

文章编号:1673-0062(2013)04-0071-06

一种基于体裁互文性的 P2P 信息检索模型

刘志明,马家宇,万亚平,冯如晓,陈晓凡

(南华大学 计算机科学与技术学院,湖南 衡阳 421001)

摘 要: P2P 信息检索的研究大多集中于资源搜索机制,而 P2P 存储系统是一个多结构数据存储系统,系统中有多个数据源,数据有各种表示形式,不同的用户对相同的现实世界的实体可能会使用不同的描述数据,有不同的表示,这样必然会产生语义异构的问题.首次将体裁互文性概念拓展到信息空间,设计和实现了一个基于体裁互文性的 P2P 信息检索模型,探讨了基本概念和性质,并以原型系统的数据为样本,对理论进行了验证.实验表明本文所设计方法在 Bpref 信息检索典型评价指标上取得了较好的效果.

关键词: 体裁;互文性;P2P 系统;信息检索模型

中图分类号: TP 391 **文献标识码:** B

A P2P Information Retrieval Model Based on Genre Intertextuality

LIU Zhi-ming, MA Jia-yu, WAN Ya-ping, FENG Ru-xiao, CHENG Xiao-fan

(School of Computer Science and Technology, University of South China, Hengyang, Hunan 421001, China)

Abstract: Current research for P2P information retrieval is mostly concentrated in the resource search mechanism. While P2P storage system is a multi-structured data storage system, which has multiple data sources, variety data representation. And different users may use different data to describe the same real world entity, which would induce the problem of semantic heterogeneity. This is the first to extend the concept of genre intertextuality to information space. It designed and implemented a P2P Information retrieval model based on a genre intertextuality, discussed its basic concepts and features, and the theory has been verified by the sample data in the prototype system. Experiments show that good results are achieved in the design method in bpref, which are typical evaluation indicators of P2P information retrieval.

key words: genre; intertextuality; P2P systems; information retrieval model

收稿日期:2013-06-18

基金项目:湖南省自然科学基金资助项目(11JJ6047);湖南省科技计划基金资助项目(2011FJ3087);南华大学创新团队建设基金资助项目(NHCXTD16)

作者简介:刘志明(1972-),男,湖南浏阳人,南华大学计算机科学与技术学院教授,博士.主要研究方向:大数据存储、信息检索.

0 前言

P2P 信息检索是指从分布在各个独立节点上的 P2P 网络信息资源中高效地索引、查找、检索出用户所需信息的过程。由于 P2P 系统数据的分布性、无结构化和缺乏面向语义的表达与描述,导致其检索效率低下,检索成本也随着网络信息量的飞速增加而爆炸性增长。用户在 P2P 系统当中提交的查询存在部分信息丢失、简短、特征稀少的特点,往往只能使用有限数量的词汇来抽象和概括他们的需求;不同的用户可能会用不同的词汇来表达类似或者相同的信息需求,使得基于关键字进行检索的 P2P 系统不太容易很快地提供给用户满意的查询结果。因此,在 P2P 系统当中如何根据资源的特征进行分类,如何从用户的表达中提取表达用户查询意图的特征词汇进而返回查询结果给用户,如何体现用户表达的完备性和独立性,这些都是 P2P 信息检索中面临的主要问题和挑战。在信息检索模型上,目前 P2P 检索几乎都使用布尔检索模型,模型最大的缺点是不支持模糊检索,无法对检索结果进行相关度排序,同时无法区分检索概念或词语在文档中所占比重,而且其刚性匹配使漏检较为严重。为获得更好的检索效率,有许多研究者也开始尝试基于概念的向量空间模型、潜在语义模型或基于语义的概念模型等新模型。

P2P 信息检索主要根据底层拓扑结构的不同,而设计不同的信息检索机制,这方面的研究相对来说比较成熟,本文试着从另外一个角度,即文档表示层的角度,通过文档间的相关性,来探讨 P2P 信息检索质量的提高。

1 相关工作

目前对 P2P 信息检索的研究大多集中在资源搜索机制上面,例如以 Gnutella 为典型代表的非结构化 P2P 普遍采用洪泛搜索,或者是改进的洪泛搜索。很多针对无结构 P2P 网络的资源搜索改进方法,主要集中在如何利用系统中的信息引导查询转发过程尽快到达目标节点,如随机漫步、启发式宽度优先搜索、本地索引、语义覆盖网等。近年来的一些研究力图依据查询内容本身及节点的知识更新来提高搜索性能^[1],如以构建语义覆盖网络(SON)为基础的路由策略^[2],但在资源的语义表达上不够准确。PeerSearch^[3]将语义查询对应到语义空间,但资源的矢量化过程代价很大;而

以智能动态路由为基础的路由策略^[4],每个节点维护 1 个本地路由表(知识库),通过不断学习了解其所在局部的资源分布情况。

由上可见对 P2P 搜索机制的研究已经取得了许多成果,但是单纯的从搜索机制优化的角度考虑问题不一定会给搜索效率带来很大提高的空间。有些学者从节点组织角度出发考虑给用户查询和检索需求提供更好的信息检索结果。例如小世界模型及其后续研究为兴趣域分域模型提供了理论基础。Social-P2P 是一种基于兴趣域的 P2P 搜索算法,它的特点是引入了社会网理论,通过模拟社会网中人与人之间的关系,建立节点之间的联系,聚集了兴趣相近的节点,以体现小世界的特性。通过实验证明 Social-P2P 能更好聚集兴趣相近的节点,提高了 P2P 网络搜索效率。PeerDB 项目使用关键字标注关系名和属性名的方法实现异构数据源的语义映射。同时对每一组关键字都建立了相应的物化视图。Hyperion 项目通过定义复杂的映射表、映射表达式和 ECA 规则控制节点间的信息交换,高效地解决了 P2P 环境下的语义异构问题。还有一些研究者提出按相似特征对节点聚类的解决方案,文献^[5]先给网络中每个节点贴上 1 个标签,该标签是通过对该节点的某些特征进行分析后得到的;然后将节点进行聚类,并划分成不同的簇,在簇内和簇间采用不同的查询请求转发机制。文献^[6]中的方法是在系统中设立中央服务器,每个节点拥有一个由中央服务器分配的聚类中心,节点上采用 DHT,将文件的属性映射为键(key),然后通过逐步调整将节点索引记录按照一定中心进行聚类,但这样会造成系统对中央服务器的依赖。小世界模型中的方法是将节点上的资源通过 DHT 进行描述,然后在节点本地进行聚类,在节点间建立相似链接,但这种方法对资源的描述不够准确。

近年来,体裁在信息检索领域得到了大量的应用。体裁对于文档的分类、查询目标的表示、或者匹配和过滤搜索引擎的查询返回结果方面都能显著改善信息检索质量。因此,将其成功的案例移植到 P2P 信息检索领域也具有天然的亲和力。本文以语言学体裁互文性理论为基础,尝试将体裁互文性概念拓展到任信息空间,利用文档之间的互文关系解决文档体裁的自动识别问题;同时,用体裁表达文档目标及查询目标,解决传统方法中查询目标用体裁显式表达的困难。这对于 P2P 信息检索技术的实现和完善有一定借鉴意义,同时

为语言学的体裁互文性理论研究提供新视野和方法.

2 体裁互文性基本概念、性质

体裁 (Genre) 起源于传统的语言学范畴,是计算语言学和修辞学的交叉研究课题. 体裁是一种共享概念模型的明确的形式化规范说明,与写作风格、句法分析联系紧密,对文章的写作有着明显的制约和规范作用. 在研究中,如何识别、描述、利用体裁是一项复杂而具有挑战性的工作. 体裁是人类思维的抽象归纳,研究者认知受限和它自身动态演变等因素使得全面、准确、有效地概括表述庞杂的体裁类别相当困难. 其次,体裁的分类边界模糊,随着时间发生变化,只在一个特点的时期具有一定的稳定性. 最后,利用体裁自动分类交叉于传统修辞学与计算语言学之间,需要有较深的语言学功底和计算语言学理论基础.

“互文性”(Intertextuality) 概念首先由法国符号学家朱丽娅·克里斯蒂娃提出:“任何作品的文本都像许多行文的镶嵌品那样构成的,任何文本都是其它文本的吸收和转化.”互文性意指文档间语符形式与意义等多层次的交互、涉及、延异等关系,主张将文档置于动态发展过程和开放背景下加以审视,强调文档间的联系与影响,为文档研究提供了新的角度和视野^[7]. 费尔克兰福进一步将文档的互文性分为文档表层的互文性和文档深层的互文性. 文档表层互文性(显性互文)是指一个文档中标明的与其它文档的互文关系,文档深层的互文性(隐式互文)指一个文档中各种体裁或文档类型规范的那种复杂关系,它“是文档生成中涉及的那些话语规范的组合”. 文档之间的互文性可以分为两个层次:一是字面上的,称为具体互文性或显式互文,表现为文字间相互引用;另一个是隐蔽或深层次的互文性,称为体裁互文性或隐式互文,表现为语义、题材、结构、文体或形式特征等. 体裁互文性是指不同风格、语域或体裁的混合交融^[8-9].

将体裁互文性概念拓展到信息空间,基于语言学中互文性的基本定义和性质,利用信息任务体现的文档体裁互文性,开展信息科学中体裁互文性的研究,加以量化,并在体裁互文性定量研究的基础上,探讨体裁互文性的度量方法,建立相关理论. 应该说明的是,这在语言学中也是一个未被系统研究的领域.

2.1 信息空间体裁互文性基本概念

信息空间中,体裁互文性可定义为:文档在信息任务中共现超过一定的阈值,导致文档之间存在的风格 (Style)、语域 (Register) 或体裁 (Genre) 的混合交融,称之为体裁互文性. 由于具体互文性存在于文档表层,体裁互文性蕴藏于文档深层,直观可推得性质 1:

性质 1 如果文档之间存在具体互文性,则它们之间也存在体裁互文性,反之不成立.

文档是通过对其他文档的引用、改写、吸收、扩展或在总体上加以改造而获得的,则有性质 2.

性质 2 一般新的体裁都是由旧的体裁转换颠倒、置换、合并而来.

体裁转换、颠倒、置换、合并的程度可以通过体裁互文性的关系强度量化,得到性质 3.

性质 3 文档间的体裁互文性可以量化为互文关系强度.

2.2 信息空间体裁互文性基本性质

美国斯坦福大学社会学家格兰诺维特 (Mark Granovetter) 在群体、组织关系的研究中提出了关系强度的概念,将关系分为强关系和弱关系. 强关系维系着群体、组织内部的关系,弱关系在群体、组织之间建立了纽带联系. 体裁的分类与社会群体的分类密切相关. 某一群体垄断一些体裁,另一群体垄断另一些体裁^[10]. 由性质 3 推论,可得到推论 4.

推论 4 文档间的体裁互文性依关系强度分为强互文关系和弱互文关系.

强互文关系维系体裁内部的关系,保证文档具有相同的体裁. 弱互文关系在体裁之间建立了纽带联系,说明文档体裁之间有交叉、融合的部分.

将性质 1、性质 2、性质 3 和推论 4 映射到信息空间,得到:

性质 5 如果文档之间存在引用关系,则它们之间存在体裁互文性,反之不成立.

文档之间的引用最终会通过体裁互文性表现出来. 所以,对任务上下文产生的体裁互文性的研究将涵盖文档引用带来的具体互文性.

性质 6 文档的体裁可以通过体裁选择、交叉、变异产生.

文档的体裁可以使用体裁向量表示,根据遗传算法的特点,可以应用遗传算法产生文档的体裁. 事实上,周昌乐等基于遗传算法用宋词诗歌体

裁生成宋词,获得了很好的效果^[11]. 互文性源于文档之间显式或隐式地“引用”,“引用”在文献引用分析、复杂网络和社区发现等研究中已发现符合幂律分布.

性质 7 体裁互文性依关系强度分为强、弱和潜互文关系.

信息空间中,体裁互文性是由任务上下文或信息行为产生的. 互文关系强度与文档在任务中的共现概率有关. 假如文档偶尔在信息任务中共现,可称之为具有潜互文关系;文档较多地在任务中共现,互文关系强度达到了一定阈值,导致体裁出现了交叉、融合,可称之为弱互文关系;假如文档的互文关系强度很强,属于同一体裁,可称为强互文关系.

基于以上体裁互文性性质的研究分析,可为体裁互文性的度量提供基础. 根据体裁互文性的性质,体裁互文性可以由互文强度度量. 互文强度的计算,除考虑在任务中的共现次数外,还需考虑互文链所带来的体裁传递.

3 基于体裁互文性的 P2P 上下文信息检索模型

3.1 实验原型系统

实验场景为一个数字化校园,企业级搜索、WWW、FTP、论坛、业务管理系统、统一身份认证等功能齐全. 实验方案包括三个部分:实验系统、数据集、实验设计. 实验原型系统如图 1 所示,主要包括三个部分组成:任务上下文服务器、客户端和搜索引擎. 独立的任务上下文服务器主要对文档体裁的生成和管理,包括体裁管理器、任务上下文管理器、互文管理器以及相应数据库和接口等. 客户端主要对用户当前和历史任务进行监控和分析,推测用户查询意图,与上下文服务器一起识别兴趣文档体裁,包括事件监视器、任务分析器、体裁推测器等模块. 搜索引擎主要通过爬虫收集 Intranet 文档,建立内容和体裁索引,进行文档——体裁二维相关性匹配,返回用户感兴趣的文档,包括内容索引库、体裁索引库、排序模块等.

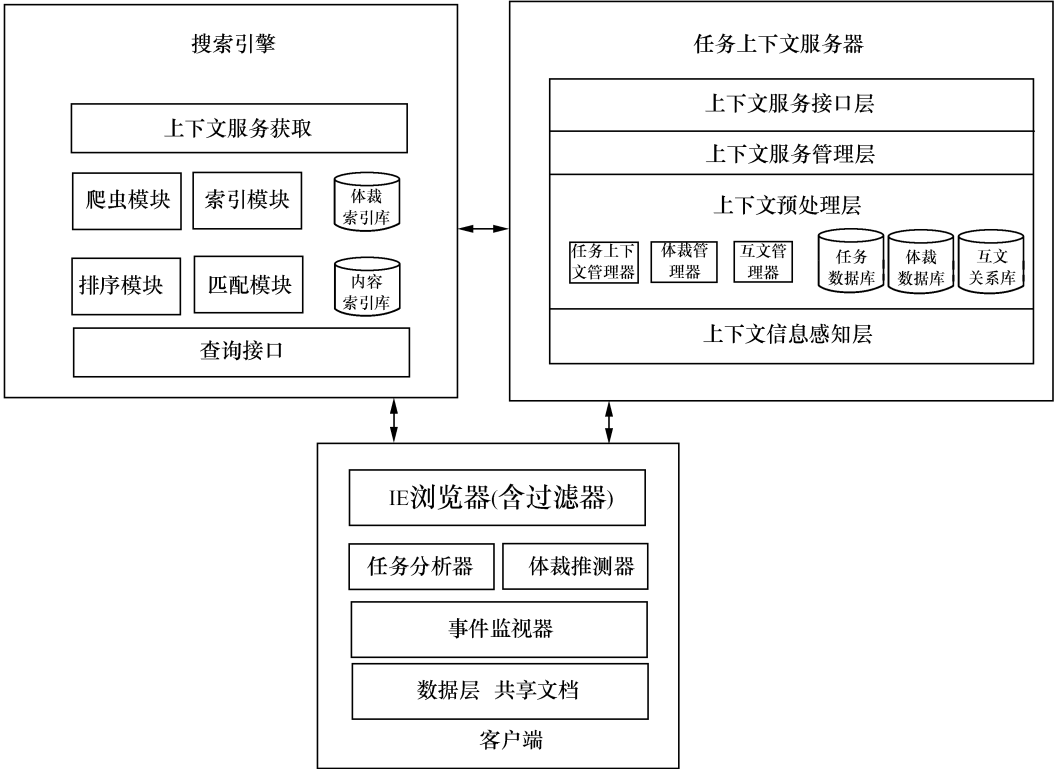


图 1 基于体裁互文性的 P2P 上下文信息检索模型

Fig. 1 A P2P information retrieval model based on the intertextuality of genre

本文 P2P 原型系统基于实验室已自主研发的石鼓 SCool 对等文件共享系统开发,系统基于 Java 技术,在此基础上,增加和完善了体裁索引、

体裁匹配等功能. 其中,任务上下文服务器基于 SOA 架构设计,客户端的任务监控与分析借鉴国际上比较成熟的 Swish 系统方法来建立,体裁过

滤器和预测器采用基于 JADE 标准的代理开发框架.

3.2 数据集

3.2.1 测试集

目前主流 TREC、NTCIR、CLEF、SEWM、搜狗和语言学领域的语料库均未见有带有体裁标注的文档语料集,数据集的来源于已运行 5 年的 SCool 校园搜索引擎所抓取的文档资源,以及若干不同专业背景的师生志愿者实验学习,依照

TREC HARD (High Accuracy Retrieval from Documents) 和 Enterprise 集标准和方法构建,借鉴 Pooling 做法,进行人工标注,建立约 3000 篇文档和 200 个查询,形成测试集.

3.2.2 训练集

实验数据来源于原型系统的任务数据库、互文关系库、体裁数据库,选取约 500 篇文档进行人工的目的性体裁标注,形成基本体裁集.图 2 显示了系统体裁生成与演化的过程.

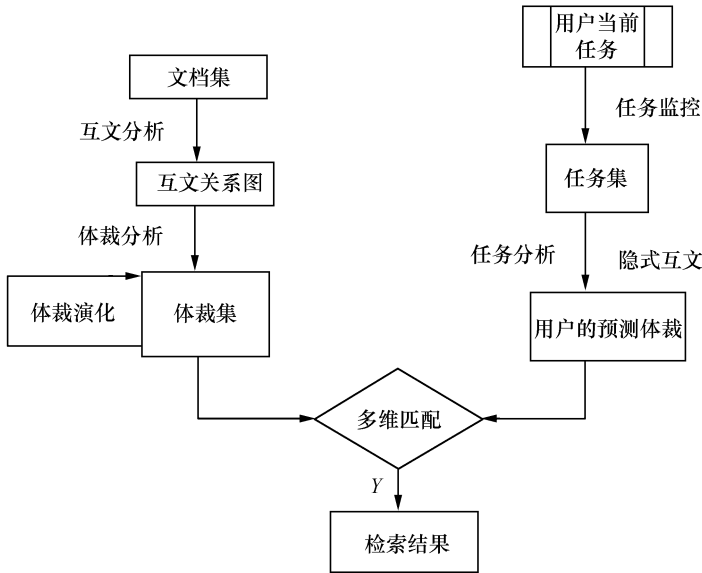


图 2 体裁生成与演化流程图

Fig. 2 Flow of the genre generation and evolution

4 实验比较及结果分析

实验中选取 Gnutella 作为对比,Gnutella 是使用宽度优先搜索方式,这种方式鲁棒性强,但是搜索效率欠佳,其扩散搜索将导致消息呈指数增加.本文从 Bpref (Binary Preference) 指标来考察本文设计方法的搜索性能.指标 Bpref 考虑返回结果列表中经过判断后的文档进行评价,在测试集相关性判断不完全的情况下具有较好的应用效果.

实验收集了实验室开发的 P2P 信息检索系统 SCool 在真实情况下运行了 24 个小时的数据.在实验中假设用户点击的文档为相关文档,其他为非相关文档.对查询日志和点击日志进行处理可以获取训练数据,然后可使用 Ranking SVM 学习排序函数.文件访问日志中记录了用户访问文件的信息,这些信息代表了文件自身的权重.比如最近访问时间、文件的访问频率等.通常越是最近访问的文件,用户更倾向于再次访问,在排序时应

该排的更前.同样访问次数越频繁的文件,也应该排的更前.利用最近访问时间,访问频率等信息,可以计算文件的权重用于排序.

实验: SCool 与 Gnutella、Social-P2P 的 Bpref 指标比较

随着测试集的增加,对所有文档进行相关性判断的难度越来越大. Bpref 对相关性判断不完整的文档集合具有很好的作用, Bpref 的计算公式如下所示:

$$Bpref = \frac{1}{R} \sum_r \left(1 - \frac{r\text{-之前的非相关文档数 } n}{R} \right)$$

(R 为相关文档数)

图 3 显示了三种方法在 Bpref 这项指标上的比较结果.由图可以看出, Gnutella 在不考虑相关性的前提下性能会有所提升,其最大 Bpref 值接近 0.6. 而 Social-P2P 和 SCool 相比也没有很明显的差别,有时候甚至非常接近.这说明对于现在发展迅速的互联网而言,随着文档数目的迅速增加,

需要建立更好的机器学习方法来判断文档集合的相关性. 由于 SCool 使用体裁互文性进行文档相

关性判断, 因此可以获得较多的相关文档数目, 因此 Bpref 还是可以获得一个较好的数值.

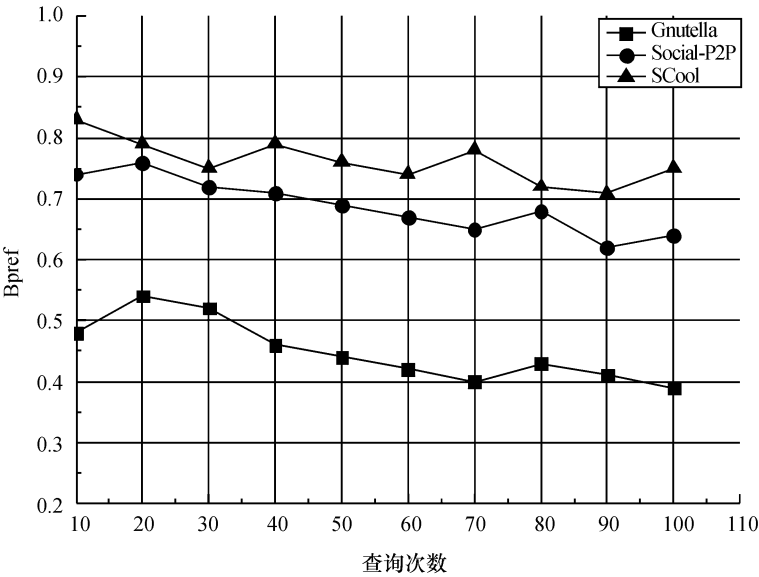


图 3 三种方法 Bpref 指标评价效果

Fig. 3 The evaluation effect of three methods for Bpref indicators

5 结论及展望

石鼓 SCool 方法使用体裁来对文档进行分类, 使用体裁互文性度量表示文档集合的相关性, 在实际应用当中具有较好的性能结果和可操作性. 由于实验当中不可避免的加入了一些主观因素, 对实验结果或多或少会有一点影响, 因此还需要在后续大量的实验中对其加以验证. 另外, SCool 对等存储系统仍然需要进一步的优化和完善, 考虑部署到校园网环境以得到较为广泛的实际应用, 收集跟过的用户数据, 扩大数据和测试集, 用实际应用检验理论设计的当中存在的问题并加以解决. 本文首次以语言学中体裁互文性理论为基础, 为解决信息检索中查询目的预测这个关键问题提供了新的概念和方法, 对于第四代信息检索技术的建立有一定的意义, 同时为语言学的体裁互文性理论研究提供了新的视野和方法.

参考文献:

[1] Shen H T, Shu Y F, Yu B. Efficient semantic-based content search in P2P network[J]. IEEE Trans on Knowledge and Data Engineering, 2004, 16(7): 813-825.

[2] Tang C, Xu Z, Mahalingam M. Peer-to-peer information retrieval using self-organizing semantic overlay networks

[C]//Proceedings of the ACM SIGCOMM. Karlsruhe: ACM Press, 2003.

[3] Sripanidkulchai K, Maggs B, Zhang H. Efficient content location using interest-based locality in peer-to-peer systems[C]//INFOCOM. San Francisco: IEEE Computer Society, 2003.

[4] Daniel A M. Scalable P2P search[J]. IEEE Internet Computing, 2003, 7(2): 83-87.

[5] Zhu Y, Yang X, Hu Y. Making search efficient on gnutella-like p2p systems[C]//Proceedings of the 19th IEEE International Parallel & Distributed Processing Symposium. New York: ACM Press, 2005.

[6] Huang J C, Li X Q, Wu J. A class-based search system in unstructured P2P networks[C]//21st International Conference on Advanced Networking and Applications. New York: ACM Press, 2007.

[7] [法]朱丽娅·克里斯蒂娃. 符号学: 意义分析研究[M]. 朱立元著, 上海文艺出版社, 1993.

[8] 纪卫宁. 文学体裁互文性的深层含义[J]. 求索, 2010(9): 188-190.

[9] 辛斌. 语篇互文性的语用分析[J]. 外语研究, 2000(3): 14-16.

[10] 辛斌. 体裁互文性的社会语用学分析[J]. 外语学刊, 2002(2): 15-21.

[11] 周昌乐, 游维, 丁晓君. 一种宋词自动生成的遗传算法及其机器实现[J]. 软件学报, 2010, 21(3): 427-437.