量化词的领域特征

刘冬明[1]，杨尔弘[2]

（1.中北大学，山西省太原市 030051；2. 北京语言大学，北京市 100083）

摘要：词作为最小的语义单位，同领域之间具有复杂的关系，特别是较为常用的词，通常难以明确界定其所属领域。在某些应用中并非必须确定词和领域的明确关系，仅仅依赖词的领域性的量化值就能够取得较好的效果。本文根据大规模语料库中词的关联信息，采用无指导的方法，对词的领域性进行量化，其结果可以作为词的一种特征应用于文本分类、话题检测、信息检索等相关的自然语言处理中。最后，通过和常用的特征——TF*IDF 在话题检测应用中进行对比，证明了其有效性。

关键词：词的领域性；话题检测；TF*IDF

中图分类号：TP391　　**文献标识码**：A

Quantize the domain property of the word

LIU Dongming[1], YANG Erhong[2]

(1.North University of China ,Taiyuan ,Shanxi　030051, China ; 2.Beijing Language and Culture University, Beijing　100083 ,China)

Abstract: Word, as the smallest semantic unit, has complex relationship with domain. Especially, it is often difficult for the more commonly used word to be clearly defined its domain belongs . But it is not necessary to establish a clear relationship between the word and the domain in some applications. We can achieve better results only relying on the quantized value of the domain property of the word. In this paper, we propose an unsupervised method for quantifying the domain property of words, based on word association information in the large-scale corpus. The results can be applied as the feature of the word to natural language processing such as text categorization, topic detection, information retrieval, etc. Finally, we compare the quantized value of the domain property of the word with common characteristics - TF * IDF in the topic detection application, and prove its effectiveness.

Key words: the domain property of the word; topic detection; TF * IDF

1 引言

随着自然语言可计算性研究的发展，在词汇级别上，人们不再满足于仅仅在语法层次上的处理，当前更加注重于语义及语用层次上的处理。在不同的表述领域，人们使用的词不尽相同，也即词在语义和语用层次上具有领域特征。如果能够在信息处理中有效的定义、描述和使用这种特征，那么对于当前许多自然语言应用的研究都具有重要的意义，如信息检索、话题检测、文本分类、自动文摘等等。

国内外研究人员构建了众多的以词为基本语义单位的知识库，如 WordNet[1]、HowNet[2]等等，这些知识库大多致力于语义角度对词形进行详细的可计算性的描述，如果要从中获取词的可计算性领域特征比较困难。由于领域本身是一个较为模糊的概念，领域之间存在着包含关系、交叉关系等，不可能采用一个清晰的结构来刻画所有的领域，因此词和领域的关系就更为模糊和复杂。还有许多研究人员开发了各种各样的算法来确定词的所属领域，这类研究的一个前提便是有一个确定的领域结构体系——一般来讲是一个元素各自独立领域集合，并且假设每个要作为研究对象的词只能属于其中一个领域，然后通过各个领域的文本样本，采用有指导的机器学习算法、自举等方法，充分利用具有明显领域属性的词的边界特征、内部结构特征等进行领域词的抽取，提取的结果可以有效的应用于后续的文本挖掘中，如文本分类、自动文摘等等[3][4][5][6]。这类研究的一个明显的局限性在于领域结构的划分，其抽取

结果只能应用于研究开始设定的领域结构体系，因此这类研究对于专业领域的深层次信息挖掘具有重要意义，但是如果应用于领域特性较差的语料中，如大众媒体，则由于领域的模糊性，不可能具有好的效果。

本文致力于量化词的领域特性，并不把词归于某一个特定的领域，而是在不假定领域体系结构的前提下，给予每一个词的一个具体的可比较的数值，表明该词的领域特性。其值可以作为该词固有的特征，应用于话题检测、文本分类、自动文摘等等自然语言处理领域，同时也能够作为一个重要的特征应用于上述特定领域词抽取研究中。

下面第二节介绍词的领域性度量的研究思路，第三节详细词的领域度的获取方法，第四节进行实验结果分析，最后是总结和展望。

2 研究思路

词的领域特性，即词和领域之间关系的性质。通俗来讲，即描述某一领域是否会用到该词，或者该词的出现是说明了文本正在描述某一领域。有些词语的领域性较强，如文本中出现“积分”，极大概率说明该文本描述数学领域，而另外一些词如“变量”，可以出现在数学领域也可以出现在计算机领域，还有一些词如“变化”，几乎所有的领域都会出现。本文所指词的领域特性的度量，定义为从词本身可以体现的领域特性程度。以上例子可以用如下方式形式化表示：

f（“积分”）＞ f（“变量”）＞ f（“变化”）　　其中 f 是一个自变量为词的实函数。

本文的目标就在于实例化 f——词和实数的映射关系。结合人类的认知层面，通常人工进行文本分类、信息检索、话题检测、关键词提取等过程中，明显会注重于 f 值较大的词，而忽略 f 值较小的词，当然不同的人根据掌握知识程度的不同，这个函数取值会不同，但是只要不涉及到个人的专业领域，那么大多数人的认识还是较为一致的。设想如果在机器自动文本分类、信息检索、话题检测、关键词提取等过程中，存在这一函数，那么必定有助于效率和效果。特别是当前这类应用中通常将文本作为“词袋”看待，这一函数将更加有助于降低噪音，获取更加准确的文本特征。

人类掌握这一函数依靠自己本身的经验和知识，并不存在明确的学习过程。这些经验通常是哪些词经常在一起描述某一领域，即便某个词自己并不知道其准确的定义，也能通过经常关联的词联系到某一领域。根据这一特性，本文从大规模语料库中获取词之间的关联关系，领域性强的词通常和它同现的词不多，并且互信息较大；而领域性弱的词，即在众多领域都出现的词，同现词较多且和每个同现词的互信息较小。因此，本文获得关联关系之后，利用这一特征计算出映射关系 f。

3 领域特征的量化方法

3．1 关联关系的获取

直观来看，词的关联关系应该源于词的领域同现关系。在不能明确划定领域的前提下，我们近似以词义的影响范围作为同现关系的提取范围。词的同现可以分为词的相邻同现、句子同现、段落同现和文章同现，相邻同现更多的反映出词的语法关系，而段落同现和文章同现往往会超出该词的意义影响范围，带来更多的噪音，句子作为具有完整意义的语言单位，可以近似的作为每个词的意义影响范围。

信息论中，互信息表达了一个事件的发生蕴含了关于另一个事件的发生的信息量的大小。同一领域中的词其同现的互信息值应较大，而不同领域的词之间的同现互信息值相对较小。因此，本文以词之间的互信息作为同现关系的度量值。

任一个词与其他词的句子同现互信息可以表示为一个向量，本文称之为关联向量，如词

$$w_i : (e_{i1}, e_{i2}, \_ e_{ik}, \ldots . e_{in}) \qquad w_i \in W \qquad \text{（公式 3.1）}$$

其中，e_{ik}表示词 w_i 和 w_k 的句子同现互信息值，W 为所有词构成的集合，n=|W|。这个向量实际上也就蕴含了关于词 w_i 的领域信息。如前所述，可以根据此向量获取该词的领域特征值。

3．2 词的不同之处——方差

如果一个词使用的领域范围有限，它仅仅和同它在同一领域的词的互信息较大，而和其他词的互信息较小，其关联向量中少数分量较大，大多分量较小甚至为 0；反之，则所有的分量较为平均。因此，关联向量的所有分量的方差可以近似的表示词的领域特征值。理论上，词 w_i 的方差 g

$$g(w_i)=\frac{\sum_{k=1}^{n}(e_{ik}-\bar{e_i})^2}{n} \quad \text{其中} \bar{e_i}=\frac{\sum_{k=1}^{n} e_{ik}}{n} \qquad \text{（公式 3.2）}$$

在实际的计算中，出于计算效率以及不同词的领域特征值数值上差异明显，我们采用如下公式计算：

$$g'(w_i)=\frac{1}{n_i'}+\left(\frac{(\sum_{k=1}^{n_i'} e_{ik}^2)}{n_i'}-\frac{(\sum_{k=1}^{n_i'} e_{ik})^2}{n_i'^2}\right) \qquad \text{（公式 3.3）}$$

其中n_i'为实际上与词 w_i 关联的词数，对大多数词来说n_i'要远小于 n。上式中的第二项可以从公式 3.2 推导得出，而第一项是为了弥补n_i'替换 n 之后损失的关于词 w_i 关联词数的信息。

3．3 基于图的领域特征转化

直观来看，领域同词之间根据关联关系所形成的团块结构密切相关。这就说明一个词领域特性不仅取决于单独针对这个词的关联关系的统计信息，而且同这个词具体关联哪些词有关。例如，对于词 w_i 和 w_j，$g'(w_i)=g'(w_j)$，如果同词 w_i 关联的词的领域性要强于同词 w_j 关联的词的领域性，那么合理的结果应该是 $f(w_i)>f(w_j)$。

基于以上推断，很自然的想到了图和迭代，具体方法如下：

词之间的关联关系也可以作为图的形式表示：G（W，E），其中 W 表示所有词的集合，E 表示词之间的关联关系，其值即为互信息。图的形式能够更加直观的体现词集的领域特征，团块结构明显的子集通常代表了一个领域。因此，为了更好地反映词的领域特征，本文利用 Google 基于图的排序算法 PageRank[7]，使词的领域特征值在$g'(w_i)$的基础上进一步强化其领域信息。

由于 PageRank 算法基于有向图，而词的互信息关联图是无向图，在此本文以$g'(w_i)$值来确定关联图中边的方向：

如果$g'(w_i)<g'(w_j)$，那么确定e_{ij}的方向为：$w_i \rightarrow w_j$

领域特征明显的词入边数大于出边数，反之则出边大于入边。这样在从无向图转化为有向图的过程中融合了前述基于方差的领域特征。

将e_{ij}的初始值置 1，采用 PageRank 算法迭代，最终获取了词的领域特征映射关系 f。

4 实验及结果分析

4．1 实验设计

词的领域特征值可以直接融合与各种应用，如自动文本分类、信息检索、话题检测、关键词提取、术语识别等等。简明起见，本文仅将其应用于话题检测，并与传统的词的特征提取方法 TF*IDF 对比，显示其效果。

实验中关联关系的获取来自 2012 年一月到十月的 18 份中文报刊，为了提高效率，分词之后根据 HowNet 中的词类标识仅提取其中的动词作为实验词集。评测语料下载自新浪网站中的专题板块，由于仅仅使用动词，因此人工标识了话题类别信息，分别为"出访"、"岛屿争端"、"航天"、"获奖"、"科技发布"、"枪击事件"、"事故"、"逝世"、"体育"、"选举"、"演唱会"、"娱乐婚恋"、"自然灾害"、"盗窃案件"、"金融",每一类别中随机抽取 50 篇文本作为测试语料库。

话题类别检测采用聚类工具 Cluto[8]，聚类方法采用 *k*-way clustering solution，相似度计算采用余弦相似度,准则函数定义为类内相似度最大,评测准则采用 Cluto 使用的 ***entropy*** 和 ***purity***[9]，详细定义如下：

对于聚类结果中的第*r*个聚类 S_r，设 n_r 为其包含的文档数量，那么该类的***entropy***定义为：

$$E(S_r) = -\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r} \qquad \text{（公式4.1）}$$

其中，*q*是测试语料包含的类别数，n_r^i是S_r中包含的第*i*类的文本数。整个聚类结果的***entropy***则定义为结果中每个聚类的加权平均值：

$$Entropy = \sum_{r=1}^{k}\frac{n_r}{n}E(S_r) \qquad \text{（公式4.2）}$$

其中，*k*为聚类结果的类别数，本文实验仅用于比较采用不同特征值的效果，因此在聚类参数中设定 *k*=*q*。 ***entropy***越小，则聚类效果越好，如果每一个聚类中仅仅包含一类文档，那么***entropy***值将为1。

对于聚类结果中第*r*个聚类S_r的***purity***定义为：

$$P(S_r) = \frac{1}{n_r}\max_i(n_r^i) \qquad \text{（公式4.3）}$$

其中参数的定义同式4.1。其意义在于用S_r中文档数最多的一类代表该聚类。同样，整个聚类结果的***purity***则定义为结果中每个聚类的加权平均值：

$$Purity = \sum_{r=1}^{k}\frac{n_r}{n}P(S_r) \qquad \text{（公式4.4）}$$

其中参数的定义同式4.2。从以上两式可以看到，***purity***值越大，那么聚类结果越好。

为了使结果更加直观，本文另外定义了一个综合指标 F-value：

$$\text{F-value}=\frac{purity}{1+entropy} \qquad \text{（公式 4.5）}$$

为了和 TF*IDF 对比效果，本文实验主要关注不同数量对比的两类话题检测，如从前述话题类别中抽取两类，分别指定不同数量文本进行聚类，下表为部分结果：

表 4-1：部分聚类结果

类别名称		文本数		***entropy***			***purity***			F-value		
1	2	1	2	TF*IDF	领域特征	综合	TF*IDF	领域特征	综合	TF*IDF	领域特征	综合
盗窃案件	金融	50	10	0.478	0	0.439	0.833	1	0.833	0.563599	1	0.578874
盗窃案件	金融	50	50	0	0	0	1	1	1	1	1	1
出访	岛屿争端	50	10	0.637	0.587	0.541	0.833	0.833	0.833	0.508858	0.52489	0.540558
出访	岛屿争端	50	50	0.429	0.567	0.166	0.9	0.82	0.97	0.629811	0.523293	0.831904

表中综合是指词的特征值以领域特征为基础在其上叠加上 TF*IDF 运算结果。

将所有的类别分别以数量 50：50 和 50：10 测试，其平均结果如下：

表 4-2：聚类结果各种指标的平均值

类别比例（不同类别试验次数）	***entropy***			***purity***			F-value		
	TF*IDF	领域特征	综合	TF*IDF	领域特征	综合	TF*IDF	领域特征	综合
50： 50（105）	0.0634	0.073619	0.053343	0.988695	0.986324	0.9908	0.937453	0.927219	0.946487
50： 10（210）	0.397557	0.258386	0.401224	0.850381	0.912043	0.8479	0.614948	0.755329	0.61096

4．2 结果分析

从上表可以看出在不同类别文本数量一致的情况下，各种特征选取的结果几乎相差不大，其中综合采用 TF*IDF 和领域特征所得的结果最好，当不同类别文本数量分布不均时，采用领域特征要比其余二者高了许多。

原因在于：

首先，TF*IDF 算法是建立在这样一个假设之上的：对区别文档最有意义的词语应该是那些在文档中出现频率高，而在整个文档集合的其他文档中出现频率少的词语，所以如果特征空间坐标系取 TF 词频作为测度，就可以体现同类文本的特点。另外考虑到词区别不同类别的能力，TFIDF 法认为一个词出现的文本频数越小，它区别不同类别文本的能力就越大。因此引入了逆文本频度 IDF 的概念，以 TF 和 IDF 的乘积作为特征空间坐标系的取值测度，并用它完成对权值 TF 的调整，调整权值的目的在于突出重要单词，抑制次要单词。但是在本质上 IDF 是一种试图抑制噪声的加权 ，并且单纯地认为文本频率小的单词就越重要，文本频率大的单词就越无用，显然这并不是完全正确的。例如：如果某类文本的数量占据了测试语料的绝大多数，那么其中本该作为特征的词因为 IDF 导致取值较小，使结果下降，这种情况从上面结果可以看出。而领域特征相对于测试语料为一恒定值，不会受到类别比例的影响。

再次，不论是 TF*IDF，还是当前流行的各种概率模型如 LSI、LDA，其获取特征完全来自于待测语料，没有知识库的支持，就如同考试仅仅根据考题特征，结合应试技巧来通过考试一样，难以获取实质性的进展。而领域特征来自于大规模训练语料库，提取的领域特征就相当于已有知识的简约表示，因此结果较好。

5 总结和展望

本文从领域和词的关系出发，提出了词的领域特征量化方法，明确指出了这种量化值在自动文本分类、信息检索、话题检测、关键词提取、术语识别等等研究领域的意义。同时以

简明的实验展示了其有效性。

本文将词的领域特征作为了一个可以比较的词的特有属性，其实通过基于句子同现的词的关联关系，还能够获得更为具体的关于词的领域知识，如哪些词可以代表一个领域，同时一个领域和另一领域的关联关系等等，这将是我们下一步研究的重点。

参考文献：

[1]George A.Miller. the WordNet project[DB].[2012]. http://wordnet.princeton.edu/

[2] 董振东, 董强. 知网[DB]. [2013]. http://www.keenage.com/.

[3] Fabrizio Sebastiani, Machine Learning in Automated Text Categorization[C]//ACM Computing Surveys (CSUR), March 2002, vol. 34, doi:10.1145/505282.505283.

[4] Navigli R, Faralli S, Soroa A, et al. Two birds with one stone: learning semantic models for Text Categorization and Word Sense Disambiguation[C] //Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011: 2317-2320.

[5] Gu H, Zhou K. Text classification based on domain ontology[J]. Journal of Communication and Computer, 2006, 3(5): 29-32.

[6]Reeve L H, Han H, Brooks A D. The use of domain-specific concepts in biomedical text summarization[J]. Information Processing & Management, 2007, 43(6): 1765-1776.

[7] S. Brin and L. Page. The anatomy of a large-scale hypertextual web searchengine[C] // In Proc. 7th International WWW Conference, 1998: pages 107－117.

[8] Karypis, George. CLUTO-a clustering toolkit[J]. No. TR-02-017. MINNESOTA UNIV MINNEAPOLIS DEPT OF COMPUTER SCIENCE, 2002.

[9] Ying Zhao and George Karypis. Criterion functions for document clustering: Experiments and analysis[C] //Technical Report TR #01－40, Department of Computer Science, University of Minnesota, Minneapolis, MN, 2001.

作者简介：刘冬明（1972——），男，讲师，主要研究领域为自然语言处理。Email:dmliutomorrow@gmail.com；杨尔弘（1965——），女，教授，主要研究领域为自然语言处理，计算语言学。Email:yerhong@126.com。