基于互联网用户心理挖掘的网站深翻系统

王 征 徐培文
(西南财经大学信息学院 & 金融研究院 成都 611130)

摘 要 基于人工检索和顶置的网站优化系统工作效率低下,响应速度慢。为消除上述问题,提出并设计了基于互联网用户心理挖掘的网站深翻系统。该系统通过相关网站及自身的观点挖掘活动,检索近期网络热点,并根据历史信息对客户需求进行测度和提取;最终将按照客户的观点需求和当前热点,从历史数据库中提取既往信息进行网站优化。仿真实验表明,该系统能够较好地实现互联网用户心理挖掘及网站的搜索引擎优化,提高网站点击率。

关键词 网络,用户心理特征,心理挖掘,搜索引擎优化,匹配
中图法分类号 TP18,TP393 **文献标识码** A

Site Deep-digger System Based on Internet Client Psychology Mining

WANG Zheng XU Pei-wen
(School of Information, Financial Research Institute, Southwest University of Finance and Economics, Chengdu 611130, China)

Abstract Site optimization systems based on artificial retrieval and topping work slowly and inefficiently. In order to deal with these problems, a site deep-digger system was proposed with internet client psychology mining. The system utilizes Web opinion mining technologies to search Internet hot points. And it can extract and estimate client requests from historic data. According to these requests and historic information, it optimizes the site by creating related Webs. Simulation results show that the system can perform better to mine internet opinions and do SEO (search engine optimization) better than the traditional methods do so.

Keywords Network, Client psychology character, Psychology mining, Search engine optimization, Match

1 前言

互联网用户心理挖掘是以心理学经典理论为基础、以数据挖掘与处理为手段,研究互联网相关情景下网络用户的心理、行为及其规律性的应用学科。互联网用户心理挖掘可以用于处理网络中海量的在线评论;针对产品、音乐、电影和博客等信息源,分析产品的属性评价,生成产品的评价摘要;也可以利用主观评价,结合用户的行为进行信息推荐。而相关的研究也包括统计博客的支持率和反对率,从而计算博主的个人声誉度[1]。此外,它结合话题跟踪和检测技术,发现感兴趣的话题,建立话题的传播模型,计算话题在网络中的心理倾向性,能够进行广域的网络舆情分析。而搜索引擎优化能够针对搜索引擎对网页的检索特点,让网站建设中的各项基本要素适合搜索引擎的检索原则,从而使搜索引擎收录尽可能多的网页,并在搜索引擎自然检索结果中排名靠前,最终达到网站推广的目的[2,3]。

当前,众多的研究者、电信运营商以及网站管理者都在研究如何将心理挖掘与搜索引擎优化两种方法紧密结合,为网站提供自动化的信息更新服务。通常采用的网络信息获取流程通常如下:首先,确立采集目标,即由用户选择目标网站;然后,提取网站特征信息,根据目标网站的网页格式,提取出目标数据的共性;最终进行网络信息获取,即利用工具自动地把页面数据存到数据库[4-6]。基于此流程,研究人员开发了各类网络信息抽取工具,主要可分为:(1)开发包装器的专用语言(Languages for Wrapper Development):用户可用这些专用语言方便地编写包装器,例如 Minerva, TSIMMIS, Web-OQL, FLORID, Jedi 等[7]。(2)以 HTML 为中间件的工具(HTML-aware Tools):这些工具在抽取时主要依赖 HTML 文档的内在结构特征,代表工具有 Knowlesys, MDR。(3)基于 NLP (Natural language processing) 的工具 (NLP-based Tools)[8,9]:这些工具通常利用 filtering、part-of-speech tagging、lexical semantic tagging 等 NLP 技术建立短语和句子元素之间的关系,推导出抽取规则,代表工具有 RAPIER, SRV, WHISK[10,11]。(4)基于模型的工具(Modeling-based Tools):这些工具让用户通过图形界面,建立文档中其感兴趣的对象结构模型,代表工具有 NoDoSE, DEByE。(5)基于本体的工具(Ontology-based Tools):这些工具首先需要专家参与,人工建立某领域的知识库,然后基于知识库去做抽取操作,代表工具有 BYU, X-tract。

针对这种实际应用需求和研究现状,以及现有方法存在

到稿日期:2011-12-05 返修日期:2012-06-20 本文受教育部人文社科项目(10YJCZH169),四川省金融智能与金融工程重点实验室项目(FIFE2010-P05),西南财经大学校管课题(2010XG068)资助。

王 征(1979—),男,副教授,硕士生导师,主要研究方向为分布式系统,E-mail:wangzheng151400@163.com;徐培文(1984—),男,博士生,主要研究方向为金融行为理论。

的种种问题，本文提出了一种基于互联网用户心理挖掘和搜索引擎优化的网站深翻系统。本文第2节分析当前应用系统中存在的问题，并提出破解方案；第3节提出深翻系统的模型结构与功能模块；第4节论述系统实现中的关键技术与流程；第5节是系统性能分析与仿真实验结果分析；最后是结论。

2 问题分析与解决

在网页信息采集系统中，搜索引擎采用 Robot 程序进行信息抽取，它依照 HTTP 协议读取 Web 页面并根据 HTML 文档中的超链在互联网上进行自动漫游；但 Robot 只能获取 Web 上的静态页面，而有价值的信息往往存放在网络数据库中，人们无法通过搜索引擎获取这些数据，只能登录专业信息网站，利用网站提供的查询接口提交查询请求，获取并浏览系统生成的动态页面。因此，智能的网络信息抽取系统应该主动通过网站提供的查询接口对网络数据库中的信息进行遍历（即网络深翻）；并根据知识库对遍历的结果进行自动的分析整理，最后导入本地的信息库[12]。由于数据采集量和处理工作量巨大，搜索引擎和信息抽取系统的这种工作模式并不适合中小型网站进行内部信息优化。为此，本系统采用了一些新型的解决方案。

首先是基于搜索引擎的反向心理信息抽取，即外部心理挖掘。从实现原理上，普通网站不可能自建搜索引擎以获取其他网站的信息，要获取相关的心理信息，必须另辟捷径。因此，本模型采用了基于搜索引擎的反向心理信息抽取技术，该项技术基于两个方面的事实：(1)普通网站可以通过搜索引擎的缓存（例如：百度快照）获取其他网站的相关信息；(2)尽管网络中相关网页较多，但是搜索引擎能提供主题较为集中、数量有限度的链接（例如：百度针对一组关键词最多提供 760 个相关链接；而 google 最多提供 740 个，其余的一概省略），因此网站完全可以通过“有限的心理主题词＋有限的搜索引擎”获得数量有限但主题集中的信息，以提高自身网站的内容质量。

其次是系统内部基于用户心理需求的挖掘和匹配的网站信息优化，即内部心理挖掘。目前，SEO（Search Engine Optimization，搜索引擎优化）工作主要针对用户的表层浏览行为和习惯，很少对深层次的用户心理元素进行发掘。这种方法带来的负面影响是：SEO 的过程大部分滞后于用户的兴趣转变，挖掘到的优化线索易丢失；而用户心理挖掘技术能够追寻持久的优化线索，并通过匹配操作，获取和发布符合用户心理变化趋势的信息，达到网站优化的目的。

最后是基于心理推测的站内信息发布。传统的网站主要通过人工检索和关键词匹配进行重要信息和精华历史信息的发掘，不但效率低下，而且准确性和实时性都显得不足。本模型采用心理阈值激发和定时触发两种方式，由服务器软件进行自动筛选和匹配，无需人工干预，从而提高了信息更新的实时性和效率。

3 系统结构与处理流程

该模型通常应用在网页服务器中，其系统结构和单元如图 1 所示。

(1)基于需求的外部心理线索采集单元：该单元通过系统外部的搜索引擎进行反向检索，主要获取符合网站用户心理需求与变化趋势的相关网页快照（从百度的“百度快照”和谷歌的“网页快照”中获得）。该单元以所在网站内部的用户心理特征作为检索关键词，通过定时对几大通用搜索引擎的信息收集（包括全文搜索引擎 Baidu 和 Google，分类搜索引擎 Yahoo 和新浪），及搜索引擎的自动排名来定位本网站的心理特征是否符合当前网络的发展趋势；同时，该模块归纳各大搜索引擎中排名 TopN 位的相关检索词（例：在 Baidu 的搜索结果页面末尾，有相关或相近的高频检索词列表），将其作为后续反向检索操作的依据与心理线索。

(2)用户心理发掘单元：该单元负责发掘两个方向上的用户心理信息，以便为网站优化和信息发布提供依据。

首先，通过浏览记录和网页内容的挖掘，发现本网站用户的心理倾向与相关线索。本系统基于 OCC 人工心理模型进行用户心理发掘，即不使用原始情感集和连续的多维空间处理心理问题，而是以认知诱发条件为依据将用户心理进行分类。具体流程包括：提取用户高频访问网页中的关键词；通过 OCC 函数（Ortony 等人在《情感认知结构》一书提出）进行用户心理识别并记录，将其作为网页和客户的心理特征；并结合历次心理发掘的既有结果，采用 OCC 心理激发函数进行心理倾向的预测。其中，心理特征和心理倾向采用 OCC 心理状态结构进行表述（22 种心理类型构成的向量）。

其次，通过外部心理线索采集单元获取常用外部检索词和当前互联网用户心理倾向，具体流程与前一步骤相同。

最终，将发掘获得的心理倾向与心理需求，以及与本网站用户心理相关的外部高频检索字，通过同一化处理后注入系统中的多维度心理空间中供网站优化使用。

(3)用户和网站心理特征匹配单元：该单元处理各大搜索引擎中和网站中的共有流行词（以及链接）之间的心理相关度，以便在实际的网页优化过程中，根据这些网页与用户心理需求的匹配程度，最终为自动的网络信息发布提供依据。另外，该单元还向网站的检索页面提供相应的主题词列表，作为其在搜索引擎中进行排名搜索的依据。

(4)网站深翻与优化单元：该单元同时针对两个方向上的信息进行心理向量匹配与处理。首先是对外方向上，该单元将对用户的心理需求与网站内的心理特征进行匹配，并进行网站中既有网页信息的深翻，作出网站的优化决策；其次是对系统内部各公开数据库（表）中的心理特征词进行深翻，提取本网站中的心理特征信息，并根据用户的心理需求在外部进行发布。该单元将根据两方面提供的优化决策信息进行心理特征词和网页排序的调整，从而起到网站优化的作用（例如：网页中的特征词密度正常为 1%到 7%，太低或太高都将导致网页排名落后）。其处理主要通过定时和激发两种模式，不断从用户和网站的心理叠加空间中提取心理线索，并将提取的相关内容进行两部分工作：首先对于既有网页，该单元将根据当前的心理线索进行心理特征词的频度调整（冲淡或者提高特征词密度），并且改变其在网页树中的层次位置，以提高搜索引擎的检索排名；其次对于已经从网页中剥离、存储在数据库中的信息，该单元将根据各种既有网页模板将它们重新生成网页，以便搜索引擎和用户获取。

该模型的运行流程如图 1 所示。

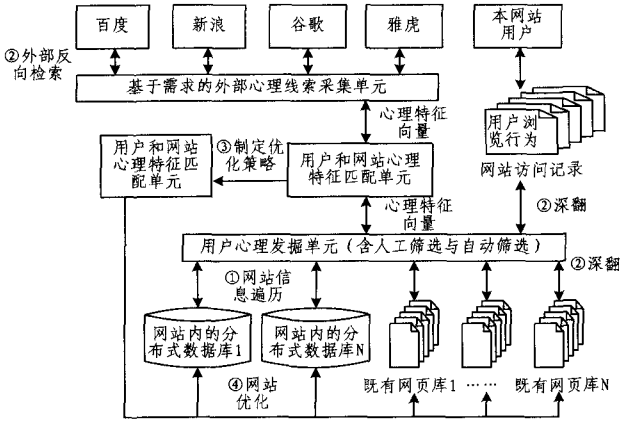

图 1　深番系统结构模型与运行流程

首先，进行内部心理元素的发掘：本模型通过以下 3 个步骤进行心理发掘：第 1 步，用户心理发掘单元遍历网站中各公开数据库与数据字典文件的相关信息，提取数据表和视图的元数据（含高频关键词等）；第 2 步，经自动筛选和人工筛选（避免数据量过大和“表不加密，字段加密”等情况）后，该模型将用户关注度较高和当前网络中流行的高频词出现的备用字段反馈给系统，从而避免模型遍历网站中的所有字段，以节省系统资源；第 3 步，该模型将各字段中自动检索的和管理员人工分拣出的用户心理特征词归并入心理向量，其表述为：$V(E$:Emotion; U:URL; K:Keys; R:Rank…$)$：E 为人工心理空间中的状态量；U 为网页地址；K 为心理特征词列表；R 为网页的搜索引擎排名。

然后，根据网站内外的心理线索进行网站内外的二次深翻：本模型利用上一步骤中的心理特征向量的排名、高频检索词等信息，通过外部心理线索采集单元和用户心理发掘单元，分析得出网站内外用户的心理需求和心理线索，并将这些信息也进行向量化处理，保存在匹配空间中备用。

接着，根据用户心理需求的变化，及时制定网站优化策略：通常网站数据库可以发布的信息较多，但不可能全部生成网页，因而必须从心理特征空间中获取既符合本网站心理特征，又顺应网络心理发展趋势的信息对象；用户和网站心理特征匹配单元通过相关度分析得到网站内外心理元素之间的相关度，并根据网站更新的具体需求，提交网页生成需求列表。

最后，该模型自动完成网页生成和网站优化：根据客户的心理需求与网站外部的心理趋势，及深翻检索出的各类信息，生成、修改和调整网站中的网页内容，从而实现网站优化的目的，例如：深翻出的既有数据库信息可以生成相关网页链接在热点新闻中，作为相关阅读提供给用户。

4　关键技术

访问频率是网页内容是否受到用户关注的衡量标志之一，但如果仅以某时间段内的访问次数作为网页是否会在未来受到用户关注的决策标准，将会造成很多用户心理信息的缺失。因此，本模型中采用了基于“新鲜度”的匹配决策模型。网页信息的新鲜度与其在网站中的驻留时间长短紧密相关，例如：某网页可能并不属于关注度很高的信息，但由于经常被低密度的访问，其新鲜度仍可能较高。

由于大多数网站通常采用树状结构，其中的相关或相近的信息往往被分置于一个信息簇内，因此深翻系统也采用了类似的数据结构收集和归纳相关信息。一个簇聚集的关键信息元素通常是由一批或者多批次的信息组成的；而深翻出来的信息往往也是成簇出现，并且由这些关键信息元素决定。即，如果簇内的关键信息元素（可能是关键词、链接、图片等形式）在一定的时间内未被访问或检索，那么这些元素的老化将导致整个簇被“湮没”（通常是置换到外存的数据库中，而不保留网页）。而与网站内外需求匹配的信息簇，也一定会具有良好的新鲜度，因此完全有必要基于新鲜度进行心理特征匹配决策的研究。本模型中设定，对于特定的簇 i，如果其中的关键元素 a_i 在时间段 j 内，其估计值 $Value_{i,n}(j)$ 的最新提交时间为 $Time_{i,n}(j)$。如果一个信息簇 i 内的一个心理挖掘决策将在特定时间段内发生，则对此信息元素的新鲜度用下述表达式度量：$F_i(i,j)=[t_{i,n}(k)-t_{i,n}(j)]^{\eta}$；此式中 η 为系统参数，当它大于 0 时，该方程自反映了信息元素 a_i 在 j 时间段内的新鲜程度。信息元素的新鲜程度随着时间和访问量变化，可以将上式标准化为：$\Phi_i(j)=[(time_{i,n}(i)-time_{i,n}(j))/(time_{i,n}(i)=t_{i,n}(0))]^{\eta}$，其中，0 时刻为信息簇开始收纳信息的时间。

根据网络信息采集的实际情况，在信息簇作出和完成决策的过程中，从各个单元才寄来的相关信息也在不断注入，必须在不过分干扰模型运行的基础上，对此加以处理。整合新到信息的模型如图 2 所示。

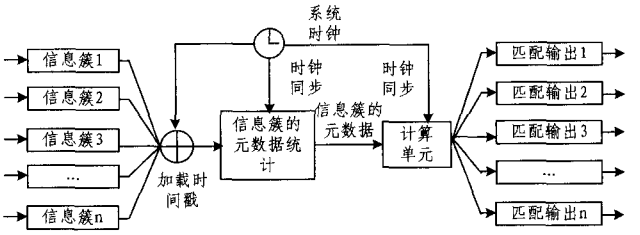

图 2　信息簇更新模型

传统的算法执行过程是原子态，即算法要么执行，要么不予执行。若采用这种方式，则新收集到的心理特征信息不能被处理，将会造成信息浪费，并延误处理时间。本模型中借鉴软中断的思想，实现了可插入新信息的计算过程。首先，每个信息簇在进行处理前，需根据当前的系统时钟以及进入计算单元的先后顺序加载时间戳；本模型根据系统时钟完成时间同步，并得到信息簇的元数据统计（某一时段内信息簇的大小、子节点数量等）；最终，计算单元根据信息簇以及及时更新的元数据进行匹配决策，选择新鲜度和相关指标较高的信息簇进行深翻。

在计算过程中，时间 t 内，N 为不断增加的信息簇的信息元素数量，该信息簇在此时段内收纳信息记录分别为 $\{(\hat{u}_1,\hat{o}_1^2),(\hat{u}_2,\hat{o}_2^2),\cdots,(\hat{u}_N,\hat{o}_N^2)\}$，并定时进行匹配决策计算。此时，对于信息簇有 $u_{k+1}=(\hat{u}_k o_k^2+u_k\hat{o}_k^2)/(o_k^2+\hat{o}_k^2)$，而 $o_{k+1}^2=(o_k^2*\hat{o}_k^2)/(o_k^2+\hat{o}_k^2)$。其中，$(\hat{u}_0,\hat{o}_0^2)$ 是初始设定的经验值。在计算过程中，可能仍有信息元素不断被信息簇收纳，这种情况之下，传统算法模型只能丢弃收纳的信息；而本模型可以通过类似软中断的方法进行处理，具体步骤如下：首先当计算单元和信息簇的元数据统计单元接到新信息到来的消息（加载在“步进”消息中）时，暂时停止计算；然后统计单元对更新后的信息簇进行数据更新，重新生成一份元数据列表，并更新计算单元中的相关参数；随后当“步进”消息到达计算单元时，计算单元重新启动运算，循环，直到没有新消息到达且计算完毕

为止。此时，新鲜度为 0 的信息簇可以直接作为深翻依据进行处理；而新鲜度为 1 的信息簇，则需要搜索是否有内外客户的及时需求，其最终将会被丢弃或者作为深翻依据。

5 实验结果与分析

本模型的验证实验在某高校内门户网站的“相关新闻”和 BBS 中的“校园轶事”等版块中进行，其中主观测试（通过对普通用户进行问卷调查）调查的普通用户人数为 271 人，回收有效调查问卷 229 份；调查的版主及管理员人数为 33 人，回收有效调查问卷 32 份。实验主要从用户的心理满足程度和信息的相关程度等内容对模型性能进行了测试，另外还对该模型对系统性能的“扰动”进行了测量，实验结果（与未用之前的测量结果进行对比）如图 3 所示。

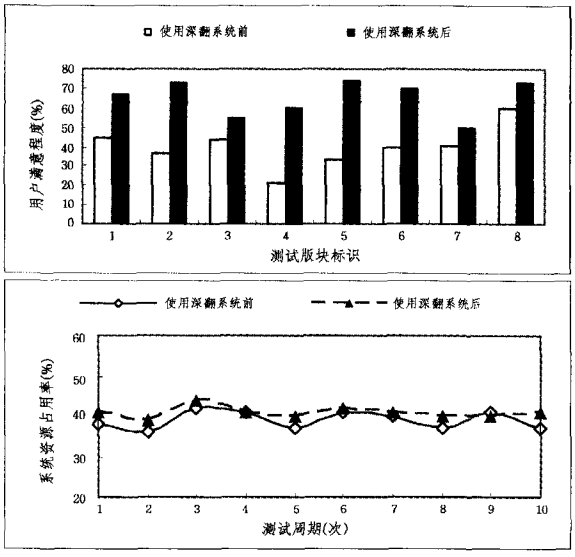

图 3 深翻系统仿真实验结果

图 3(上)中的测试结果显示使用深翻系统后，网站点击率明显提高，且用户对网站中的各版块用户满意度均有不同程度的提高，即深翻系统能够较好地从内外两个方向上获取用户的当前心理，并根据当前用户的心理需求从历史数据中“深翻”出相关度较高的内容。此外，版主与网站管理员的问卷调查也显示：深翻系统能够提供与用户心理匹配较好的信息列表，且自动化程度较高，能够节省大量的人力资源。

图 3(下)中显示了 10 个监测周期内系统的平均负载变化（网站流量、内存与 CPU 为主要监测对象），其中，测试环境采用了联想万全 R520 服务器和 Windows2003 操作系统。由图中可知，使用了深翻系统之后，网站系统的整体负载并未大幅度攀升，与使用前基本持平；通过进一步的分析，可究其原应为：由于深翻系统减轻了用户的信息检索工作量和人机交互通信量，使得网站系统占用的系统负载下降，部分抵销了网站内部“深翻”所产生的系统负载。

结束语 基于互联网用户心理挖掘的网站深翻模型在实际应用中取得了良好的效果，具有一定的推广价值。进一步的研究工作包括：多信息源辅助的深翻决策模型、海量和多批次的反向心理线索发掘模型等，并将其应用于公开股票信息挖掘等领域。

参 考 文 献

[1] 王辉，王晖昱. 观点挖掘综述[J]. 计算机应用研究，2009，26(1)：25-29

[2] 杨超，冯时，等. 基于情感词典扩展技术的网络舆情倾向性分析[J]. 小型微型计算机系统，2010，31(4)：691-695

[3] 张顺香，朱广丽，陆奎. 基于 Web 挖掘的主页多主题更新模型[J]. 计算机应用，2009，29(10)：2796-2799

[4] 李晓亚，赫枫龄，左万利. 基于网页分块技术主题爬行器的实现[J]. 吉林大学学报，2007，45(6)：959-965

[5] 阮光册. 基于兴趣度策略的启发式 Web 挖掘算法[J]. 计算机工程与应用，2009，45(35)：148-150

[6] 章剑锋，张奇，吴立德. 中文观点挖掘中的主观性关系抽取[J]. 中文信息学报，2008，22(2)：55-60

[7] 葛育祥，熊励. 整合文本挖掘的商务智能系统结构研究[J]. 计算机技术与发展，2009，19(4)：1-4

[8] 杨频，李涛，赵奎. 一种网络舆情的定量分析方法[J]. 计算机应用研究，2009，26(3)：1066-1069

[9] 周红芳，冯博琴，等. 基于语义模型的 Web 挖掘算法研究[J]. 哈尔滨工业大学学报，2009，41(11)：212-214

[10] 查凯莱巴蒂. Web 数据挖掘[M]. 北京：人民邮电出版社，2009：23-82

[11] Chou P-H，Li P-H. Integrating Web mining and neural network for personalized e-commerce automatic service[J]. Expert Systems with Applications，2010(37)：2898-2910

[12] Hung S-H，Lin C-H，et al. Web mining for event-based commonsense knowledge using lexico-syntactic pattern matching and semantic role labeling[J]. Expert Systems with Applications，2010(37)：341-347

（上接第 159 页）

[3] Elmagarmid A K，Panagiotis G，et al. Duplicate record detection：a survey[J]. IEEE Transactions on Knowledge and Data Engineering，2007，19(1)：1-16

[4] 张昌年. 一种基于 VSM 的检测相似重复记录的方法[J]. 微电子学与计算机，2008，25(8)：184-187

[5] 马翔. 粒子群优化 BP 神经网络用于重复记录检测[J]. 辽宁工程技术大学学报：自然科学版，2010，29(5)：959-963

[6] 朱恒民，王宁生. 一种改进的相似重复记录检测方法[J]. 控制与决策，2006，21(7)：805-813

[7] Elmagarmid K，Panagiotis G. Duplicate record detection：a survey [J]. IEEE Transaction on Knowledge and Data Engineering，2007，19(1)：1-16

[8] Minton S N，Nanjo C，Knoblock C A. A heterogeneous field matching method for record linkage[C]//Proceedings of the 5th International Conference on Data Mining. Washington：IEEE Computer Society，2005：314-321

[9] 巩知乐，张德贤，胡明明. 一种改进的支持向量机的文本分类算法[J]. 计算机仿真，2009，26(7)：165-168

[10] 杨福刚. 基于人工免疫算法的最小二乘支持向量机参数优化算法[J]. 计算机应用研究，2010，27(5)：1702-1704

[11] 李旭芳，王士同. 基于 QPSO 训练支持向量机的网络入侵检测[J]. 计算机工程与设计，2008，29(1)：34-36

[12] 吴文欢，张少辉，李巍. 分阶段进化的粒子群优化算法[J]. 重庆理工大学学报：自然科学，2012，26(6)：67-70