一种领域语料驱动的句子相关性计算方法研究

李 峰[1,2] 黄金柱[3] 李舟军[1] 杨伟铭[2]

(北京航空航天大学软件开发环境国家重点实验室 北京 100191)[1]

(中国人民解放军后勤科学研究所 北京 100166)[2] (中国人民解放军外国语学院语言工程系 洛阳 471003)[3]

摘 要 句子相关性计算在自然语言处理的多个实践应用中均具有十分重要的作用,如舆情监测、信息检索、统计机器翻译等。在明确相似性与相关性之间的关系之后,设计了一种基于领域语料驱动的句子相关性计算方法,该方法基于同一领域的语料构建一个"句-段-篇"3层的领域语义空间,通过度量词语在各个层级间的共现概率、共现平均距离和句长等因子来测量词间的主题相关性。与基于字面特征、HowNet和同义词词林的方法进行了实验对比,结果表明该方法具有较好的实践应用价值。

关键词 句子相关度,语料驱动,主题相关性,计算模型

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.5.034

Study on Domain-corpus Driven Calculation Method of Sentence Relevance

LI Feng[1,2] HUANG Jin-zhu[3] LI Zhou-jun[1] YANG Wei-ming[2]

(State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China)[1]

(Logistics Science Research Institute of PLA, Beijing 100166, China)[2]

(Department of Language Engineering, PLA University of Foreign Languages, Luoyang 471003, China)[3]

Abstract Sentence relevance calculation plays a very important role in various fields of NLP, such as public opinion monitoring, information retrieval and statistical machine translation(SMT) etc. This paper, after a clear definition of relationship between similarity and relevance, designed a domain-specific corpus-driven calculation model of sentence relevance. The model applies the linguistic data of the same domain to construct a "sentence-paragraph-article" three-level domanial semantic space. The topic relevance of words can be figured out through calculating different factors of various levels such as co-occurrence probability, co-occurrence average distance and sentence length etc. The paper made comparative experiments between the model and methods based on literal features, HowNet and Tongyici Cilin respectively and the results show that this model has great practical value.

Keywords Sentence relevance, Corpus driven, Topic relevance, Calculation model

1 引言

在自然语言处理中,通常将使用句号、问号或惊叹号隔开的文本片段视为句子,同时文本中句子相关性的计算在语言信息处理及相关领域中得到了广泛的应用,并成为许多方法的研究基础。例如,在话题检测研究中通过句子相关性计算进行话题的识别与划分[1,2];在信息检索领域中根据句子的相关性返回检索结果[3];在自动问答系统中通过句间的相关性计算进行问题与答案的匹配[4];在文本自动摘要领域中,句子相关性计算能够为主题判断、冗余信息消除等提供依据[5]。

当前,常见的句子相关性计算方法可分为两类,一种是基于统计的方法,一种是基于语言资源的方法。基于统计的方法,如基于句子间词的重叠数量或句子向量的夹角余弦的方法,其优势是构建成本相对简单,并具有广泛的适应性,没有语种依赖,可以应用于多个场景;缺点是计算的精度较差,受统计对象的数据规模和平衡性影响较大;基于语言资源的方法,如基于句法结构、同义词词林或HowNet等语言资源的方法,这类方法的优势是相关性计算较为精准,尤其表现在低概率词句间的相关性计算上;缺点是前期语言资源构建成本较高,数据资源建成后更新缓慢,对于专业领域数据覆盖面较差,且具有较强的语种依赖性。

自然语言处理领域的研究往往是任务驱动的,例如商品评论情感倾向性判断、微博话题分析、体育新闻推荐、政治舆情分析等,这些具体的应用往往具有一定的领域倾向性,语言风格不一、专业术语较多,同一词词义也可能因领域不同而差别巨大。例如:评论句子"(数据线)内线质量杠杠的!银灰色更是高大上!",微博句子"钟汉良、孙菲菲曾主演内线",政治报道"通过内线电话进行了磋商",体育新闻"火箭赢球难掩内

到稿日期:2015-04-16 返修日期:2015-08-23 本文受国家自然科学基金项目(61170189,61370126),高等学校博士学科点专项科研基金(20111102130003),软件开发环境国家重点实验室自选课题(SKLSDE-2013ZX-19)资助。

李 峰(1982-),男,博士,工程师,主要研究方向为计算语言学、数据挖掘,E-mail:li_bopr@126.com;黄金柱(1982-),男,博士生,主要研究方向为自然语言处理、本体资源库建设;李舟军(1963-),男,博士,教授,主要研究方向为智能信息处理、网络安全;杨伟铭(1982-),男,硕士,工程师,主要研究方向为软件工程、大数据处理。

线隐患"4 个句子中，若使用常规的基于统计的方法时，除了会因词间重叠不多而计算精度不高，而且重叠的词间可能并无太大关系，如"内线"；而在使用基于语言资源的方法时，也会因语言资源难以覆盖专业性较强的术语或更新缓慢，而损失了计算的精度。鉴于此，本文基于现有的研究方法，提出并设计了一种基于领域语料驱动的句子相关性计算方法，该方法基于同一类型的语料构建一个"句-段-篇"3 层的领域语义空间，通过度量词语在各个层级间的共现概率、共现平均距离和句长等因子来测量词的相关性，该方法属于基于统计的方法，它不依赖于具体的语种，能够给出零字面重叠的句子间的相关性。下文首先明确了相似性与相关性的关系；随后概述了常见的句子相关性计算方法；然后提出了一种领域语料驱动的句子相关性计算方法，并进行了算法描述；最后与基于字面、《知网》(HowNet)和《同义词词林》的方法进行了详细的实验对比和分析，指出了本文所提方法的优势及适用场景，并明确了下一步研究的方向。

2 相关性与相似性

相关性与相似性是一对极为相近而又容易为人们所混淆的概念，在许多研究领域，出于习惯人们总把"相关"等同于"相似"，就自然语言处理领域而言，人们多以句子间的相似度计算来替代句子间的相关性计算，或者即使计算的是句子间的相关性却也描述为句子的相似性计算，这主要是因为许多学者对"相关"与"相似"两个概念存在模糊认识。在《现代汉语词典》中，对"相关"一词的解释是："彼此关连；互相牵涉"，强调的是"关联"。而对于"相似"一词，则解释为："相类；相像"，强调的是"像"。显然，"像"虽是一种密切的"关联"却不是"关联"的全部，"关联"还应该包含除了"像"以外的其它关系，如包含关系、对立关系等。例如，图 1 中左侧表示相似，而右侧则表示相关。

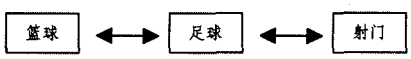

图 1 相似与相关实例

显然，就概念范畴而言，相关与相似虽有交集，却并不完全相等，因此研究句子的主题相关性，除了应该考察句子间字面重叠的比率(相似性)，还应该把非字面重叠的词纳入考虑(相关性)，才会使计算结果更加趋于合理。

3 典型的句子相关性计算方法

在自然语言处理领域中，许多学者常常把句子的相关性计算直接等同于句子的相似性计算，本文在梳理时把常见的句子相似度计算方法也纳入考虑，统一作为句子相关性计算的常见方法。归纳起来主要有 4 种：1)基于关键词信息；2)基于依存句法；3)基于语义信息；4)多特征融合。

3.1 基于关键词信息的方法

最为基本的是基于字面重叠的方法，它通过统计两个句子之间相同关键词的数目以及两个句子的关键词数目之和来度量句子相关度，计算公式为 $2c/(a+b)$，其中 a、b 是两个句子各自关键词数，c 是两个句子中相同关键词数。另一种方法是基于向量空间模型(Vector Space Model，VSM)的 TF-IDF 方法。这种方法假设所有句子中包含的词为 W_1，W_2，…，W_n，则每一个句子都可以用一个 N 维的向量 $T=(W_1, W_2, \cdots, W_n)$来表示。$T_i(1\leqslant i\leqslant n)$的计算方法为：

$$T_i = n \times \log \frac{m}{M} \tag{1}$$

其中，n 为 W_i 在这个句子中出现的个数，m 为其他所有句子中含有 W_i 的句子的个数，M 为句子的总数。在该方法中，任意两个句子间的相关度可以表示为其对应向量 T 与向量 T' 之间的夹角余弦值，计算公式为：

$$Similarity(T,T') = \frac{\sum_{i=1}^{n} T_i \times T_i'}{\sqrt{(\sum_{i=1}^{n} T_i^2) \times (\sum_{i=1}^{n} T_i'^2)}} \tag{2}$$

基于关键词信息的方法对句子间的字面重叠有着高度的依赖，例如 3 个短句："卫星发射中心"、"火箭常规赛事"和"火箭发射成功"，采用这两种方法，3 个句子两两间拥有相同的相关度，而事实并非如此。同时基于关键词的方法仅考虑了词的表层统计特征，没有涉及句子深层次的分析，如结构、语义等，因此具有较大的局限性。

3.2 基于依存句法的计算方法

依存句法是由法国语言学家 L. Tesniere 在其著作《结构句法基础》(1959 年)中提出的。依存文法通过分析语言单位内成分之间的依存关系揭示其句法结构，主张句子中动词是支配其他成分的中心成分，而它本身却不受其他任何成分的支配，所有受支配成分都以某种依存关系从属于支配者[6,7]。依存结构的计算方法仅考虑那些有效搭配对之间的相似程度。所谓有效搭配对是指全句核心词和直接依存于它的有效词组成的搭配对，这里有效词定义为动词、名词以及形容词。因而，使用这种方法的关键在于找出句子各成分间的依存关系，其相应的计算公式如下：

$$Similarity(s_1, s_2) = \frac{\sum_{i=1}^{n} W_i}{\mathrm{Max}\{PairCount_1, PairCount_2\}} \tag{3}$$

其中，$\sum_{i=1}^{n} W_i$ 为句子 1 和句子 2 有效搭配对匹配的总权重，$PairCount_1$、$PairCount_2$ 为两个句子有效搭配对数。这种方法从句子深层结构进行分析，考虑到了词之间的深层关系，对句子的理解更进一步，但这种方法对句子各单元之间的语义相关性考虑不足，同时受限于当前句法分析技术，因此其应用范围和效果都十分有限。

3.3 基于语言资源的计算方法

基于语言资源的句子相关性计算方法目前多为基于语义词典或领域本体的方法，其通过词语间的语义距离获取词语间的语义相关度，然后求和得到句子的整体相关性值。在面向中文的语言信息处理中，多以《知网》(HowNet)、《同义词词林》为语言计算资源[8,9]，面向英文的句子相似度计算多采用 WordNet 为语言计算资源[10]，在其它类别语言处理中，有的采用了词典作为语言计算资源或对应的本体知识库等[11,12]。

该类方法考虑了句子中词语的同义、近义等语义关系，在一定应用场景下能够取得较好的效果。但这类方法多限于某个具体的语种，资源更新缓慢，难以应用于专业的领域，且难以免费获取，运用技术难度也较大。

3.4 多特征融合的计算方法

为了取得更为令人满意的效果，许多研究者使用融合的方法来计算句子的相关性，这种方法往往会考虑句子的诸多特征，如语句的长度、词序、语义距离等[13]。在一些应用领域

内，这种方法可以等同于基于融合的句子相似度计算。如，中国石油大学张培颖在文献[14]中给出了一种多特征融合的语句相似度计算模型，文章综合考虑了句子的词形、词序、结构、长度、距离和语义这 6 种特征，实验结果表明这种方法的准确率要优于基于向量空间模型的 TF-IDF 方法和语义依存方法。

多策略融合的方法能够综合考虑句子的多项特征，较为全面地反映句子所包含主体信息的相关性。但这种方法也存在一些不足，例如各种特征参数值的设定、方法的计算复杂度等。

另外一些句子相关性计算方法，如编辑距离（又称 Levenshtein距离）[15]、浅层语义索引（Latent Semantic Indexing）[16]、语义角色标注[17]以及上述各种方法的改进与变形[18,19]，由于在效果上并没有太多的提升或者应用代价较高而使用较少，本文不再一一赘述。

4 领域语料驱动的句子相关性计算方法

4.1 相关假设

当人们需要把握一个词的准确含义时，总是要看其所在的语境（句子、段落或篇章上下文），而这些语境的形成又由表达主题的各个意义单元组成。换言之，文本中的某个意义单元的具体含义总是受其所在语境约束，而语境又因各语义单元间的相互作用而存在。据此，本文作出如下假设：词间的相关性强弱与其在句子、段落、篇章 3 个层面的共现频率与平均空间距离有关；句子间相关性的强弱正比于实词间的相关性强弱。

例如，在句子“北约将在地中海北岸盟国部署几十架战斗轰炸机、加油机、侦察机和无人驾驶飞机”中，若同等条件下，“侦察机”与“加油机”的相对共现频率高于“侦察机”与“部署”，可认为前者相关度高于后者；若同等条件下，“侦察机”与“加油机”在篇章中共现，而“侦察机”和“无人驾驶飞机”在句子中共现，则可认为后者相关度高于前者。

类似地，可以把句子间的相关性计算转化为组成句子的词语间的相关性计算，计算公式为：

$$Relativity(S_i,S_j)=\sum_{k=0}^{MinLen(S_i,S_j)} Relativity(W_i,W_j)/MaxLen(S_i,S_j) \tag{4}$$

上述公式考虑了句子的长度特征，其中 $MinLen(S_i,S_j)$，$MaxLen(S_i,S_j)$分别对应句子 S_i，S_j 中短句子与长句子的长度，$Relativity(W_i,W_j)$则表示 S_i，S_j 中对应实词的相关性值。

4.2 “句-段-篇”三层领域空间的构造

依据上述假设，基于领域语料库构建了“句-段-篇”3 个层面的领域空间，通过测量词在不同层面的共现频率与平均空间距离来计算词间的相关性。所构造的三层领域空间如图 2 所示。

该空间包含多篇文本语料，分为篇章、段落和句子 3 层。在篇章层，每一篇文本作为一个独立的要素，互不关联；在段落和句子层，每段或每一句作为一个独立的要素，并都有一个唯一的篇章父层。在“句-段-篇”空间中，每个词可以在多个句子、段落或篇章中出现，并拥有不同的词性标记，词与词间通过字面与词性区分合并。如图 2 所示，句子 n 与句子 l 处于同一段落 j，且包含相同的词 q，则句子 n 与句子 l 间的相关性要大于与段落 i 中句子 m 的相关性。在实际的算法设计中，除了需要考虑两个句子中所有实词在整个语义空间中各个层级间的共现概率外，还需要考虑其平均空间距离。

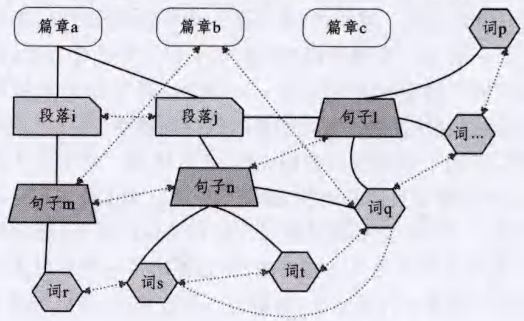

图 2 “句-段-篇”三层领域空间示意图

为了验证假设的可行性，并方便后续的实验设计，采集了新浪军事新闻频道[20]中 8500 份军事新闻作为构建领域语义空间的语料，使用微软 MSSQL2008 数据库存储相关数据。在构建过程中，使用 NLPIR2015[21]中文分词组件进行中文的分词与标注，并过滤常见的停用词；使用倒排索引算法分别以词、句子和段落为索引项构建“句-段-篇”三层领域空间，在构建过程中使用数据库索引压缩技术提升检索与使用效率。

4.3 句子相关度计算方法算法描述

基于统计的方法通常把句子相关性计算转化为实词间的相关性计算来得到相关度值，较少考虑不同语境下词义的变化。本文亦通过考察句子间实词的语义相关性来计算句子的相关度，不同的是，为提高实词语义相关性计算的准度，构建并引入“句-段-篇”三层领域空间来测量实词在不同级别的共现距离与频度等信息，强化语境对实词实际意义的约束，以期提高句子相关度计算的准确性。

设句子 S_i 与 S_j 之间的相关性为 R_s，其所包含的实词分别为 $W_i(w_{i1},w_{i2},\cdots,w_{im})$和 $W_j(w_{j1},w_{j2},\cdots,w_{jn})$，设 S_i 与 S_j 中词 w_{im} 与 w_{jn} 之间的相关性为 R_w。本文使用线性求和以及引入参数的方法来计算整个句子的相关性，公式如下：

$$R_w=\alpha\times R_{sent}+\beta\times R_{para}+\gamma\times R_{txt} \tag{5}$$

其中，

$$R_{sent}=\lambda\times\frac{2c}{a+b}+(1-\lambda)\times\left(1-\frac{\sum_{i=1}^{c}D(w_{im},w_{jn})}{c}\right) \tag{6}$$

$$R_{para}=R_{txt}=\frac{2c}{a+b} \tag{7}$$

上述公式中 R_{sent}、R_{para}、R_{txt} 表示词 w_{im} 与 w_{jn} 在句子、段落、篇章级空间中的相关程度，α、β、γ 为权重调节因子，三者之和为 1，且 $\alpha>\beta>\gamma$。在 R_{sent} 的计算中，c 表示同时包含 w_{im} 与 w_{jn} 的句子总个数，a 和 b 分别表示包含 w_{im} 和 w_{jn} 的句子个数，$\sum_{i=1}^{c}D(w_{im},w_{jn})/c$ 表示两个词在句子层面的平均空间距离，λ 为权重调节因子，且 $\lambda>0$；R_{para}、R_{txt} 中参数的意义与 R_{sent} 中的参数类似，区别在于统计计算的范围有差异。主要算法流程如下。

(1)对输入句 s_1，s_2 进行分词与标注处理，去除单个字、停用词及虚词，仅保留形容词、名词、动词。

(2)统计 s_1，s_2 中的实词数，令 s_1，s_2 实词数分别为 m，n。

(3)初始化矩阵 $W(m,n)$，计算 s_1，s_2 中任意两个实词的相关度，并填充到矩阵中。若两实词字面相同，直接填充值

1，若不同，则按上述词语相关度计算公式计算。

(4)对 W 中的值进行从高到低排序，并按 s_1，s_2 中最短实词长度进行遍历：从 W 中取出当前最大相关度值〈$Word(s_1)$，$Word(s_2)$，$RValue$〉放入 $FinalList$ 列表中；删除短句子中对应的 $word$；当短句子中所有 $word$ 完全删除后停止遍历。

(5)最终句子 s_1、s_2 的相关度值 R_s 等于 $FinalList$ 中所有 $RValue$ 之和除以较长句子的实词个数。

算法记录了“相同字面”实词的相关度值，通过领域语料构建的“句-段-篇”三层空间测量了“不同字面”实词之间的相关度，同时通过实词个数融入了句长特征，因此其属于基于统计的多特征融合的句子相关性计算方法。

5 相关实验与结果分析

5.1 评价方法

为检验算法的可行性以及算法的实际效果，设计了评测实验。目前，句子相关性计算的评价方式主要有两种：1)采用人工判别的方法，即使用相关算法计算一组句子的相关性，然后依靠人的认知来评价输出结果；2)将句子相关性计算应用到某个实践应用领域，观察在其他指标不变的情况下，其对目标应用的改进效果。前者主观性较强，受测评人员的知识与认知水平影响较大；后者则相对客观，但目标应用的效果必须与句子相关性计算强相关。本文采用了后一种方法，通过句子相关性计算对文本摘要质量的改进来评价方法的效果，为突出句子相关性对摘要质量的影响，对文本摘要进行了简化处理，并选取基于字面、《知网》(HowNet)和《同义词词林》3种句子相关性计算方法进行实验对比。

5.2 实验设计与分析

实验采用了两部分语料共 100 份作为测试数据，其中 50 份来自新浪军事频道国际军情栏目[22]，其它 50 份是从凤凰网新闻频道[23]随机抓取的其它领域新闻。为降低实验的复杂度，本文对自动摘要处理过程进行如下简化：

首先，为确保实验数据具有较高的参考价值，组织几名硕士生人工为每篇新闻生成一份标准摘要，要求组成标准摘要的句子必须与标题密切相关，且能在原文中找到；在抽取过程中不考虑文本或句子的其它特征，如段落首句、是否包含专有名词、日期数字等；抽取比率为原文句子数的 30%，个数向上取整。

其次，使用本文所提方法以及选取的另外 3 种方法从原文中抽取与标题相关的句子，句子个数与标准摘要句子个数保持一致。在实验评价时，主要测量摘要中的句子与人工摘要句子的重合率。

在实验中，基于字面的句子相似度计算方法采用张培颖在文献[14]中使用的方法，其考虑了句子的词重叠、语序、连续词块长度、句子相对长度等特征，在本文中各特征值比重系数取值分别为 6/11，3/11/，1/11，1/11；基于《知网》(HowNet)的计算方法参考刘群、李素建在文献[8]中提供的方法；基于《同义词词林》的计算方法与数据来自于夏天的开源项目 X-Similarity[24,25]。在各种方法中，均只考虑句子中所包含的形容词、名词与动词，并取句子相关性的最大值。在本文所提的算法中，α、β、γ、λ 取值分别为 16/20、3.5/20、0.5/20、15/20。最终的实验结果分别如图 3 和图 4 所示。

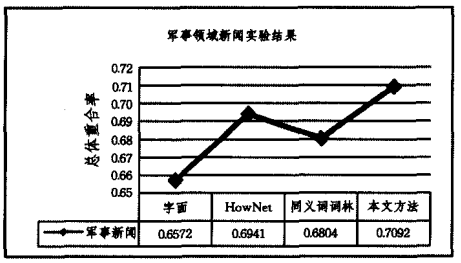

图 3 4 种句子相关性计算方法实验结果一

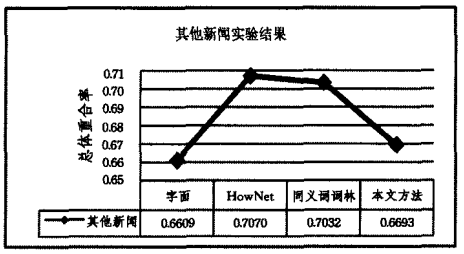

图 4 4 种句子相关性计算方法实验结果二

图 3 的实验结果表明，当处于同一受限领域时，领域语料驱动的句子相关性计算方法能够取得相对较好的实验结果，比基于字面的方法提升了约 5.2 个百分点，同时也取得了优于基于《知网》(HowNet)和《同义词词林》方法的结果。通过仔细分析发现，影响基于语言资源方法实验结果的因素主要有两个：1)专业领域词汇缺失，本实验中，部分军事词汇(如“无人机”、“火箭弹”)在《知网》(HowNet)或《同义词词林》中的缺失，导致了基于语义资源计算时词间的相关性计算被赋予了零分值；2)中文分词和标注问题，实验采用了 NLPIR2015 进行中文分词和标注，分词和标注结果与《知网》(HowNet)或《同义词词林》中的词及词性不统一，例如有些实词在《知网》(HowNet)中被视为虚词，导致在计算中存在一些偏差。

而当测试语料为非军事语料时，从图 4 中可以看出，本文所提出的方法与基于字面的方法取得的结果较为接近，同时效果略逊于基于《知网》(HowNet)与《同义词词林》的方法。这主要是因为，当缺少前文构建的“句-段-篇”三层领域空间的支撑时，领域语料驱动发挥不了作用，导致整个算法几乎等同于基于字面的相似度计算，而当有了该领域空间支撑时，本文所提方法能够在考虑句子间重叠字面的同时考虑非重叠字面的实词在领域空间中的关联程度。

为考察本文所提方法的实用性，在实验中记录了各种方法的计算效率，结果如图 5 所示。由于本文所提方法需要计算句子中实词在“句-段-篇”三层领域空间中的关联度，因此耗时最高，其他 3 种方法因计算简单或者预置了计算资源，并设计了较为良好的数据结构，计算效率相对较高。

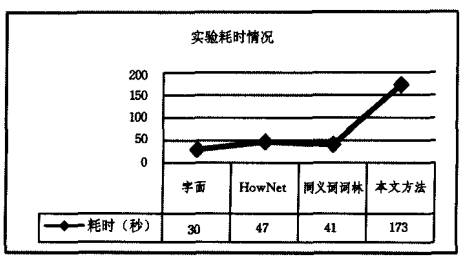

图 5 4 种方法的计算效率比较

5.3 实验结论

综上实验结果表明，本文设计的领域语料驱动的句子相关性计算方法主要具有如下优势及应用价值：

1)在同一专业领域中，本方法能够取得较其他常用方法更好的结果，并具有良好的适用性。自然语言处理实践多是面向专业领域的应用，因而本文所提方法具有较高实践应用价值。

2)本文从统计的视角提出了一种新的句子相关性计算方法，进一步丰富了句子相关性计算手段；同时该方法不依赖于具体的语言，因此能够为中、英以外的其他语种自然语言处理提供一种现实可行的句子相关性计算方案。

结束语 句子相关性计算在诸多自然语言处理应用中均具有十分关键的作用，本文在明确相似性与相关性之间关系之后，分析了典型的句子相关性计算方法，如基于关键词信息、依存句法、语言资源等多种方法；随后提出了一种领域语料驱动的句子相关性计算方法，该方法通过构建“句-段-篇”三层领域空间，并基于该空间来测量句子间词的相关性，从而计算句子的相关性值。实验表明，在同一专业领域，本文所提方法与基于字面特征、HowNet 和同义词词林的方法相比，取得了最好的实验结果，具有良好的实践应用价值；同时该方法不依赖于具体的语种语言资源，能够服务于中、英之外的其他语言信息处理。在后续的研究中，我们一方面要进行算法的数据结构设计，提高算法的处理效率，另一方面拟扩大语种的测试范围，如俄语、朝语等。

参 考 文 献

[1] Fei Yue, Hong Yi-hong, Yang Jian-wu. Handling Topic Drift for Topic Tracking in Microblogs [M]// Advances in Information Retrieval: Proceedings of 37th European Conference on IR Research. 2015: 477-488

[2] Zhou Gang, Zou Hong-cheng, Xiong Xiao-bing, et al. MB-Single-Pass: Microblog Topic Detection Based on Combined Similarity [J]. Computer Science, 2012, 39(10): 198-202(in Chinese)
周刚，邹鸿程，熊小兵，等. MB-SinglePass：基于组合相似度的微博话题检测[J]. 计算机科学，2012，39(10)：198-202

[3] Dan Ş, Rajendra B, Vasile R. A Sentence Similarity Method Based on Chunking and Information Content [M]// Computational Linguistics and Intelligent Text Processing: Proceedings of 15th International Conference on CICLing 2014. 2014: 442-453

[4] Chung-Hsien W, Chao-Hong L, Po-Hsun S. Sentence extraction with topic modeling for question-answer pair generation[J]. Soft Computing, 2015, 19(1): 39-46

[5] Ercan C, Igor K. Multi-document Summarization via Archetypal Analysis of the Content-graph Joint Model [J]. Knowledge and Information Systems, 2014, 41(3): 821-842

[6] Zhou Fang. Study and Application on Chinese Sentence Similarity Computation [D]. Zhengzhou: Henan University, 2005 (in Chinese
周航. 汉语句子相似度计算方法及其应用的研究[D]. 郑州：河南大学，2005

[7] *Li Bin, Liu Ting, Qin Bing, et al. Chinese Sentence Similarity* Computing Based on Semantic Dependency Relationship Analysis [J]. Applications Research of Computers, 2003, 20(12): 15-17

[8] Liu Qun, Li Su-jian. Calculation of lexical semantic similarity based on the "HowNet"[C]// *Proceedings of the 3rd Chinese* Lexical Semantics Workshop. Taipei, Taiwan, 2002: 59-76

[9] Tian Jiu-le, Zhao Wei. Words Similarity Algorithm Based on Tongyici Cilin in Semantic web Adaptive Learning System [J]. Journal of Jilin University (Information Science Edition), 2010, 28(6): 602-608(in Chinese)
田久乐，赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报(信息科学版)，2010，28(6)：602-608

[10] Mostafa A, Mahmoud S, Ehsan Y. Semantic similarity assessment of words using weighted WordNet [J]. International Journal of Machine Learning and Cybernetics, 2014, 5(3): 479-490

[11] Zhang Hui-chang. Chinese Text Similarity Matching Based on Domain Dictionary [D]. Qiandao: Shandong University, 2014 (in Chinese)
张会昌. 基于领域词典的中文文本相似度匹配[D]. 青岛：山东大学，2014

[12] Liu Hong-zhe. Ontology based Sentence Similarity Measurement [J]. Computer Science, 2013, 40(1): 251-256(in Chinese)
刘宏哲. 一种基于本体的句子相似度计算方法[J]. 计算机科学，2013，40(1)：251-256

[13] Li Su-jian. Research of Relevancy between Sentences Based on Semantic Computation [J]. Computer Engineering and Applications, 2002(7): 75-76(in Chinese)
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(7)：75-76

[14] Zhang Pei-ying. Model for sentence similarity computing based on multi-features combination [J]. Computer Engineering and Applications, 2010, 46(26): 136-137(in Chinese)
张培颖. 多特征融合的语句相似度计算模型[J]. 计算机工程与应用，2010，46(26)：136-137

[15] Liu Bao-yan, Lin Hong-fei, Zhao Jing. Chinese sentence similarity computing based on improved edit-distance and dependency grammar [J]. Computer Applications and Software, 2008, 25(7): 33-34(in Chinese)
刘宝艳，林鸿飞，赵晶. 基于改进编辑距离和依存文法的汉语句子相似度计算[J]. 计算机应用与软件，2008，25(7)：33-34

[16] Wang Dong, Wu Jun-hua. Based on LSI and Dictionary Text Semantic Similarity Algorithm [J]. Coal Technology, 2010, 29(12): 217-218(in Chinese)
王栋，吴军华. 基于 LSI 和词典的文本语义相似度算法[J]. 煤炭技术，2010，29(12)：217-218

[17] Wang Zhi-qing. Chinese Sentence Similarity Based on Semantic Role Labeling [D]. Beijing: Beijing University of Posts and Telecommunications, 2014(in Chinese)
王志青. 基于语义角色标注的句子相似度计算[D]. 北京：北京邮电大学，2014

[18] Vu H H, Villaneau J, Saïd F, et al. Sentence Similarity by Combining Explicit Semantic Analysis and Overlapping N-Grams [M]// Text, Speech and Dialogue: Proceedings of 17th International Conference on TSD 2014. 2014: 201-208

[19] Jiang Min, Xiao Shi-bin, Wang Hong-wei, et al. An Improved Word Similarity Computing Method Based on HowNet [J]. Journal of Chinese Information Processing, 2008, 22(5): 84-89 (in Chinese)

(下转第 208 页)

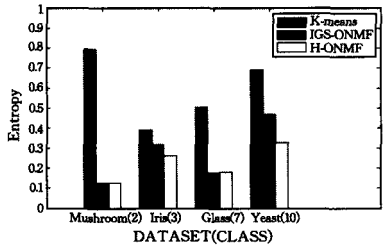

图 6　不同数据集下算法对应的 Entropy 值

图 5 和图 6 给出了基于 3 种算法在不确定程度和维度的 4 种数据集上进行聚类的结果。在数据维数较低的数据集 Mushroom 和 Iris 上，传统的 K-means 聚类算法在指标 Purity 上与本文中所提出的 IGS-ONMF 和 H-ONMF 下的 K-means 算法性能较为相似；但当数据的维数增加时，传统的 K-means 聚类算法的性能明显降低，而对于基于 IGS-ONMF 和 H-ONMF 的 K-means 算法来说，数据对应的维数越高，对应聚类算法的处理性能和鲁棒性就越高。H-ONMF 的 K-means 算法由于保证了数据的数值特征和完全正交性，能够在维数中等的数据集上有优越的聚类效果；而另一方面，IGS-ONMF 的 K-means 算法设置了一个权重系数用于控制正交性，在聚类过程中具有更强的适用性和鲁棒性，更适合维数较高的数据集。

结束语　本文在非负矩阵分解的基础上对高维数据进行处理，并将其数据原型矩阵分别通过改进的 Gram-Schmidt 正交化和 Householder 正交化加入了正交约束，由此提出了基于正交非负矩阵分解的高维数据 K-means 聚类算法。对 IGS-ONMF 和 H-ONMF 两种算法在真实和模拟数据集下进行了 K-means 聚类比较，实验结果表明，两种算法在一定程度上改善了高维数据下聚类的纯度和熵，具有较强的适应性和鲁棒性。

参 考 文 献

[1] Seung H S, Lee D D. Learning the parts of objects by non-negative matrix factorization[J]. Nature, 1999, 401(6755): 788-791

[2] Chu M, Plemmons R J. Nonnegative matrix factorization and applications[J]. Image, 2005, 34: 1-5

[3] Liu Wei-xiang, Zheng Nan-ning, You Qu-bo. Non negative matrix factorization and its application in pattern recognition[J]. Chinese Science Bulletin, 2006, 51(3)(in Chinese)
刘维湘，郑南宁，游屈波. 非负矩阵分解及其在模式识别中的应用[J]. 科学通报，2006，51(3)

[4] Tang J, Ceng X, Peng B. New Methods of Data Clustering and Classification Based on NMF[C]// 2011 International Conference on Business Computing and Global Informatization. IEEE Computer Society, 2011: 432-435

[5] Cichocki A, Amari S I, Zdunek R, et al. Extended SMART Algorithms for Non-negative Matrix Factorization[M]// Artificial Intelligence and Soft Computing(ICAISC 2006). Springer Berlin Heidelberg, 2006

[6] Ma Hui-fang, Zhao Wei-zhong, Tan Qing, et al. Orthogonal Nonnegative Matrix Tri-factorization for Semi-supervised Document Co-clustering[C]// Proceedings of the 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2010). 2010: 189-200

[7] Chen Y H, Wang L J, Dong M. Semi-supervised Document Clustering with Simultaneous Text Representation and Categorization[C]// Buntine W, Grobelnik M, Mladeni D, eds. Machine Learning and Knowledge Discovery in Databases: Lecture Nates in Computer Science. Springer, Heidelberg. 2009: 211-226

[8] Stadlthanner K, Theis F J, Puntonet C G, et al. Extended sparse nonnegative matrix factorization[M]// Computational Intelligence and Bioinspired Systems. Springer Berlin Heidelberg, 2005: 249-256

[9] Lee D D, Seung H S. Algorithms for Non-negative Matrix Factorization[J]. Nips, 2000, 32(6): 556-562

[10] Ding C, Tao L I, Peng W, et al. Orthogonal Nonnegative Matrix Tri-Factorizations For Clustering[M]// Sigkdd. 2006: 126-135

[11] Li H, Adal T, Wang W, et al. Non-negative Matrix Factorization with Orthogonality Constraints and its Application to Raman Spectroscopy[J]. Journal of Vlsi Signal Processing Systems for Signal Image & Video Technology, 2007, 48(1/2): 83-97

[12] Cao B, Shen D, Sun J T, et al. Detect and Track Latent Factors with Online Nonnegative Matrix Factorization[C]// International Joint Conference on Artificial Intelligence. 2007: 2689-2694

[13] Redko I, Bennani Y, Redko I. Controlling orthogonality constraints for better NMF clustering[C]// 2014 International Joint Conference on Neural Networks (IJCNN). IEEE, 2014: 3894-3900

[14] Li Le, Zhang Yu-jin. A Survey on Algorithms of Non-Negative Matrix Factorization[J]. Acta Electronica Sinica, 2008, 36(4): 737-743(in Chinese)
李乐，章毓晋. 非负矩阵分解算法综述[J]. 电子学报，2008，36(4)：737-743

[15] Li B, Zhou G, Cichocki A. Two Efficient Algorithms for Approximately Orthogonal Nonnegative Matrix Factorization[J]. Signal Processing Letters IEEE, 2015, 22(7): 843-846

(上接第 192 页)

江敏，肖诗斌，王弘蔚，等. 一种改进的基于《知网》的词语语义相似度计算[J]. 中文信息学报，2008，22(5)：84-89

[20] 新浪军事新闻[EB/OL]. (2014-03)[2015-04-05]. http://mil.news.sina.com.cn

[21] NLPIR[EB/OL]. (2015-03-27)[2015-04-05]. http://www.nlpir.org

[22] 新浪国际军情新闻[EB/OL]. (2015-04)[2015-04]. http://roll.mil.news.sina.com.cn/col/gjjq/index.shtml

[23] 凤凰网新闻频道[EB/OL]. (2015) [2015-04]. http://news.ifeng.com

[24] Xia Tian. Study on Chinese Words Semantic Similarity Computation[J]. Computer Engineering, 2007, 33(6): 191-194(in Chinese)
夏天. 汉语词语语义相似度计算研究[J]. 计算机工程，2007，33(6)：191-194

[25] Xia Tian. X-Similarity[EB/OL]. (2014-08) [2015-04-07]. http://code.google.com/p/xsimilarity