●庞景安（中国科学技术信息研究所　北京　100038）

Web 文本特征提取方法的研究与发展

摘　要：本文对当前有关 Web 文本特征提取方法的研究和试验进行了简要的综述和分析，比较了每类方法的优势和不足，指出研究中存在的难点和共同探讨的问题，并在此基础上，对该领域未来研究的发展方向和趋势进行了预测。

关键词：Web 文本；文本特征；特征提取；学习算法

Abstract: With a brief summary and analysis of the present research and experiment on Web text feature extraction, this paper compares the advantages and shortages of each method, points out the difficulties and problems in the research, and based on this, forecasts the future development and tendency of the research in this domain.

Keywords: Web text; text feature; feature extraction; learning algorithm

当前，因特网正在以前所未有的速度飞速发展，Web 已经发展成为拥有数以亿计页面的分布式信息空间，而且这个数字仍以每 4 至 6 个月翻一番的速度增加。人们迫切需要从这些海量的、异构的 Web 信息资源中，快速、有效地发现和利用有价值的知识和信息。

Web 文本挖掘就是从大量的 Web 文档中发现隐含知识和模式的一种方法和工具，它从数据挖掘发展而来，但与传统的数据挖掘又有许多不同。Web 文本挖掘的对象是海量、异构、分布的 Web 文档；文档内容是人类所使用的自然语言，缺乏计算机可理解的语义。传统数据挖掘所处理的数据是结构化的，而 Web 文档都是半结构或无结构的。所以，Web 文本挖掘面临的首要问题是如何在计算机中合理地表示文本，使之既要包含足够的信息以反映文本的特征，又不至于过于复杂使学习算法无法处理。

近年应用最多的 Web 文本特征表示方法是向量空间模型（VSM）。在该模型中，文档空间被看作是由一组正交词条向量所组成的向量空间，每个文档表示为其中的一个范化特征向量 $V(d) = (t_1 w_1(d); \cdots\cdots; t_i w_i(d); \cdots\cdots; t_n w_n(d))$。其中 t_i 为词条项，$w_i(d)$ 为 t_i 在 d 中的权值。可以将 d 中出现的所有单词作为 t_i，也可以要求 t_i 是 d 中出现的所有短语，从而提高文本特征表示的准确性。

利用向量空间模型（VSM）表示 Web 文档时，特征向量的维数经常会达到几十万，即使删除停用词和低频词，仍会有大量的特征留下。统计学、模式识别和机器学习中都有许多特征选择的方法，但是都不适用于 Web 文本挖掘，因为 Web 文本的特征数量实在太大。若特征数为 F，优化时要搜索的特征空间大小为 2^F，这样的计算复杂度是难以实现的。于是关于 Web 文本特征提取的研究就显得非常重要，成为 Web 文本挖掘进行的必要前提和基础。

1　基于评估函数的特征提取方法

这类型算法是在特征独立的假设基础上，通过构造评估函数，对特征集合中的每个特征进行独立评估，并对每个特征打分。然后将所有特征按分值大小排序，提取预定数目的最优特征作为提取结果的特征子集。显然，对于这类型算法，决定 Web 文本特征提取效果的主要因素是评估函数的质量。常用的评估函数有以下几种：

1）文档频数（Document Frequency）。

$$\text{DocFreq}(F) = P(W|C_i) = \frac{DF}{|C_i|}$$

2）信息增益（Information Gain）。

$$\text{InfGain}(F) = P(W)\sum_i P(C_i|W)\log\frac{P(C_i|W)}{P(C_i)} + P(\overline{W})\sum_i P(C_i|\overline{W})\log\frac{P(C_i|W)}{P(C_i)}$$

3）期望交叉熵（Expected Cross Entropy）。

$$\text{CrossEntryTxt}(F) = P(W)\sum_i P(C_i|W)\log\frac{P(C_i|W)}{P(C_i)}$$

4）互信息（Mutual Information）。

$$\text{MutualInforTxt}(F) = \sum_i P(C_i)\log\frac{P(W|C_i)}{P(W)}$$

5）文本证据权（The Weight of Evidence for Text）。

$$\text{WeightofEvidTxt}(F) = P(W)\sum_i P(C_i)\left|\log\frac{P(C_i|W)(1-P(C_i))}{P(C_i)(1-P(C_i|W))}\right|$$

6）优势率（Odds Ratio）。

$$\text{OddsRatio}(F) = \log \frac{P(W|pos)(1-P(W|neg))}{P(W|neg)(1-P(W|pos))}$$

7）词频（Word Frequency）。

$$\text{Freg}(F) = TF(W)$$

在上述公式中，F 为对应于单词 W 的特征，P（W）为单词 W 出现的概率，$\overline{W}$ 表示单词 W 并不出现，P（C_i）为第 i 类词的出现概率，P（$C_i|W$）为单词 W 出现时属于第 i 类的条件概率，P（W|pos）为在类 pos 中单词 W 出现的条件概率，P（W|neg）为在类 neg 中单词 W 出现的条件概率，TF（W）为单词 W 在文档集中出现的概率。

这些评估函数在 Web 文本挖掘中被广泛使用，特征选择精度普遍达到 70% ~ 80%，但也各自存在缺点和不足。例如，“信息增益”考虑了单词未发生的情况，对判断文本类别贡献不大，而且引入不必要的干扰，特别是在处理类分布和特征值分布高度不平衡的数据时选择精度下降。“期望交叉熵”与“信息增益”的唯一不同就是没有考虑单词未发生的情况，因此不论处理哪种数据集，它的特征选择精度都优于“信息增益”。与“期望交叉熵”相比，“互信息”没有考虑单词发生的频度，这是一个很大的缺点，造成“互信息”评估函数经常倾向于选择稀有单词。“文本证据权”是一种构造比较新颖的评估函数，它衡量一般类的概率和给定特征类的条件概率之间的差别，这样在文本处理中，就不需要计算 W 的所有可能值，而仅考虑 W 在文本中出现的情况。“优势率”不像前面所述的其他评估函数将所有类同等对待，它只关心目标类值，所以特别适用于二元分类器，可以尽可能多地识别正类，而不关心识别出负类。

从考虑文本类间相关性的角度，可以把常用的评估函数分为两类，即类间不相关的和类间相关的。“文档频数”（DF）是典型的类间不相关评估函数，DF 的排序标准是依据特征词在文档中出现篇数的百分比，或称为篇章覆盖率。这种类型的评估函数，为了提高区分度，要尽量寻找篇章覆盖率较高的特征词，但又要避免选择在各类文本中都多次出现的无意义高频词，因此类间不相关评估函数对停用词表的要求很高。但是，很难建立适用于多个类的停用词表，停用词不能选择太多，也不能选择太少，否则都将会影响特征词的选择。同时，类间不相关评估函数还存在一个明显的缺点，就是对于特征词有交叉的类别或特征相近的类别，选择的特征词会出现很多相似或相同的词条，造成在特定类别间的区分度下降。

类间相关的评估函数，例如期望交叉熵、互信息、文本证据权等，综合考虑了词条在已定义的所有类别中的出现情况，可以通过调整特征词的权重，选择出区分度更好的特征，在一定程度上提高了相近类别的区分度。但是，该区分度的提高仅体现在已定义的类别间，而对于尚未定义的域外类别，类间相关评估函数的选择效果也不理想。因此，在评估函数选择问题上，提高对域外类别文本的区分度是十分重要的研究课题。

许多学者对这一问题进行了研究和探讨。例如，基于相对词频的文本特征提取方法[1]，更加强调提高已定义的类别与域外类别的区分度，通过比较类别内与类别外的用词特点，选出更能反映该类别特征的特征词条，建立文档频数与相对词频结合的特征提取评估函数：

$$P(W|C_i) = \frac{DF}{|C_i|} \times \frac{a}{F_w}$$

其中 DF、C_i与前面定义相同，F_W为单词 W 在现代汉语中的使用频率，a 为归一化参数。这种评估方法实际是采用了一个柔性的停用词表，不仅仅挑选出现次数最多的特征词，而是利用相对词频，即特定词在某一类别中的词频与该词在现代汉语中常用词频的比值，选出对于该类别与常见类别有明显区别的特征词。这种方法将大大降低对停用词表的要求，甚至可以不用停用词表，自动删除高频无意义词汇。

另一种基于纯统计学无意义文本分解技术建立的文本特征提取方法是 N—Gram 算法。它的基本思想是将文本内容按字节流进行大小为 N 的滑动窗口操作，形成长度为 N 的字节片段序列。每个字节片段称为 gram，对全部 gram 的出现频度进行统计，并按照事先设定的阈值进行过滤，形成关键 gram 列表，即为该文本的特征向量空间，每一种 gram 则为特征向量维度。由于 N—Gram 算法可以避免汉语分词的障碍，所以在中文文本处理中具有较高的实用性。中文文本处理大多采用双字节进行分解，称之为 bi-gram。但是 bigram 切分方法在处理 20%左右的中文多字词时，往往产生语义和语序方面的偏差。而对于专业研究领域，多字词常常是文本的核心特征，处理错误会导致较大的负面影响。基于 N—Gram 改进的文本特征提取算法[2]，在进行 bigram 切分时，不仅统计 gram 的出现频度，而且还统计某个 gram 与其前邻 gram 的情况，并将其记录在 gram 关联矩阵中。对于那些连续出现频率大于事先设定阈值的，就将其合并成为多字特征词。这样通过统计与合并双字特征词，自动产生多字特征词，可以较好地弥补 N—Gram 算法在处理多字词方面的缺陷。

2 基于特征相关性的特征提取方法

基于评估函数的特征提取方法是建立在特征独立的假设基础上，但在实际中这个假设是很难成立的，因此需要考虑特征相关条件下的文本特征提取方法。

J．Pearl 提出马可夫条件集的概念[3]，对特征空间进

行后向搜索，删除那些当已知其他特征时，其所含类信息最少的无关特征。M. Singh 实现了一种前向特征选择算法[3]。初始集合 Y 为空，以“信息增益”作为评估函数，每次循环都将使 P（C|Y）和 P（C | Y∪ {X_i}）间的期望相对熵最大的特征加入。前向选择开始并没有特征，只是每次加入特征后试图使新的分布和原分布相差最远，但这不能保证一定是向正确的方向前进最大。相比之下，后向搜索可能更加优越，但困难的是马可夫条件集的寻找和建立。

支持向量机（Support Vector Machines，SVMs）是由 Vapnik 等人提出的一种基于统计学习理论的机器学习方法。Joachims 等人将 SVMs 应用于文本分类和特征提取研究中[4]，他们认为 SVMs 对于特征相关性和稀疏性不敏感，并且处理高维问题具有其他机器学习方法不可比拟的优势，不必利用评估函数进行特征选择，线形支持向量机就可以达到很好的分类效果。传统评估函数的特征提取方法独立地对每个特征评估打分，虽然可以选出各个类中的重要特征，但是却不能判断噪音特征和删除无效特征。基于支持向量的文本特征提取方法能够识别每个类别的重要特征和噪音特征。一个文本特征是不是噪音特征，可以由该特征在支持向量中的权值以及支持向量的性质决定，利用支持向量对文本特征的重要性进行评估。

3 基于遗传算法的特征提取方法

Web 文本实际上可以看作是由众多的特征词条构成的多维空间，而特征向量的选择就是多维空间中的寻优过程，因此在 Web 文本特征提取研究中可以使用高效寻优算法。遗传算法（Genetic Algorithm，GA）是一种通用型的优化搜索方法，它利用结构化的随机信息交换技术组合群体中各个结构中最好的生存因素，复制出最佳代码串，并使之一代一代地进化，最终获得满意的优化结果。在将 Web 文本特征提取问题转化为 Web 文本空间的寻优过程中，首先对 Web 文本空间进行遗传编码，以文本向量构成染色体，通过选择、交叉、变异等遗传操作，不断搜索问题域空间，使其不断得到进化，逐步得到 Web 文本的最优特征向量。

Hills[5]进一步把协同演化的思想引入到遗传算法中。基于协同演化的遗传算法不是使用固定的环境来评价个体，而是使用其他的个体来评价特定个体。个体优劣的标准不是其生存环境以外的事物，而是由在同一生存竞争环境中的其他个体来决定。协同演化的思想非常适合处理同类文本的特征提取问题。由于同一类别文本相互之间存在一定相关性，因而各自所代表的那组个体在进化过程中存在着同类之间的相互评价和竞争。因此，每个文本的特征向量，即该问题中的个体，在不断的进化过程中，不仅受到其母体（文本）的评价和制约，而且还受到种族中其他同类个体的指导。所以，基于协同演化的遗传算法不仅能反映其母体的特征，还能反映其他同类文本的共性，这样可以有效地解决同一主题众多文本的集体特征向量的提取问题，获得反映整个文本集合某些特征的最佳个体。

4 基于语义理解的特征提取方法

当前，关于 Web 文本特征提取方法的研究主要包括基于统计计算的方法和基于语义分析的方法。统计方法具有算法简单、易于实现、过滤速度快、不依赖具体领域和语言等优点，但由于缺乏对文档的语法语义分析，不能深层次地理解文本所表达的主题思想，因而很难取得较好的选择效果和系统性能。而基于概念理解的方法则更多地运用了自然语言理解、人工智能，以及语言学等方面的知识和技术，可以更深入地分析文档语法语义和主题思想，充分考虑语言中大量存在的同义和多义现象，以及褒贬倾向等在特征提取中起关键作用的因素，提高特征提取和文本过滤的精度。

1）基于语境框架的文本特征提取方法[6]。这是一种新的处理 Web 文本的语义形式化模型。语境框架是一个三维的语义描述，把文本内容抽象为领域（静态范畴）、情景（动态描述）、背景（褒贬、参照等）三个框架。在语境框架的基础上，从语义分析入手，实现了 4 元组表示的领域提取算法、以领域句类为核心的情景提取算法和以对象语义立场网络图为基础的褒贬判断。该方法可以有效地处理语言中的褒贬倾向、同义、多义等现象，表现出较好的特征提取能力。

2）基于本体论的文本提取方法[7]。应用本体论（Ontology）模型可以有效地解决特定领域知识的描述问题。具体针对数字图像领域的文本特征提取，通过构建文本结构树，给出特征权值的计算公式。算法充分考虑特征词的位置以及相互之间关系的分析，利用特征词统领长度的概念和计算方法，能够更准确地进行特征词权值的计算和文本特征的提取。

3）基于知网的概念特征提取方法[8]。对于 Web 文本的处理，尤其是中文文本处理，字、词、短语等特征项是处理的主要对象。但是字、词、短语更多体现的是文档的词汇信息，而不是它的语义信息，因而无法准确表达文档的内容；大多数关于文本特征提取的研究方法只偏重考虑特征发生的概率和所处的位置，而缺乏语义方面的分析；向量空间模型最基本的假设是各个分量间正交，但作为分量的词汇间存在很大的相关性，无法满足模型的假设。

（下转第 367 页）

中的知识资源并创造新的知识。当社区成员向知识库中存储知识后，系统会自动触发一个事件，并通过电子邮件系统把它传给每个指定的成员。通知的发出取决于早期的系统记录，这样就可以保证通知发给了真正感兴趣的成员，以协助他们利用新知识。否则，这个新知识对其他成员是没有价值的，系统就不会把它传递给这些成员。

对于监督者或者管理员来说，其主要职责是对系统中的知识资源进行甄别和组织。为了监管知识社区中知识传播与利用的质量，他们可以随时进入知识系统，对新加入社区的知识进行组织和甄别[5]。他们是社区中某方面知识专家或者权威，在知识社区中监督者可以找专人专管，也可以由社区成员兼顾管理。在进行知识监管中，为了保证安全，系统会给每个知识监督者以不同的管理权限。

3 本解决方案的成功之处及其亟待解决的问题

本解决方案中的代理通过两种不同的事件来支持知识社区中的知识管理：其一是当新知识生成时，用代理通知整个知识社区中的成员；其二是当社区成员访问存入的新知识时，用代理更新用户登录文件以保证知识社区的成员可以得到自己感兴趣的知识，同时过滤掉不感兴趣的知识。只要用户通过系统利用和创造新知识，代理就可以随时随地地为知识社区中的用户服务，这样就能随时更新知识社区中的知识状态，并通知相应的用户，以保证知识资源的最佳利用率。

此外，我们在知识社区实际设计与使用中发现，为了保证代理的正常使用以及对知识社区协同学习的支持，还有很多问题值得研究：如加强知识社区成员的团队意识的问题，也就是如何促进他们把自己的隐性知识贡献出来存入社区知识库，供社区其他成员共享；再如系统结构和技术问题，包括社区成员必须对整个知识系统负责，以保证系统能够稳定运行，因为代理将全天候地运行在计算机网络上和基于客户机/服务器结构的网站上，系统的稳定性是一个亟待解决的问题。□

参考文献

1 Chae B，Koch H，Paradice D. Exploring Knowledge Management Using Network Theories. The Journal of Computer Information Systems，2005，45（4）：62

2 Heier H，Borgman H P，Manuth A. Expanding the Knowledge Management System ShareNet to Research & Development. Journal of Cases on Information Technology，2005，7（1）：92

3 Tirpak T M. Five Steps to Effective Knowledge Management. Research Technology Management，2005，48（3）：15

4 姜继娇，杨乃定. 基于 Multi-Agent 的虚拟组织知识管理研究. 计算机工程与应用，2005（15）：114～116

5 张建华，刘仲英. 当前知识管理系统模型问题与对策分析. 情报学报，2004（1）：73～77

作者简介：刘高勇，男，1975 年生，博士生，讲师。
汪会玲，女，1978 年生，博士生。

收稿日期：2005－12－26

（上接第 340 页）

基于概念特征的特征提取方法是在 VSM 的基础上，对文本进行部分语义分析，利用知网获取词汇的语义信息，将语义相同的词汇映射到同一概念，进行概念聚类，并将概念相同的词合并成同一词。用聚类得到的词作为文档向量的特征项，能够比普通词汇更加准确地表达文档内容，减少特征之间的相关性和同义现象。这样可以有效降低文档向量的维数，减少文档处理计算量，提高特征提取的精度和效率。

5 结束语

Web 文本特征提取方法的研究是 Web 文本分类、过滤、挖掘，以及知识发现的重要基础和核心问题，它是人类在知识社会中应用互联网面临的新的挑战和机遇，已引起专家学者们的高度重视和普遍关注。目前的研究主要集中于探讨、创新统计评估算法和语义概念分析方法。可以预见，随着网络知识组织、人工智能等学科的发展，Web 文本特征提取将向着数字化、智能化的方向深入发展，在社会知识管理方面发挥更大的作用。□

参考文献

1 张鹏飞等. 基于相对词频的文本特征抽取方法. 计算机应用研究，2005（4）：23～26

2 于津凯等. 一种基于 N-Gram 改进的文本特征提取算法. 图书情报工作，2004，48（8）：48～50

3 Sahami M. Using Machine Learning to Improve Information Access. Stanford：Stanford University，1999

4 Joachims T. Text Categorization with Support Vector Machines：Learning with Many Relevant Features. In：Proc. of the 10th European Conf. on Machine learning. [s. l]：[s. n]，1999

5 刘明吉. 基于协同演化的文本特征获取算法. 计算机工程，2005，31（4）：85～87

6 晋耀红，苗传江. 一个基于语境框架的文本特征提取算法. 计算机研究与发展，2004，41（4）：582～586

7 唐晓文. 基于本体论的文本特征提取. 电脑与信息技术，2005，13（1）：36～38

8 赵林等. 基于知网的概念特征抽取方法. 通信学报，2004，25（7）：46～53

作者简介：庞景安，男，研究员。

收稿日期：2005－12－14