本体驱动的文本虚拟样本构造方法研究*)

王晓东 郭 雷 方 俊
(西北工业大学自动化学院 西安 710072)

摘 要 构造虚拟样本能够为机器学习中的训练集融入先验知识,从而改善标注瓶颈问题。提出了一种本体驱动的文本虚拟样本构造方法。在确保类别不变性的前提下,该方法依据领域相关本体所明晰表达的领域知识,基于本体树的点、边、子树,从同义、父子、语义同构的多个词义关系角度实现了文本虚拟样本的构造。初步实验表明,该方法与原分类及类似方法相比具有更好的分类精度和推广能力。

关键词 虚拟样本,文本分类,本体,本体树,领域知识

Research on Ontology-driven Text Virtual Sample Constructing

WANG Xiao-Dong GUO Lei FANG Jun
(College of Automation, Northwestern Polytechnical University, Xi'an 710072)

Abstract Constructing virtual examples can incorporate prior knowledge into training set in machine learning, so as to alleviate the labeling bottleneck. An Ontology-driven scheme to construct text virtual sample is proposed. Under the precondition of label invariability, the proposal constructs virtual samples according to the domain knowledge explicitly formalized by domain-specific Ontology. Based on the different Ontology tree structures, namely nodes, edges, and sub-trees, various lexical-semantic relations, including synonymy, paternity, and semantic isomorphs, are applied into text virtual example constructing. The primary experimental results show the scheme outperforms original text categorizations and other similar ones in precision and generalization ability.

Keywords Virtual example, Text categorization, Ontology, Ontology tree, Domain knowledge

1 引言

虚拟样本是对机器学习中的实际样本集进行某种知识驱动下的扩展,所获得的人工生成样本。通过虚拟样本对训练集的扩充作用,能够使标注瓶颈问题[1]得到有效改善。

虚拟样本的概念由 Poggio[2]在 1992 年首次提出,目前已经被成功应用到 3D 图像[3]、手写体数字[4]、语音[1]识别等多个研究方向。本文主要关心的是虚拟样本在自动文本分类中的应用。在该类应用中,虚拟样本构造技术的研究较其它领域来说还相对薄弱。虽然国内外学者[5,6]先后进行了一些尝试,使得原分类算法在推广性与分类精度上得到了一定提高,实践证明了虚拟样本构造在文本分类中的可行性,但总体上看,这些已有方法还显得相对粗糙[5,6],更重要的是缺乏有力的理论支持,致使文本虚拟样本在构造方法类型和分类知识蕴含度上都十分有限。因此,寻找更加系统有效的文本虚拟样本构造手段及支持理论,成为一个非常值得研究的问题。

构造虚拟样本知识输入是关键,在自动文本分类中可以利用领域内的词义关系(lexical semantic relation)作为文本虚拟样本构造的知识输入,并且这种关系不仅只是目前已利用到的同义关系,还应包括上下位、整体部分、实例,甚至更为复杂的关系。之前之所以未能引入,很大程度上就是因为没有非常适合的工具去形式化地描述该知识。随着近些年来计算机领域中本体技术的迅速崛起与发展,本体所具有的形式化、明晰的知识描述已为该设想提供了理论与实践基础。基于上述考虑,本文提出利用领域相关本体所承载的、丰富、具体的领域知识作为输入,基于本体树的点、边、子树从同义、父子、语义同构的多个词义关系角度出发,构造蕴含更丰富分类信息的虚拟样本到文本分类训练集,进而达到提高文本分类精度及推广能力的目的。

1 基本概念

1.1 本体

本体概念源于哲学,在人工智能领域中是指共享概念模型的、明确的、形式化规范说明。作为一种新的知识表示工具,本体能有效承载领域知识。

定义 1(本体) 本体 O 是一个 5 元组。$O:=(L,F,C^*,H,Root)$。L:词典,自然语言词汇集;C^*:概念集;F:参照函数,$F:2^L \mapsto 2^{C^*}$;H:层次关系,$H \subset C^* \times C^*$,H 是有向的、无环的、传递的、自反的,$H(c_1,c_2)$ 代表 c_1 是 c_2 的下层概念;$Root$:顶层概念,$Root \in C^*$。

依据对领域 DM 的相关知识 K_{dm} 的专署度不同,本体可分为通用(domain-independent)本体和领域相关(domain-specific)本体[7]。领域相关本体比通用本体对专业领域知识描述更精确。本体中的概念与关系还可以构成一棵树,称为本体树。

定义 2(本体树) 本体树 $T_o=(V,E,Root)$ 是特殊的有向无环图,节点(node)集 $V \subseteq C^*$,并满足除了根节点 $Root$ 外其余节点入度皆为 1,有向边(edge)集 $E \subseteq H$。从 $Root$ 到节点 v 的边数称为 v 的层数,最大的层数称为本体树的高。若节点 u 到 v 有直接连线 $\langle u,v\rangle \in E$,则称 u 为 v 的父亲,v 为 u 的儿子,同父节点称为兄弟。图 1 为体育领域相关本体树的片段。

*)国家自然科学基金资助项目(60675015)。王晓东 博士生,研究方向为信息检索、本体、数据可视化;郭 雷 博士生导师,研究方向为神经计算、计算机视觉、图像处理等;方 俊 博士生,研究方向为本体、语义网技术。

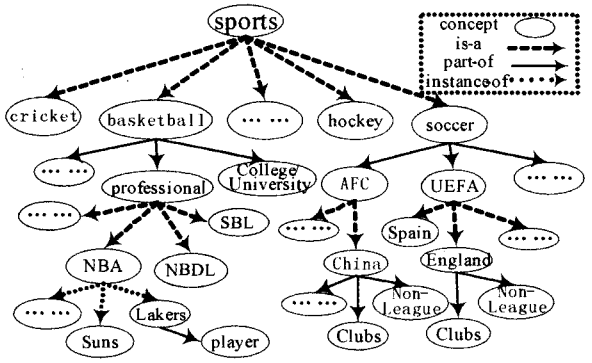

图 1　体育领域相关本体树片段

本体树中节点对应的概念 c 由同义词集(synset,记为:$syn(c)$)表示,由释义(gloss)解释,由属性(attribute)集合$\{a_1, a_2, \cdots, a_N\}$来约束。词汇间的同义关系连同 E 表示的 is-a、part-of、instance-of 等词义关系构成了领域内的主要词义关系。

定义 3(本体子树)　设 v 为本体树 T_o 上的任意一点,如果 v 及其若干后代导出的子图 T_o^v 亦为树,则称 T_o^v 为以 v 为根 T_o 的一棵子树(sub-tree)。如图 2 的 T_o^{NBA}, T_o^{NBDL} 均为图 1 本体数的子树。

定义 4(子树同构)　若本体 T_o 的两子树 T_o^v 与 T_o^u 满足:$T_o^v \neq T_o^u$,顶点集合之间存在双射,对应有向边的指向及类型相等,则称两子树同构,记为:$T_o^v \cong T_o^u$。如图 2 中的 T_o^{NBA}, T_o^{NBDL} 就是一对同构子树。

1.2　虚拟样本

在自动文本分类中,自然语言语义表达的多样性与复杂性使得仅靠人工搜集的有限训练样本不可能非常完备地反映文本类实际的数据分布。虚拟样本方法可以利用先验知识构造类别中可能出现而未出现的样本,在数量与分布上有针对性地丰富训练集对实际数据分布的模拟。

定义 5(虚拟样本)　$e=(x,y), x\in R^n, y\in\{-1,+1\}$为任意训练样本,若以知识 K 及 e 为输入,经变换 M 生成新样本集 $E_v=\{e_v^i\}, e_v^i=(x_v^i, y_v^i), x_v^i\in R^m, m\geqslant n, y_v^i\in\{-1,+1\}$ 并有 $y_v^i\equiv y$,则称 e_v^i 为 e 的虚拟样本,e 为 e_v^i 的原样本,M 为虚拟样本构造方法,记为:$M(e,K)\Rightarrow E_v$, $n_a=|E_v|$ 为虚拟样本的个数。

2　基于本体的虚拟样本构造

2.1　基本构造方法

文本虚拟样本构造一般采用对原样本中的文本子段进行替换[5]或增删[6]的基本方法,具体包括:(1)对换替换;(2)增量替换;(3)减量替换;(4)不变替换,等。根据替换子段的类型,构造方法又可分为词操作和文本块操作。本文以方法(1)、(2)为主,操作方法综合采用词操作和文本块操作。

2.2　基于本体的虚拟样本构造原理

构造虚拟样本必须确保类别不变,否则构造将失去控制和意义。为了保证类别不变性,文[5,6]依靠训练集中同义词/词块或类关键词的替换、增删的方法实现虚拟样本的构造,具有片面性和随机性。

依据概念以及概念之间的语义关系是文本类别的决定因素这一事实,加之二者又可被本体形式化地描述出来,通过逆向思维可以得到以下法则:

法则 1　以原样本中的领域相关本体概念为基点沿词义关系路径为原样本增加或替换新的本体概念,则生成的虚拟样本与原样本类别一致。

基于类似思想,本体已经被成功应用到信息检索中的查询扩展来解决查询与目标的失配问题[7]。扩展后的查询与原始查询在主题上通常能够保持一致。同理,我们考虑以本体表达的领域知识对各原样本进行扩展实现 $M(e,O)\Rightarrow E_v$。

由于本体概念在文本中主要以领域关键词 k_i($k_i\in\{syn(C^*)\}$),既可以是词也可是短语,简称为关键词)的形式存在,对于 2.1 节的两种基本操作方法,本文分别选择关键词 k_i 和关键词相关词集 $block(k_i)$ 作为词操作和文本块操作对象单元。$block(k_i)$是由关键词 k_i 连同 k_i 的相关词构成的词集,可以通过回溯 k_i 在样本中上下文窗口内(本文取 k_i 左右的±10个实词)的词集构成。若训练集中未出现 k_i,则以 k_i 对应概念的释义作为 $block(k_i)$。

2.3　关键词定位与概念匹配

由 2.2 节可知,关键词是本虚拟样本构造的基点,因此在进行构造之前,需要对样本文本矢量 x_l 中的关键词进行定位并与 C^* 进行概念匹配。

若 k_i 与 $c_m(c_m\in C^*)$ 的同义词集 $syn(c_m)=\{l_1^m, l_2^m, \cdots, l_n^m\}$ 中的 l_j^m 有重叠,则认为 k_i 为关键词且与概念 c_m 相匹配,记为:$k_i\vdash c_m$,匹配度 $Score(c_m)=\max[\|k_i\|/\|l_j^m\|]$, $\|\cdot\|$ 为词数。若 k_i 与 C^* 中不止一个概念相匹配,即:存在歧义,则需进行消歧处理来找出 x_l 中 k_i 所对应的概念。本文采用文[8]提出的消歧算法改进形式,如(1)式:

$$Match_{c_i\in C_s}(k,c_i)=\max[Score(c_i)+\sum_{j=1}^{m}\frac{Score(c_j)}{SD(c_i,c_j)}]\geqslant \gamma_{Score} \quad (1)$$

其中,k 为关键词,$C_s(C_s\subset C^*)$是消歧候选概念集,$c_i\in C_s$; c_j 属于集合$\{c_1, c_2, \cdots, c_m\}$,该集合是 k 的相关概念集,本文使用 x_l 中的无歧义匹配概念,所以令 $Score(c_j)\equiv 1$; $SD(c_i,c_j)$是概念 c_i 与 c_j 在本体树 T_o 中的最短路径长度; γ_{Score} 为概念匹配判定门限值,大于该值则认为匹配成功。

2.4　基于本体树的虚拟样本构造方法

依据法则 1,现正式提出以利用本体树结构的点、边、子树对原样本进行扩展构造虚拟样本的方法,等价于对原样本在同义、父子、语义同构三个方向(如图 2)的扩展,对应生成Ⅰ、Ⅱ、Ⅲ型虚拟样本。

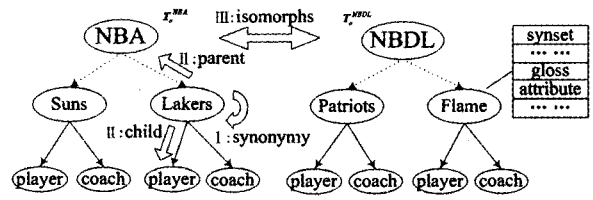

图 2　基于本体树的虚拟样本构造方向

2.3.1　基于本体树点的虚拟样本构造

基于点的方法利用本体中同一概念下的同义词/文本块对原样本相应部分进行替换,模拟"一义多词"形成的样本分布(图 2 中方向Ⅰ),例如:"MUFC"(曼联)与"the red devil"(红魔)的替换。

虚拟样本并非越多越好,而应当是规模有限并具有针对性,因此对替换候选对象 l_i 应当有所选择,尽量选择文集中出现频率低、原样本中词频高,以及对于概念相对更重要的词,因而定义 c_m 的同义词替换优先级 S_1 如下:

$$S_1(c_m)=cons+idf_i-tf_i-Index(l_i)/|syn(c_m)| \quad (2)$$

$l_i\in c_m$,其中 $cons>0$ 为常数用于确保 $S_1>0$, idf_i 为 l_i 在训练集中的逆文献频率(l_i 未在训练集中出现时 tf_i 取一

大数∧），tf_i 为 l_i 在原样本中的词频，$Index(l_i)$ 为 l_i 在 c_m 同义词集中的序号，$|syn(c_m)|$ 为 c_m 同义词的个数。替换后需重新计算词频及权值。词块的优先级在确定词后再比较各词块上下文词的 idf_i 与 tf_i 值来确定，下同。

2.3.2 基于本体树边的虚拟样本构造

基于边的方法以本体的概念层次关系 H 作为延伸路径，沿概念之间的有向边对原样本关键词进行扩展，模拟父子关系形成的样本分布（图 2 中方向Ⅱ），如：用“NBA”或“Kobe”对“Lakers”进行的增量替换。替换词选择时，父概念同义词仍然可以采用(2)式，对子概念 c_m 同义词替换时还应当考虑 c_m 的整体频率分布，因此定义优先级 S_2 如下：

$$S_2(c_m)=S_1(c_m)+\sum_{l_j\in c_m} idf_j-\sum_{l_j\in c_m} tf_j \tag{3}$$

其中 $\sum_{l_j\in c_m} idf_j$ 为 l_i 子概念 c_m 同义词的总逆文献频率，$\sum_{l_j\in c_m} tf_j$ 为 l_i 子概念 c_m 同义词的词频。

2.3.3 基于本体子树的虚拟样本构造

基于点与边的方法可以沿原样本集涉及概念所处的本体树枝进行纵向扩展，但对于未涉及树枝上的概念却无法覆盖。为此，本方法利用语义同构的词义关系来实现不同枝上概念间的横向扩展（图 2 中方向Ⅲ）。语义同构是指领域中区域相似概念间的语义组织的相似性，如图 1 中的 NBDL 与 NBA 就是职业篮球下的区域相似概念组织，它们的子树存在同构关系，且概念之间存在双射关系。利用该关系，我们可以用原样本匹配概念构成的子树的同构树进行对换替换，实现树枝间的横向扩展。下面给出寻找同构子树时节点映射关系的判定条件如下：

$$d(c_1,c_2)=\sum_{k=1}^{k_1}\beta_k|a_{1k}-a_{2k}|+\sum_{k=1}^{k_2}\beta_k d(a_{1k},a_{2k}) \tag{4}$$

β_k 为权重因子，k_1 为概念有序属性的个数，k_2 为概念无序属性的个数。(4)式通过对概念对的属性集进行相似距离比较来实现判定，当所有的对应点的相似距离小于某阈值则认为双射关系成立。多项式第一部分为有序属性（如属性 a_i 的取值范围是 $\{professional; semiprofessional; amateurish\}$ 的有序值）距离，只需对属性序号进行减运算。第二部分为符号属性（如属性 a_i 的取值范围是 $\{soccer; hockey; basketball\}$ 的无序符号）距离，按照(5)式计算：

$$d(a_{ik},a_{jk})=\begin{cases}0 & a_{ik}=a_{jk}\\ 1 & a_{ik}\neq a_{jk}\end{cases},i\neq j \tag{5}$$

同构子树的判断还应满足双射点的层数、对应边的类型相等的条件。通常，以兄弟节点为根的子树之间同构的可能性相对较大，可以优先考虑。

替换的优先级 $S_3=\sum_{c_i\in T'} S_1(c_i)$，$T'$ 为候选子树。

2.3.4 算法

算法 1 给定同类别支持向量集合 $E=\{e_1,e_2,\cdots,e_n\}$，$e_i=(x_i,y_0)$，领域相关本体 O_{dm} 连同本体树 T，临时变量 $e_v^{temp}=Null$，控制参数 n_a^1,n_a^2,n_a^3；求虚拟样本集 E。

① 顺次取样本 e_i 并得到对应文本矢量 x_i；

② 依 S_1 对 x_i 实现基本替换(1)，得到 e_v^{temp}；$E_v^1+\{e_v^{temp}\}\rightarrow E_v^1$；

③ $|E_v^1|>n_a^1$ 或所有同义词 l_j^i 均被替换使用过，则执行下一步；否则执行②；

④ 依 S_1 得到 x_i 各关键词父概念词集 $\{l_j^{upper}\}$ 实现基本替换(2)，得到 e_v^{temp}；$E_v^2+\{e_v^{temp}\}\rightarrow E_v^2$；

⑤ 依 S_2 得到 x_i 的各关键词的子概念词集 $\{l_j^{down}\}$ 实现基本替换(2)，得到 e_v^{temp}；$E_v^2+\{e_v^{temp}\}\rightarrow E_v^2$；

⑥ $|E_v^2|>n_a^2$ 或所有子概念同义词均被替换使用过则执行下一步，否则执行第④步；

⑦ 得到以文本矢量 x_i 对应本体子树 T'，在 O_{dm} 中查询该子树的所有同构子树得到集合 $T_{sub}=\{T_1,T_2,\cdots,T_k\}$；

⑧ 依 S_3 取同构子树 T_i 对 x_i 进行基本替换(1)，构造虚拟样本 e_v^{temp}；$E_v^3+\{e_v^{temp}\}\rightarrow E_v^3$；$T_{sub}-\{T_i\}\rightarrow T_{sub}$；

⑨ $|E_v^3|>n_a^3$ 或 $T_{sub}=\phi$，则执行下一步，否则执行⑦；

⑩ $E-\{e_i\}\rightarrow E$；如果 $E\neq\phi$ 执行第①步，否则 $E\cup E_v^1\cup E_v^2\cup E_v^3\rightarrow E$ 结束。

关于文本块替换与词替换相类似，算法从略。

3 实验

本节实验在支持向量机文本分类对本文方法加以应用（记为：Onto-SVM），并在精度和推广性方面同原 SVM 及文[5,6]方法进行比较。

3.1 实验配置

实验采用 movies，software，vehicles，basketball 共 4 个本体，各本体主体框架均剪裁自 http://rdf.dmoz.org/的开放式分类目录（DOMZ）对应部分，以 is-a、part-of、instance-of 三种主要关系定义概念之间的词义关系，并利用 WordNet 提供的概念同义词集和释义对本体概念节点进行填充（WordNet 未出现的概念自行填补相关项），再为各概念定义：领域、类型、组织、形式等 12 个属性参量，最后以 OWL 方式存储。实验语料来自 Yahoo 分类目录中对应类别网站上的英文新闻网页各 1200 篇，共计 4800 篇，长度适合，主题集中。

实验程序主体部分采用 Matlab7.0 实现（仅对正例样本进行构造），SVM 采用 LIBSVM[9] 提供的工具，并选择高斯径向基函数（RBF）$K(x,z)=\exp\{-\gamma\|x-z\|^2\}$ 作为核函数。实验机器采用 PentiumⅣ 3.0G CPU、512M 内存、Windows XP 操作系统、NTFS 文件系统。

3.2 预处理

在去除标记和停用词之后，即按照 2.1 节的方法进行概念匹配（$\gamma_{Score}=0.6$）。然后选择 4 个本体中的全部同义词、释义中的实词/词组以及文集中出现的其它未被匹配的词构成向量空间特征词集合，采用 TF-IDF 模式表示文本。取各 1000 个 4 类文本矢量组成训练集（分类时以同类为正例，异类为负例），其余共计 800 个矢量构成测试集。

3.3 实验结果与分析

SVM 参数最佳经验值[9] 取 $\gamma=0.5$；$n_a=6$，$\beta_k=1$，同构子树概念双射判定门限取 5；其余参量取各自最优值，Onto-SVM 各型虚拟样本的比例为：$n_n^1=n_n^2=n_n^3$，词操作和文本块操作各占一半。

3.3.1 分类精度比较

各分类方法在四类文本分类应用中的 F_1 精度如图 3 所示。

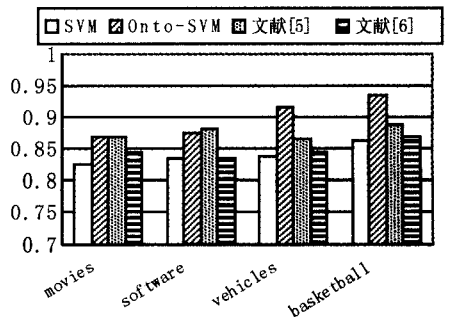

图 3 精度比较

由图3可知，三种虚拟样本生成方法的精度总体高于原SVM分类，其中Onto-SVM相对最好，这得益于Onto-SVM对样本数量和质量提升的良好表现。但在稳定性方面Onto-SVM表现一般，如对software，movies的分类Onto-SVM在精度上的优势并不突出。分析认为Onto-SVM对分类精度的提升作用与本体的领域知识描述质量有很大关系。由于DOMZ在software，movies的分类颗粒较粗，导致相应构建的本体在领域知识的表达详细程度上不十分理想，因而构造出来的虚拟样本知识蕴含量也相对较低。

3.3.2 推广能力比较

考虑到训练样本数量应当对分类精度有所影响，设计实验依次增加训练样本的数量进行分类精度测试。分别取50、100、200、400、600、800、1000个正例（负例数量不变）分别构成训练集，测试四种方法对四类文本的平均F_1精度，实验结果如图4所示。

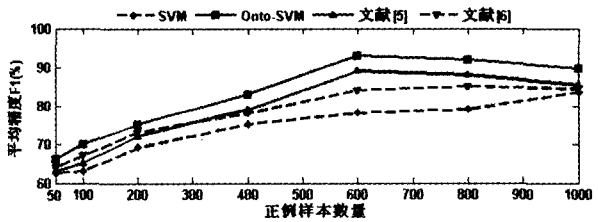

图4 样本数量与精度比较

由图4可知，在不同正例样本数量条件下，虚拟样本构造方法对推广能力的提升作用呈先升后降的趋势，但总体上还是高于原SVM。其中Onto-SVM的增幅最大，而降幅相对最小。分析认为：虚拟样本的构造对原样本数量有很大依赖性，原样本数量越多则生成的虚拟样本的质量也就越好。此外，由于虚拟样本构造具有一定的风险性，很可能构造出与实际不符的样本，我们称之为假样本，会给分类带来危害。随着真样本的增加，假样本的危害作用表现越来越明显。因而可以得出结论：构造虚拟样本的方法更适合小样本空间时的机器学习，这与文[5]的结论也是一致的。Onto-SVM的增/降幅的不同表现间接说明其有效样本相对较多。

结束语 本文研究了本体驱动的文本虚拟样本构造，与同类方法相比不同之处在于：利用了本体表示的共享概念化知识作为构造虚拟样本的依据，实验表明该方法是有效的。未来的工作可以从以下几个方面展开：首先进一步对文中几种虚拟样本的构造方法进行测试，并逐步完善；其次对设计虚拟样本的选择方法进行合理的筛选。随着语义网技术的不断发展，相信本体将会成为越来越重要的知识源驱动，对机器学习进行优化。

参考文献

1 苏金树，张博锋，徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报，2006，17(9)：1848～1859

2 Niyogi P, Girosi F, Poggio T. Incorporating prior information on machine learning by creating virtual examples [J]. Proc. IEEE, 1998, 86(11): 2196～2209

3 Poggio T, Vetter T. Recognition and structure from one 2D model view: observations on prototypes, object classes, and symmetries[C]. A. I. Memo No. 1347, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1992

4 Scholkopf B, Simard P, Smola A, et al. Prior knowledge in support vector kernels[C]. Advances in Neural Information Processing Systems. MIT Press, 1998

5 李辉，等. 运用文本领域的常识改善基于支撑向量机的文本分类器性能[J]. 中文信息学报，2002，16(2)：7～13

6 Sassano M. Virtual examples for text classification with support vector machines[C]. In: Proceedings of 2003 Conference on Empirical Methods in Natural Language Processing, 2003. 208～215

7 Bhogal J, Macfarlane A, Smith P. A review of ontology based query expansion [J]. Information Processing and Management, 2007, 43(4): 866～886

8 Latifur R K, Mcleod D. Ontology-based information selection [D]. California: University of Southern California, 2000

9 Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines[CP], 2001. Software available at http://www.csie.ntu.edu.tw/～cjlin/libsvm.

（上接第21页）

强度的修复策略、惯性因子和种群规模对算法的影响，总结了参数和算法性能之间的相互关系。本文提出的基于粒子群算法的网络社区发现方法为一个网络社区二分方法，需要对网络图不断进行分割，完成最后的划分。进一步研究的研究工作可包括k-社区划分方法等。同时针对在适应度计算随问题规模增长的同时时间复杂度和空间复杂度都跟随增长的问题，可以研究相应的并行粒子群网络社区划分方法、负载均衡该方法的适值计算。这方面的工作有待进行深入的研究。

参考文献

1 Garey M R, Johnson D S. Computers and Intractability: A Guide to the Theory of NP-Completeness. San Francisco: W H Freeman, 1979

2 王晓宇，周傲英. 万维网的链接结构分析及其应用综述[J]. 软件学报，2003，14(10)：1768～1780

3 杨楠，弓丹志，李忱，孟小峰. Web社区发现技术综述[J]. 计算机研究与发展，2005，42(3)：439～447

4 Wu F, Huberman B A. Finding communities in linear time: A physics approach[J]. Euro Phys J B, 2003, 38: 331～338

5 Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks. cond-mat/0309488. 2004

6 Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs [J]. Bell System Technical Journal, 1970, 49: 291～307

7 Girvan M, Newman M E J. Community structure in social and biological networks [J]. In: Proc Natl Acad, 2002, 99: 7821～7826

8 Newman M E J, Girvan M. Finding and evaluating community structure in networks [J]. Physical Review E, 2004, 69: 026113

9 Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks. cond-mat/0309488. 2004

10 Duch J, Arenas A. Community detection in complex networks using extremal optimization. Phys Rev E 72, 2005: 027104

11 Kennedy J, Eberhart R C, Particle swarm optimization[C]. In: Proceedings of the 1995 IEEE International Conference on Neural Networks, vol 4, IEEE Press, 1942～1948

12 Eberhart R C, Shi Y, Comparison Between Genetic Algorithm and Particle Swarm Optimization[C]. Evolutionary Programming VII (1998), Lecture Notes in Computer Science 1447, Springer, 611～616

13 Shi Y H, Eberhart R C. A Modified Particle Swarm Optimizer [C]. In: IEEE International Conference on Evolutionary Computation, Anchorage, Alaska, May, 1998

14 Zachary W W. An information flow model for conflict and fission in small groups[J]. Journal of Anthropological Research, 1977, 33: 452～473

15 Lusseau D. The emergent properties of a dolphin social network [J]. Biology Letters. In: Proc R Soc, London B(suppl.). DOI 10.1098/rsbl.2003.0057. 2003

16 Newman M E J. Modularity and community structure in networks[J]. Proc Natl Acad Sci USA 2006(in press)