

自然语言是正则语言

师越¹ 师海忠²

(图科技大数据研究中心 兰州 730070)¹ (西北师范大学数学与统计学院 兰州 730070)²

摘要 自然语言由字母集、单词集、句子集、段落集和文章集5部分组成,而且,字母集包含于单词集,单词集包含于句子集,句子集包含于段落集,段落集包含于文章集。在此观点下,自然语言是正则语言。引入了字母空图和字母空图语言等10个概念。作为特例,英语由英语字母集、英语单词集、英语句子集、英语段落集和英语文章集5部分构成。在此观点下,英语是正则语言。引入了英语字母空图和英语字母空图语言等10个概念。汉语由汉字集、汉语词汇集、汉语句子集、汉语段落集和汉语文章集5部分构成。在此观点下,汉语是正则语言。引入了汉字空图和汉字空图语言等10个概念。这为计算机自然语言处理打开了一扇新的大门,开辟了语言学新的研究领域。

关键词 自然语言,英语,汉语,正则语言,自然语言处理,文章集,英语文章集,汉语文章集,文章有向图语言,英语文章有向图语言,汉语文章有向图语言

中图分类号 TP391 文献标识码 A

Natural Languages Are Regular Languages

SHI Yue¹ SHI Hai-zhong²

(Research Center of Big Data & Graph Tech, Lanzhou 730070, China)¹

(College of Mathematics and Statistics, Northwest Normal University, Lanzhou 730070, China)²

Abstract Natural languages are consisted of alphabet set, word set, sentence set, paragraph set and essay set. Moreover, alphabet set is included in word set; word set is included in sentence set; sentence set is included in paragraph set; paragraph set is included in essay set. In this way, natural languages are regular languages. As a specific example, English is consisted of English alphabet set, English word set, English sentence set, English paragraph set and English essay set. In this way, English is regular language. We also introduced ten concepts including English alphabet empty-graph, English alphabet empty-graph language, etc. Chinese is consisted of Chinese character set, Chinese word set, Chinese sentence set, Chinese paragraph set and Chinese essay set. In this way, Chinese is regular language. We also introduced ten concepts including Chinese character empty-graph, Chinese character empty-graph language, etc. These concepts and theories not only offer a new solution for natural language processing, but also establish a new research field for linguistics.

Keywords Natural language, English, Chinese, Regular language, Natural language processing, Essay set, English essay set, Chinese essay set, Essay digraph language, English essay digraph language, Chinese essay digraph language

1 引言

语言在我们的生活中起着重大的作用。语言的功能很强,使人类区别于其他动物^[1]。语言学历来被看成是典型的人文科学。数学是最重要的自然科学,这两门表面上如此不同的学科之间竟然存在着深刻的内在联系^[2]。数理语言学成了语言学、数学、计算机科学、人工智能等学科所共同关注的重要领域^[2]。乔姆斯基(Avram Noam Chomsky)最初从产生语言的角度研究语言。1956年,通过抽象,他将语言形式地定义为由一个字母表中的字母组成的一些串的集合;对任何语言 L ,有一个字母表 Σ ,使得 $L \subseteq \Sigma^*$ ^[3]。采用计算机技术来研究和处理自然语言是20世纪40年代末期和20世纪50年代才开始的,60多年来,这项研究取得了长足的发展,形成了

计算机科学中一门重要的新兴学科——自然语言处理(Natural Language Processing, NLP)^[4]。D. Jurafsky 和 J. U. Martin 指出:英语和其他自然语言是正则语言吗?在“英语是符号串的集合”的观点下,英语不是正则语言^[4]。

本文第2节为基本知识;第3节在“自然语言有字母集、单词集、句子集、段落集和文章集5部分组成”的观点下自然语言是正则语言;第4节为英语是正则语言;第5节为汉语是正则语言;最后总结全文。

2 基本知识

定义 1^[5] 称 $D=(V, A)$ 是有向图,其中 V 是非空的结点集, A 是弧集,每条弧对应 V 中结点的有序对(不必相异),在本文中, V 是有限集。

师越(1991—),男,主要研究方向为经济网络、大数据分析、(V, R)-语言等, E-mail: haizhong.shi@163.com; 师海忠(1962—),男,博士,教授,主要研究方向为互连网络设计与分析、无(有)向图语言、(V, R)-语言、图半群等(通信作者)。

定义 2^[5] 称 $D=(V,E)$ 是无向图,其中 V 是非空的结点集, E 是边集,每条边对应 V 中结点的无序对(不必相异),在本文中, V 是有限集。

定义 3^[7] 设 V 是一个非空字母表, F_V 是 V 上的自由半群, F_V^* 是 V 上的自由幺半群, E 是 V^2 的子集,且若 $xy \in E$ 则 $yx \in E, \bar{E}=V^2-E$,那么 $F_V^* \bar{E} F_V^*$ 是 F_V 的一个理想, ρ_E 为由此理想产生的 Rees 同余^[6],则称商半群 F_V/ρ_E 为由顶点集 V 和边集 E 的图半群,记为 F_V/ρ_E 。

注意,当 E 为空集时, F_V/ρ_E 即为空图半群 $\{v|v \in V\} \cup \{0\}$ 。

定义 4^[7] 设 V 是一个非空字母表, F_V 是 V 上的自由半群, F_V^* 是 V 上的自由幺半群, A 是 V^2 的子集, $\bar{A}=V^2-A$,那么 $F_V^* \bar{A} F_V^*$ 是 F_V 的一个理想, ρ_A 为由此理想产生的 Rees 同余^[6],则称商半群 F_V/ρ_A 为由顶点集 V 和弧集 A 的有向图半群,记为 F_V/ρ_A 。

定义 5^[8] 设 $G=(V,E)$ 是一个拟简单图, V 为它的顶点集, E 为它的边集, F_V/ρ_E 为 G 的图半群。 $L(F_V/\rho_E)=\{f|f(\rho_E) \in F_V/\rho_E \text{ 且 } f(\rho_E) \neq 0\}$ 是 F_V/ρ_E 的图半群语言,也是 G 的无向图语言 $L(G)$,即 $L(G)=L(F_V/\rho_E)$ 。

定义 6^[9] 设 $D=(V,A)$ 是一个拟简单有向图(可能有有向环但无重有向边)。 V 为它的结点集, A 为它的弧集。 F_V/ρ_A 为 D 的有向图半群, $L(D)=L(F_V/\rho_A)=\{f|f(\rho_A) \in F_V/\rho_A \text{ 且 } f(\rho_A) \neq 0\}$ 是 F_V/ρ_A 的有向图半群语言,也是 D 的有向图语言。

3 自然语言是正则语言

自然语言由如下 5 部分组成:

字母集= $\{\text{字母或标点符号}\}$ 。这里只考虑字母集是有限集。

单词集= $\{\text{单词}|\text{单词是合法的字母串}\}$ 。单个字母或标点符号是合法的字母串,即是单词。单词有有限长度,即单词由有限个字母组成,所以单词集是有限集。

句子集= $\{\text{句子}|\text{句子是合法的“单词”或“单词+空格”或“单词+标点符号+空格”串}\}$ 。单个单词是合法的单词串,即是句子。句子有有限的长度,即句子由有限个“单词”或“单词+空格”或“单词+标点符号+空格”组成,所以句子集是有限集。

段落集= $\{\text{段落}|\text{段落是合法的句子串}\}$ 。单个句子是合法的句子串。段落有有限长度,即段落由有限个句子构成,所以段落集是有限集。

文章集= $\{\text{文章}|\text{文章是合法的段落串}\}$ 。单个段落是合法的段落串,即段落是文章。文章有有限长度,即文章由有限个段落组成,所以文章集是有限集。

显然,字母集 $\subset$ 单词集 $\subset$ 句子集 $\subset$ 段落集 $\subset$ 文章集。由此可见,自然语言也是文章集。

字母集对应空图 $D(N,\emptyset)$, N =字母集, $\emptyset$ 是空集。

字母集产生单词有向图 $D_n(N,WA)$: N =字母集, WA = $\{\text{字母 1 到字母 2 连一有向边}|\text{字母 2 可合法跟在字母 1 的后面}\}$ 。

单词集产生句子有向图 $D_w(W,SA)$: W =单词集, SA = $\{\text{单词 1 到单词 2 连一有向边}|\text{单词 2 可合法跟在单词 1 的后面}\}$ 。

句子集产生段落有向图 $D_s(S,PA)$: S =句子集, PA = $\{\text{句子 1 到句子 2 连一有向边}|\text{句子 2 可合法化跟在句子 1 的后面}\}$ 。

段落集产生文章有向图 $D_p(P,EA)$: P =段落集, EA = $\{\text{段落 1 到段落 2 连一有向边}|\text{段落 2 可合法化跟在段落 1 的后面}\}$ 。

以上 5 种有向图依次产生 5 种有向图语言:

字母空图语言 $L(D(N,\emptyset))$;

单词有向图语言 $L(D_n(N,WA))$;

句子有向图语言 $L(D_s(W,SA))$;

段落有向图语言 $L(D_p(S,PA))$;

文章有向图语言 $L(D_p(P,EA))$ 。

引理 1^[7] 有向图语言是正则语言。

因此有,

定理 1 1)空图语言 $L(D(N,\emptyset))$ 是正则语言;

2)单词有向图语言 $L(D_n(N,WA))$ 是正则语言;

3)句子有向图语言 $L(D_s(W,SA))$ 是正则语言;

4)段落有向图语言 $L(D_p(S,PA))$ 是正则语言;

5)文章有向图语言 $L(D_p(P,EA))$ 是正则语言。

因为自然语言是文章集,而文章集是相应文章有向图语言的有限子语言,所以有,

定理 2 自然语言是正则语言。

下面以英语和汉语为例讨论它们的正则性。

4 英语是正则语言

英语由如下 5 部分组成:

英语字母集= $\{\text{英语字母或标点符号}\}$ 。显然英语字母集是有限集。

注意:考虑科技文章(例如化学、数学等)时,英语字母集里还要包含相应的科技符号。

英语单词集= $\{\text{英语单词}|\text{英语单词是合法的英语字母或标点符号串}\}$ 。单个英语字母或标点符号是合法的字母串,即是英语单词。英语单词有有限长度,即英语单词由有限个英语字母组成,所以英语单词集是有限集。

英语句子集= $\{\text{英语句子}|\text{英语句子是合法的“英语单词”或“英语单词+空格”或“英语单词+标点符号+空格”串}\}$ 。单个英语单词是合法的单词串,即是英语句子。英语句子有有限的长度,即英语句子由有限个“英语单词”或“英语单词+空格”或“英语单词+标点符号+空格”组成,所以英语句子集是有限集。

英语段落集= $\{\text{英语段落}|\text{英语段落是合法的英语句子串}\}$ 。单个英语句子是合法的英语句子串。英语段落有有限长度,即英语段落由有限个英语句子构成,所以英语段落集是有限集。

英语文章集= $\{\text{英语文章}|\text{英语文章是合法的英语段落串}\}$ 。单个英语段落是合法的英语段落串,即英语段落是英语文章。英语文章有有限长度,即英语文章由有限个英语段落组成,所以英语文章集是有限集。

显然,英语字母集 $\subset$ 英语单词集 $\subset$ 英语句子集 $\subset$ 英语段落集 $\subset$ 英语文章集。由此可见,英语也是英语文章集。

英语字母集对应空图 $ED(EN,\emptyset)$, EN =英语字母集, $\emptyset$ 是空集。

英语字母集产生英语单词有向图

$ED_n(EN, EWA)$; EN = 英语字母集, EWA = {英语字母(或标点符号)1 到英语字母(或标点符号)2 连一有向边 | 英语字母(或标点符号)2 可合法跟在英语字母(或标点符号)1 的后面}。

英语单词集产生英语句子有向图

$ED_w(EW, ESA)$; EW = 英语单词集, ESA = {英语单词 1 到英语单词 2 连一有向边 | 英语单词 2 可合法跟在英语单词 1 的后面}。

英语句子集产生英语段落有向图

$ED_s(ES, EPA)$; ES = 英语句子集, EPA = {英语句子 1 到英语句子 2 连一有向边 | 英语句子 2 可合法化跟在英语句子 1 的后面}。

英语段落集产生英语文章有向图

$ED_p(EP, EEA)$; EP = 英语段落集, EEA = {英语段落 1 到英语段落 2 连一有向边 | 英语段落 2 可合法化跟在英语段落 1 的后面}。

以上 5 种有向图依次产生 5 种有向图语言:

英语字母空图语言 $L(ED(EN, \emptyset))$;

英语单词有向图语言 $L(ED_n(EN, EWA))$;

英语句子有向图语言 $L(ED_s(EW, ESA))$;

英语段落有向图语言 $L(ED_p(ES, EPA))$;

英语文章有向图语言 $L(ED_p(EP, EEA))$ 。

由引理 1 知:

推论 1 1) 英语空图语言 $L(ED(EN, \emptyset))$ 是正则语言, 英语字母集是英语空图语言的有限子语言;

2) 英语单词有向图语言 $L(ED_n(EN, EWA))$ 是正则语言, 英语单词集是英语单词有向图语言的有限子语言;

3) 英语句子有向图语言 $L(ED_s(EW, ESA))$ 是正则语言, 英语句子集是英语句子有向图语言的有限子语言;

4) 英语段落有向图语言 $L(ED_p(ES, EPA))$ 是正则语言, 英语段落集是英语段落有向图语言的有限子语言;

5) 英语文章有向图语言 $L(ED_p(EP, EEA))$ 是正则语言, 英语文章集是英语文章有向图语言的有限子语言。

因为英语是英语文章集, 而英语文章集是相应英语文章有向图语言的有限子语言, 所以英语是有限语言, 当然也是自然语言, 由定理 2 知:

推论 2 英语是正则语言。

5 汉语是正则语言

汉语由以下 5 部分组成:

汉字集 = {汉字或标点符号}。显然汉字集是有限集。

注意: 考虑科技文章(例如化学、数学等)时, 汉字集里还要包含相应的科技符号。

汉语词汇集 = {汉语词汇 | 汉语词汇是合法的汉字串}。单个汉字或标点符号是合法的汉字串, 即是汉语词汇。汉语词汇有有限长度, 即汉语词汇由有限个汉字或标点符号组成, 所以汉语词汇集是有限集。

汉语句子集 = {汉语句子 | 汉语句子是合法的“汉语词汇”或“汉语词汇 + 标点符号”串}。单个汉语词汇是合法的汉语词汇串, 即是汉语句子。汉语句子有有限的长度, 即汉语句子由有限个“汉语词汇”或“汉语词汇 + 标点符号”组成, 所以汉

语句子集是有限集。

汉语段落集 = {汉语段落 | 汉语段落是合法的汉语句子串}。单个汉语句子是合法的汉语句子串, 即是汉语段落。汉语段落有有限长度, 即汉语段落由有限个汉语句子构成, 所以汉语段落集是有限集。

汉语文章集 = {汉语文章 | 汉语文章是合法的汉语段落串}。单个汉语段落是合法的汉语段落串, 即汉语段落是汉语文章。汉语文章有有限长度, 即汉语文章由有限个汉语段落组成, 所以汉语文章集是有限集。

显然, 汉字集 $\subset$ 汉语词汇集 $\subset$ 汉语句子集 $\subset$ 汉语段落集 $\subset$ 汉语文章集。由此可见, 汉语也是汉语文章集。

汉字集对应汉字空图 $CD(CN, \emptyset)$, CN = 汉字集, $\emptyset$ 是空集。

汉字集产生汉语词汇有向图

$CD_n(CN, CWA)$; CN = 汉字集, CWA = {汉字(或标点符号)1 到汉字(或标点符号)2 连一有向边 | 汉字(或标点符号)2 可合法跟在汉字(或标点符号)1 的后面}。

汉语词汇集产生汉语句子有向图

$CD_w(CW, CSA)$; CW = 汉语词汇集, CSA = {汉语词汇 1 到汉语词汇 2 连一有向边 | 汉语词汇 2 可合法跟在汉语词汇 1 的后面}。

汉语句子集产生汉语段落有向图

$CD_s(CS, CPA)$; CS = 汉语句子集, CPA = {汉语句子 1 到汉语句子 2 连一有向边 | 汉语句子 2 可合法化跟在汉语句子 1 的后面}。

汉语段落集产生汉语文章有向图

$CD_p(CP, CEA)$; CP = 汉语段落集, CEA = {汉语段落 1 到汉语段落 2 连一有向边 | 汉语段落 2 可合法化跟在汉语段落 1 的后面}。

以上 5 种有向图依次产生 5 种有向图语言:

汉字空图语言 $L(CD(CN, \emptyset))$;

汉语词汇有向图语言 $L(CD_n(CN, CWA))$;

汉语句子有向图语言 $L(CD_s(CW, CSA))$;

汉语段落有向图语言 $L(CD_p(CS, CPA))$;

汉语文章有向图语言 $L(CD_p(CP, CEA))$ 。

由引理 1 知:

推论 3 1) 汉字空图语言 $L(CD(CN, \emptyset))$ 是正则语言, 汉字集是汉字空图语言的有限子语言;

2) 汉语词汇有向图语言 $L(CD_n(CN, CWA))$ 是正则语言, 汉语词汇集是汉语词汇有向图语言的有限子语言;

3) 汉语句子有向图语言 $L(CD_s(CW, CSA))$ 是正则语言, 汉语句子集是汉语句子有向图语言的有限子语言;

4) 汉语段落有向图语言 $L(CD_p(CS, CPA))$ 是正则语言, 汉语段落集是汉语段落有向图语言的有限子语言;

5) 汉语文章有向图语言 $L(CD_p(CP, CEA))$ 是正则语言, 汉语文章集是汉语文章有向图语言的有限子语言。

因为汉语是汉语文章集, 而汉语文章集是相应汉语文章有向图语言的有限子语言, 所以汉语是有限语言, 当然也是自然语言, 由定理 2 知:

推论 4 汉语是正则语言。

结束语 本文指出, 自然语言由字母集、单词集、句子集、段落集和文章集 5 部分组成, 而且, 字母集 $\subset$ 单词集 $\subset$ 句子集

段落集、文章集。进一步引入了字母空图、单词有向图、句子有向图、段落有向图和文章有向图以及相应的字母空图语言、单词有向图语言、句子有向图语言、段落有向图语言和文章有向图语言。由此可知,自然语言是正则语言。作为特例,英语由英语字母集、英语单词集、英语句子集、英语段落集和英语文章集 5 部分组成,还引入了英语字母空图以及相应的英语字母空图语言等概念。由此可知,英语是正则语言。汉语由汉字集、汉语词汇集、汉语句子集、汉语段落集和汉语文章集 5 部分组成,还引入了汉字空图以及相应的汉字空图语言等概念。由此可知,汉语是正则语言。这些概念的引入为用计算机处理自然语言打开了一扇新的大门,进而形成自然语言处理的新领域。

参考文献

[1] Bloomfield L. 语言学[M]. 袁家骅,赵世开,甘世福,译. 北京:商

务印书馆,2008

- [2] 冯志伟,胡凤国. 数理语言学 [M]. 北京:商务印书馆,2012
- [3] 蒋宗礼,姜守旭. 形式语言与自动机理论(第 2 版)[M]. 北京:清华大学出版社,2007
- [4] Jurafsky D, Martin J H. 自然语言处理综论 [M]. 冯志伟,孙乐,译. 北京:电子工业出版社,2005
- [5] Bondy J A, Murty U S R. Graph Theory with Applications [M]. The Macmillan Press Ltd, 1976
- [6] Howie J M. An Introduction to Semigroup Theory [M]. Academic Press, London, 1976
- [7] 师海忠,师越. (V, R)-语言[J]. 计算机科学, 2014, 41(6A): 33-36
- [8] 师海忠. 无向图语言[J]. 计算机科学, 2011, 38(6): 259-261, 274
- [9] 师海忠. 有向图语言[J]. 计算机工程与应用, 2011, 47(22): 53-56

(上接第 42 页)

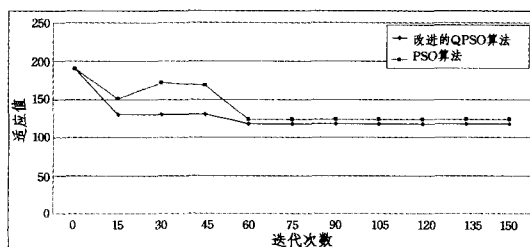


图 7 $QD=20, QDJ=30, QB=40$ 时, 两种算法适应值对比曲线图

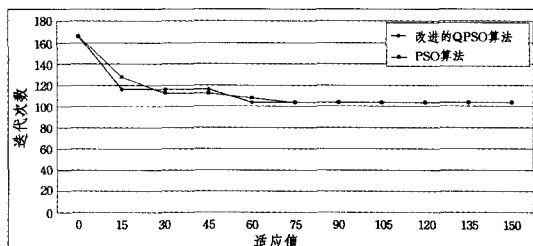


图 8 $QD=25, QDJ=35, QB=40$ 时, 两种算法适应值对比曲线图

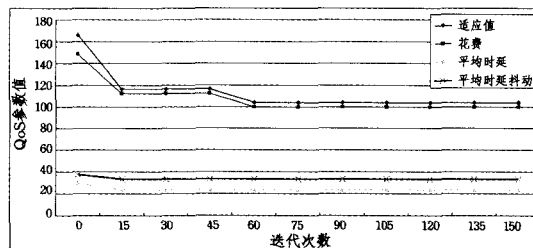


图 9 $QD=25, QDJ=35, QB=40$ 时, 改进的 QPSO 算法各 QoS 参数曲线图

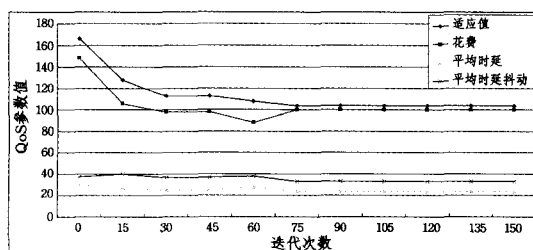


图 10 $QD=25, QDJ=35, QB=40$ 时, PSO 算法各 QoS 参数曲线图

图 9 和图 10 给出了两种算法在约束 2 条件下各 QoS 参数动态曲线图。其中平均时延和平均时延抖动表示组播树中从源节点到目的节点各单播路径的时延和时延抖动的平均值。经过对比可以发现,改进的 QPSO 算法比 PSO 算法更加稳定,且具有更快的全局收敛速度和更高的收敛效率,尤其在时延和时延抖动等 QoS 参数上表现得更加明显。

结束语 本文用量子粒子群算法解决 QoS 组播路由问题,同时引入了遗传算法中的交叉算子与选择算子,大大提高了算法的执行效率。采用预处理机制进行网络拓扑的转换,极大地简化了算法编解码操作。同时在算法运行过程中,通过动态地调整参数 α 的值来更好地平衡算法的局部开发能力和全局搜索能力。仿真实验表明:该算法高效快速,具有比粒子群算法更好的全局寻优能力。

参考文献

- [1] WANG Zheng, CORWCROFT J. Quality-of-service routing for supporting multimedia applications[J]. IEEE Journal on Selected Areas in Communications, 1996, 14(7): 1228-1234
- [2] 孙俊. 量子行为粒子群优化算法研究[D]. 无锡:江南大学, 2009
- [3] Clerc M, Kennedy J. The particle swarm: explosion, stability and convergence in a multi-dimensional complex space [J]. IEEE Transaction on Evolutionary Computation, 2002, 6(2): 58-73
- [4] Gen M, Cheng R W. Genetic algorithms and engineering design [M]. New York: John Wiley & Sons, 1996
- [5] Leela R, Thanulekshmi N, Selvakumar S. Multi-constraint QoS unicast routing using genetic algorithm (MURUGA) [J]. Applied Soft Computing, 2011(11): 1753-1761
- [6] Wang Hua, Xu Hong, Yi Shan-wen, et al. A tree-growth based ant colony algorithm for QoS multicast routing problem[J]. Expert Systems with Applications, 2011(38): 11787-11795
- [7] 刘萍, 高飞, 杨云. 基于遗传算法和蚁群算法融合的 QoS 路由算法[J]. 计算机应用研究, 2007, 24(9): 224-227
- [8] 王征应, 石冰心. 基于启发式遗传算法的 QoS 组播路由问题求解[J]. 计算机学报, 2001, 24(1): 55-61