利用 HowNet(知网)、WordNet 和同词词林等词汇语义词典。例如，李素建 [6] 以 HowNet 和同义词词林为依据提出了语句相关度的定量计算模型，该方法较为依赖语义词典的完整性，词典的不全面将直接影响计算的准确性。唐琦 [7] 在词汇语义相似的基础上，利用格语法理论获取句子之间的语义相似度。史燕 [8] 利用概念层次理论计算词语概念基元、词语的语义相似度，在此基础上考虑词语的融合得到语义块，进而得到句义相似度。基于句子结构层面的句义相似度计算主要是在句法分析基础上，按照分析出的句子结构来衡量句子之间的相似度。例如，以汉语框架网语义资源为基础，李茹等 [9] 通过多框架语义分析、框架的重要度度量、框架的相似匹配、框架间相似度计算等关键步骤来实现句子语义的相似度量。类似地，李彬等 [10] 计算句子的相似度是通过衡量句子间核心词与其直接依存成分之间的语义搭配对之间的相似度来确定的，由于该方法仅考虑句子的核心语义依存结构，未将全部语义依存信息考虑在内，因此句子相似度度量结果不够准确。为充分考虑句子的全部语义，李春梅等 [10] 尝试提出多特征的汉语句子相似度的计算模型，在基于词的基础上，不仅考虑句子中词的表层而且考虑句子中词的逻辑联系，在句子层面上，从只考虑句子的局部结构到考虑句子的整体结构，用句子的区分度、相同词的相似度、长度相似度、词性相似度及词序相似度五个方面来综合考虑两个句子相似度的计算。

（2）文本相似度计算

段与段的相似度计算，可以理解为对文本进行相似度计算。目前，文本相似度计算方法主要分为基于大规模文本集统计的传统方法和基于语义计算的方法。

基于大规模文本集统计的方法，通常采用向量空间模型 (Vector Space Model, VSM)[12] 和隐性语义索引模型 (Latent Semantic Indexing, LSI)[13] 等方法。这些方法均基于段落中的词进行相似度计算，未考虑特征项的语义信息，与句子相似度计算中基于词的统计特征的方法相似，由于只考虑词在上下文中的统计特性，而没有考虑词和段落的语义，因此该方法具有一定的局限性。基于语义计算的文本相似度计算方法的主要研究有：Sanchez J A[14] 和 Vicient C 等 [15] 提出基于本体的文本特征抽取及文本相似度计算方法。基于本体的方法由于其本身的构建是一项较为复杂的工程，不仅需要领域专家的参与还需要大量调研工作，同时该方法过于注重本体树中各个概念节点的结构分布，导致基于本体的计算文本相似度方法应用不是很广泛。基于外部语义词典的文本相似度计算方法主要有基于 Hownet(知网) 和基于 WordNet 两种，知网的研究主要集中在我国，WordNet 是由美国普林斯顿大学学者联合设计的语义词典且起步早，故而研究相对较多，利用基于语义词典的方法计算语义相似度时，可以充分考虑语义词典的结构及其语义信息。例如，Capelle M 等 [16] 提出基于内容的新闻推荐方法，该方法利用 WordNet 的词汇结构和语义信息来计算新闻内容中同义词的语义相似度。在计算词汇相似度的基础上，进而考虑句子和段落语义信息，例如，Bhagwani S 等 [17] 计算句子相似度，提出基

于语义词典和大规模统计方法结合来计算句子的相似度。有学者利用语义词典 WordNet 与概念森林 (Concept Forests) 的树型结构相结合来计算语义相似度，例如，基于 WordNet 的文本相似度计算方法 [18]，该方法从文本中构造概念森林来表示文本的语义，并通过计算概念森林之间共有部分语义相似度来度量两文本的相似性，Tsatsaronis G[19] 提出基于 Wikipedia 和 WordNet 对文本的词汇资源建立语义森林。

2.2.2 文本倾向性分析研究

随着互联网的高速发展，大量网络数据和资源出现，同时由此引发各种新的任务也产生许多新的算法，倾向性分析是近年学术界的一个热点问题之一，它涵盖了文本检索、信息抽取、机器学习等各个领域。网络上存在大量与倾向性相关的信息的文档：博客，新闻评论，商品在线评论等。从自然语言处理角度来看，针对这些倾向性的文档主要有两类任务：倾向性分类和倾向性信息抽取 [20]。

上世纪 90 年代起，国外就开始了对词汇倾向性的分析研究，Turney 提出了一种通过一组基准词计算词语的情感倾向性的方法，达到了 95% 的准确率 [21]；Kim 等人同样将工作重点放在情感词汇的倾向性分析上，在一对基准词集的基础上使用 WordNet 计算未知词汇的情感倾向性 [22]。随着研究工作和实际应用领域的发展，对整篇文档的观点抽取和倾向性判断成为研究工作的热点，情感词的上下文信息和语义搭配关系也逐渐被应用到语义倾向性计算当中。而在实际工作中，单词的倾向性与短语的倾向性往往相反，Wilson 和 wiebe 等人在后期研究中着力研究了短语级情感倾向性，并对中立情感这一实际大量存在的文本进行研究 [23]。

在国内研究中，对主观句所表达的情感倾向进行褒贬识别，主要包括两种分类方法：运用机器学习方法和基于情感词的方法。

（1）运用机器学习的方法，首先需要人工标注一些文本情感倾向并把其作为训练语料，接着对其进行训练同时构造分类器，最后通过对未知类别的文本进行分类测试得出文本的情感倾向。使用机器学习方法进行情感分类时，分类算法的选择和特征项的选取是最重要的两个方面。PangBo[24] 最早将机器学习方法应用于情感分类领域，他分别利用朴素贝叶斯、最大熵、支持向量机算法对电影评论进行分类，当以 unigram(自然语言处理中代表单个 word) 作为特征项时，支持向量机表现最好，准确率为 82.9%，朴素贝叶斯法和最大熵的计算效果相当。在特征项的选择上，崔彩霞等 [25] 提出一个特征项选择函数，用来替代传统的文档频率和互信息选择方法。另外，王素格等 [26] 研究了停用词对中文文本情感分类的影响，她构造了五种停用词表作为特征项选择的依据，实验表明停用词表的选择对文本情感分类的影响很大。

（2）基于情感词的方法，其基本思路是通过判定句子中包含情感词的语义倾向，加上句法结构等信息，间接得到句子的情感倾向。基于语义的情感倾向分析研究是对文本计算一个情感倾向值，值的符号表示其倾向性，而其绝对值的大小则反映其情感强度。如朱嫣岚等人 [27] 利用 HowNet 提供的语义相似度和语义相关场计算功能对词语的褒贬倾向度按一定计算法则进行赋值，并根据该值判别该词语义倾向，并在后续工作中利用词语倾向性进行计算文本倾向性。许歆艺等 [28] 提出了一种基于文本纹理特征的中文情感倾向性分类，通过测试多种文本纹理特征对文本情感倾向性的影响，成功将文本纹理特征融入情感分类中，通过计算各类特征与文本的情感倾向性的相关度，对特征进行降维，相对于基于词频的情感倾向性分类方法，查准率平均提高了 10% 左右。

采用基于情感词的方法判定句子情感倾向时，能否得到情感倾向准确、包含全面的情感词集是关键，同时也要考虑一些特殊的句法结构对结果的影响，如否定句、比较句等。郝玫等 [29] 提出一种中文网络评论的复杂语义倾向性计算方法，该方法在建立产品领域情感词典的基础上，首先确定特征观点窗口的度量范围，完成特征观点组的提取；然后在特征观点组中综合考虑观点词的程度、反转语义及特征评价的频数等多种因素，完成特征评价倾向性的计算。

2.3 研究评述

由以上研究可知，目前句子语义相似度计算的研究中，基于词层面特征的相似度计算方法未考虑句子的结构信息；基于句子结构特征的相似度计算方法也具有一些缺陷，有的未能全面考虑句子语义，有的利用的资源存在限制，计算句子相似度时，综合考虑句子的结构信息和词汇语义信息等因素，是十分关键的问题。上述的几种关于文本相似度计算方法在一些特定的领域中均有比较好的效果，但其中仍有许多的不足，不能被广泛的使用，总体来说，基于大规模文本集统计的传统方法，能在词汇出现的频度和频率层面上反映两个文本的相似程度。但是一个有实际意义的文本，它有其实际的语义与中心思想，这是语义层面上的概念，利用统计方法计算出来的中心思想与整篇文本实际表达的中心思想可能会相差甚远，基于语义层面的方法，这类方法利用语义词典对文本中的词汇进行语义分析，但没有考虑语义之间的内在联系，也没有考虑不同词汇对文本的重要程度的差异，因此计算的准确率受到影响，所以关于文本相似度计算的方法也有很多学者进行不断的研究与改进。在现有的相似度计算方式中，主要以句子或文本中词汇作为基本处理单元，较少考虑文本内部的句子结构特征和组合语义特征，以统计方式为主的相似度计算方法在传统的信息检索、聚类、分类等领域应用较广，可以保证相似度计算的高效性，而基于语义规则的相似度计算由于其计算的复杂性难以直接应用于大规模文本处理的领域。

由以上文本倾向性分析的研究可知，文本倾向性分析尽管已经引起了学者们的普遍关注，但尚未被广泛应用于政策领域，因此，本文认为文本倾向性分析对于科技政策的研究和分析是具有探索价值的。对科技政策的倾向性分析，应立足于基于语义的情感倾向性研究，针对政策文本的特点，如政策文本的倾向性明确，不存在局部倾向性会对整体倾向性造成效果差异，政策文本不存在一般网络评论会有的“先抑后扬”的结构等，根据政策文本的特点构建适用的算法，再通过政策情感词典的构建，计算政策文本的情感强度。

3 总结与展望

本文通过对相似度匹配算法研究现状和倾向性分析研究现状的调研，总结提炼若干算法并提出其适用性和优缺点，为相关语义分析模型的构建提供一定的参考。目前对于科技政策的系统分析较为缺乏，且大部分研究是为政策制定者提供政策建议，本文认为：科技政策的分析和应用需要综合考虑科技政策的政策影响力、政策关键要点分布与政策内的统计频次以及行文措辞等方面因素。例如：可以通过度量情感强度，并经过反复多次的调研和调整来建立科技政策的语义分析模型，这对于政策制定者、政策研究者和企业都有一定的参考价值。下一步，我们拟针对政策文本的特点，重点立足于语义相似度和情感倾向分析的深入研究，构建其适用的模型算法。

参考文献：

[1] 朱崇实 , 陈振明 . 公共政策 – 转轨时期我国经济社会政策研究 [M]. 北京 : 中国人民大学出版社 ,1999.

[2] 陈光 , 方新 . 关于科技政策学方法论研究 [J]. 科学学研究 ,2014,32(3):321-326.

[3] 黄萃 , 苏竣 , 施丽萍 , 程啸天 . 政策工具视角的中国风能政策文本量化研究 [J]. 科学学研究 , 2011(06):876-882.

[4] 汪涛 , 安暄 . 类定量化科技政策文本分析框架构建及北京市科技政策演进分析 [J]. 技术经济 ,2011,30(6):15-17,34.

[5] 仲伟俊 , 蔡琦 . 科技政策分析框架研究 [J]. 科技管理研究 ,2014(22):23-27.

[6] 李素建 . 基于语义计算的语句相关度研究 [J]. 计算机工程与应用 ,2002,38(7):75-76.

[7] 唐琦 . 基于语义分析的句子相似度计算方法 [D]. 北京 : 华北电力大学 ,2008.

[8] 史燕 . 基于 HNC 的汉语句子相似度算法的研究 [D]. 镇江 : 江苏大学 ,2009.

[9] 李茹 , 王智强 , 李双红等 . 基于框架语义分析的汉语句子相似度计算 [J]. 计算机研究与发展 ,2013,50(8): 1728-1736.

[10] 李彬 , 刘挺 , 秦兵等 . 基于语义依存的汉语句子相似度计算 [J]. 计算机应用研究 ,2003(12):15 -18.

[11] 李春梅 , 徐庆生 . 基于多特征的汉语句子相似度计算模型的研究 [J]. 计算机技术与发展 ,2014,24(6) :136-141.

[12]Han J W, Kamber M, Pei J. Data Mining: Concept and Techniques. [M].2nd Edition Amsterdam, Holland: Elsevier, 2006.

[13] 徐戈 , 王厚峰 . 自然语言处理中主题模型的发展 . 计算机学报 , 2011, 34(8): 1423-1436.

[14]Sanchez J A, Medina M A, Starostenko O, et al. Organizing OpenArchives via Lightweight Ontolog to Facilitate the Use of Heterogeneous Collections[J]. Aslib Proceedings, 2012, 64(1): 46-66.

[15]Vicient C, Sanchez D, Moreno A. An Automatic Approach forOntology-Based Feature Extraction from Heterogeneous DocumentalResources[J]. Engineering Applications of Artificial Intelligence,2013, 26: 1092-1106

[16]Capelle M, Hogenboom F, Hogenboom A, et al.Semantic NewsRecommendation Using WordNet and Bing Similarities[C]//Proc. of the 28th Annual ACM Symposium on Applied Computing. Coimbra, Portugal, 2013: 296-302

[17]Bhagwani S, Satapathy S, Karnick H. Sranjans: Semantic TextualSimilarity Using Maximal Weighted Bipartite Graph Matching[C]// Proc. of the 1st Joint Conference on Lexical and ComputationalSemantics. Montreal, Canada, 2012: 579-585

[18]Wang J Z, Taylor W. Concept Forest: A New OntologyAssistedText Document Similarity MeasurementMethod[C]//Proc. of the IEEE/ WIC/ACM International Conference on Web Intelligence, Fremont, USA, 2007: 395-401

[19]Tsatsaronis G, Varlamis I, Norvag K. SemaFor: Semantic Document Indexing Using Semantic Forests[C]//Proc. of the 21st ACM International Conference on Information and Knowledge Managemen, Maui, USA, 2012: 1692-1696.

[20] 黄萱菁 , 张奇 , 吴苑斌等 . 文本情感倾向分析 [J]. 中文信息学报 ,2011,25(6):118-126.

[21]Peter D Turney,Michael L Littman.Measuring praiseand criticism:Inference of semantic orientation fromassociation[J]. ACM Transactions on Information Systems, 2003, 21(4): 315-346.

[22]Kim,S M,E Hovy. Automatic Detection of OpinionBearing words and Sentences[C]// Companion Volume to the Proceedings of UCNLP-05, Jeiu Island,KR,2005:61-66.

[23]WILSON I WIEBEJ, HOFFMANN P. Recognizing Contextual Polarity:An Exploration of Features for Phrase-Level Sentiment Analysis[J]. Computational Linguistics,2009,35(3):399-433.

[24]PangBo,LeeL. Thumbsup-Sentimentclassific

action using machine learing techniques [C]// Procedings of the Conference on Empirical Methodsin NLP. Morristown: ACL,2002: 79-86.

[25] 崔彩霞 , 王素格 . 基于内类频率的文本分类特征选择方法 . 计算机工程与设计 ,2007,28(17)：4249-4251.

[26] 王素格 , 魏英杰 . 停用词表对中文文本情感分类的影响 . 情报学报 ,2008,27(2):175-179.

[27] 朱嫣岚 , 闵锦 , 周雅倩等 . 基于 HowNet 的词汇语义倾向计算 [J]. 中文信息学报 ,2006,20(1)：14-20.

[28] 许歆艺 , 刘功申 . 基于文本纹理特征的中文情感倾向性分类 [J]. 中文信息学报 ,2015,29(3)：106-112.

[29] 郝玫 , 王道平 . 中文网络评论的复杂语义倾向性计算方法研究 [J]. 图书情报工作 ,2014,58(22)：105-110.