摘　要

从复杂网络这一崭新的视角对汉语进行系统的探索，无论是对汉语本体研究还是对中文信息处理，都具有方法论意义上的创新性,并且涉及复杂系统、语言学、自然语言处理、机器学习、统计学等多学科的交叉，因此具有十分重要的科学意义，已成为当前自然语言处理的研究前沿与热点之一。

本文将主要介绍以下几方面的研究成果：(1) 利用目前可能得到的汉语资源，构造了覆盖词法、句法、语义不同层次的各种类型大规模汉语语言网络，他们包括汉语词同现网络、依存句法网络、标签语义网络和查询词语义网络，并对上述语言网络的性质进行分析与对比；(2) 基于汉语语言网络，提出基于二部图的资源分配算法进行查询词推荐，取得了较以往算法更好的效果。

本研究对汉语语言网络全面、深入的考察与研究, 在一定程度上丰富和深化对汉语的科学认识，得出的一系列结果或结论对汉语语言学、语言认知、中文信息处理等均具有一定参考价值。

关键词：复杂网络　语言网络　汉语　自然语言处理

Abstract

A systematic investigation of Chinese, no matter on Chinese ontology or Chinese information processing, from a brand-new perspective of complex network, is a great innovation at methodology level. It has referred to various disciplines, including complex system, linguistics, natural language processing, machine learning and statistics. Therefore, it is of great scientific significance and has become the hot frontier issue of natural language processing.

The thesis will introduce the following research results mainly: (1) I construct different types of large-scale Chinese language networks, including word co-occurrence network, syntactic network and blog tag co-occurrence network and query semantic network. Afterwards I analyze and compare the statistical, properties of these language networks. (2) I present a resource allocation method based on bipartite language network of query logs to recommend related queries, and achieve great improvement and significant features than previous methods.

These research and investigation on Chinese language networks will enrich and intensify the scientific cognition on Chinese. And the research findings and conclusions will exert some referential value to Chinese linguistics, cognitive linguistics and Chinese information processing.

Keywords: complex network language network Chinese natural language processing

目　录

第1章 引言

1.1 课题背景及意义

语言是人类进化的重要产物，也是人类区别于动物的根本标志。经过上百万年的演化，人类语言已经形成一个复杂系统。处在当今信息急遽增长的时代，如何让计算机更好地处理乃至理解人类语言，已成为迫切需要。而二十一世纪之初兴起的复杂网络理论，为我们提供了一个崭新的视角来探索人类语言的本质，并已经在许多语言上取得了初步的令人振奋的成果。而利用复杂网络理论对汉语进行系统而深入的探索，无论是对肇始于一百余年前《马氏文通》的汉语本体研究，还是对蓬勃发展于近三十年的中文信息处理研究，都具有方法论意义上的创新性。进一步地，如果我们把从汉语出发的这种研究投射到丰富多彩的人类语言坐标系中去，就很可能会超越汉语本身，从而对人类语言本质的探究做出贡献。加之这个研究课题涉及到复杂系统，语言学，自然语言处理，机器学习，统计学等多学科的交叉，因而具有十分重要的科学意义。同时，相关成果会对自然语言处理诸多应用研究(如文本信息检索，Web2.0 架构下的搜索引擎，数字图书馆，知识管理，个性化服务等)产生重要的借鉴作用，应用前景广阔，因而也已成为当前自然语言处理的研究前沿与热点之一。

语言网络已经吸引了包括来自物理学，心理学，语言学，认知科学和计算机科学等各个领域学者的关注。物理学家把语言网络作为众多复杂网络之一，研究如何建立统一的复杂网络理论。心理学家则试图通过语言网络研究人类的语言习得等能力。语言学家着眼于通过语言网络研究语言的本质和演化等性质。认知科学家尝试把语言网络和大脑联结起来，探索人类大脑运行机制。计算机科学家则在探索如何用语言网络更好地让计算机处理和理解人类语言，并在自然语言处理领域逐渐引起关注，著名计算语言学学术会议 HLT-NAACL 已经在 2006 年和 2007 年连续两年召开以基于语言网络的自然语言处理方法为主题的研讨会(TextGraphs Workshop)，并在 2008 的著名计算语言学学术会议 COLING 上继续召开第三届。语言网络为各领域提供了一个新的视角重新审视原有的问题，并受到越来越多的

学者的关注，同时目前的研究仍处于探索阶段，在理论研究和关键应用技术研究方面都非常薄弱，给我们留下了极具前瞻性的探索空间。本文将总结、分析和比较在语言网络研究方面的已有成果， 并对未来在语言网络上的研究方向进行展望。

语言网络研究内容主要包括语言网络基本统计性质，各层次上的网络结构，网络演化性质和模型以及在自然语言处理中的应用等几个方面。本文第 1.2 节主要介绍语言网络的理论基础，即复杂网络理论。在第 1.3 节我们主要介绍，分析和总结在语言网络上的研究成果。第 1.4 节我们展望语言网络的未来研究方向。

1.2 复杂网络理论与方法

众所周知，自然界中存在的大量复杂系统都可以通过网络来描述，世纪之交在真实网络上的两个重要发现， 奠定了复杂网络理论的基础。它们分别是网络的小世界现象(small world phenomenon)和无标度性质(scale free property)。真实复杂网络的平均最短路径(average path length)很短而同时聚类系数(clustering coefficient)较大，被称为小世界现象[1]。无标度性质则揭示了网络的连接度呈幂律(power law)分布这一重要性质[2]。复杂网络理论迅速受到语言学，信息科学，生物学和社会学等各领域的关注[3]，短短的几年内，取得了丰硕的研究成果。在对这个领域的研究现状进行全景式的分析和归纳后，我们认为目前复杂网络相关理论与方法的研究主要包括网络性质，网络结构及其演化三个相辅相成的基本方面[3-8]。

1.2.1 网络性质

人们研究和总结了复杂网络的许多统计性质，其中最重要的是小世界现象和无标度性质。假设一个节点总数为 N 和连边数为 M 的真实语言网络 $G(V,E)$，其中的节点集为 $V=\{v_i\}$，设网络节点的平均度为 $\bar{k}$，下面介绍语言网络研究中比较重要的几个性质。

1.2.1.1 小世界现象

20 世纪的后 40 年里，Erdos 和 Renyi 建立的随机网络模型[9](ER 模型)

一直被人们认为是真实复杂网络的最佳模型。但由于大多数实际的复杂网络并不是随机连接的，ER 模型作为复杂网络的模型，无疑存在着较大缺陷。几乎与此同时，人们还开展了对“小世界”效应的实验研究，并提出最著名的六度分离推断[10]。1998 年，Watts 和 Strogatz 将小世界模型引入对复杂网络的研究，称为 WS 模型[1]。稍后 Newman 和 Watts 对该模型进行了改进，建立了 NW 模型[11]。这两个小世界模型本质上是一样的，它们都反映了实际复杂网络的一个性质，即大部分节点只与它们的邻近节点相连，同时也有某些节点可与非邻近节点直接相连。

通常从如下两个角度观察小世界现象，平均最短路径长度和聚合系数。平均最短路径长度是网络中两节点之间的平均距离。具有小世界性质的网络的平均最短路径会很短，远小于网络规模，这也是“小世界”命名的原因。设平均最短路径为 L，对“小世界”网络有 $L \approx \ln(N)/\ln(\bar{k})$。随机图模型和小世界模型在这一点上对复杂网络的刻画都比较恰当。一个节点的聚合系数反映了其相邻节点所构成集合的聚集程度。整个网络的聚合系数 C 是每个节点 i 的聚合系数 C_i 的平均值 $0 \le C \le 1$。在极端情况下，当网络所有节点均为孤立节点时，$C=0$；当网络所有节点为全耦合节点时，每个节点与其余 $N-1$ 节点均有连接，$C=1$。对一个包含 N 个节点的 ER 随机图网络，当 N 很大时，有 $C \approx \bar{k}/N$，即其聚合系数远小于 1。而实际复杂网络表现出显著的聚合效应，即聚合系数 C 虽然小于 1，但比 $O(N^{-1})$ 要大得多。

1.2.1.2 无标度性质

对复杂网络进行研究的另一个重要方面是节点的度分布(degree distribution)。ER 模型和 WS，NW 模型给出的度分布近似泊松分布[3]。但大量研究表明实际复杂网络的度分布明显不同于泊松分布，而更接近于幂律(power law)分布，即 $\Pr(k) \propto k^{-\gamma}$，其中 $\Pr(k)$ 是度为 k 的节点出现在网络中的概率，γ 为常数。泊松分布的图形一般在网络平均度 $\bar{k}$ 处有一个峰值，网络中大部分节点都集中在附近，因此称 $\bar{k}$ 为网络的特征标度

(characteristic scale)，而在幂律分布中则没有这样的峰值，因此又被称为无标度性质[12]。

Barabasi 和 Albert[2]认为实际复杂网络的两个重要性质导致了无标度性质:(1)增长性，即网络规模不断扩大同时其自身在不断演化；(2)优先连接性，即新的节点更倾向于与那些具有更高连接度的节点连接，表现出“马太效应”。这两个性质导致了复杂网络中节点的度分布服从幂律分布，存在少量度相对很高的节点，但绝大多数节点的度相对很低(即存在所谓的“长尾”)。在此基础上，他们提出了无标度网络模型，即 BA 模型。

1.2.1.3 其他性质

除了小世界现象和无标度性质外，研究者还陆续提出很多的刻画复杂网络的统计量，包括层次性[13]，自相似性[14]，匹配性(反映节点及其邻居节点的连接度的关系)[15]，直径(网络中最短路径的最大值)[3]，效率(度量网络中信息传递的效率)[16]，介数(通过该节点或边的最短路径比例，反映该节点或边的重要程度)[17]，度的相关度[3]，可导航性(在满足小世界网络性质的基础上某节点根据局部信息是否可以有效到达指定节点的可能性)[18]，鲁棒性(移除某些节点后的网络破碎的可能性)[19]，等等。它们分别从某个角度反映了复杂网络的重要统计特征。

1.2.2 网络结构

在网络结构方面，研究者提出了模体(Motif)[20，21]，模体簇[22]和社团结构[23]等概念。网络中频繁出现的子图(一般包括 3，4 个节点)称为模体。如**错误！未找到引用源。**所示是 13 个可能的三元组模体(Triad Significance Profile， TSP)。Milo 等人在 Science 上撰文提出了模体的概念和统计算法，研究发现通过模体能够识别网络的典型局部连接模式[20，24]。特定的几个模体聚集在一起可以形成大的模体簇，这有助于理解网络的增长机制[25]。

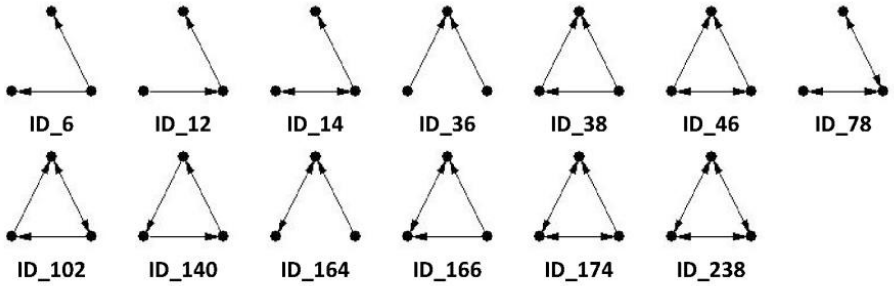

图 1.1 13 个可能的三元组模体。

从更宏观的角度来讲，实际网络都具有一个共同结构特点，即社团结构:每个社团内部连接相对紧密，而社团之间连接比较稀疏。社团结构划分是复杂网络研究的重要问题，目前的主要方法是分级聚类算法，该算法基于各节点之间相似性或连接强度划分网络，主要分为凝聚算法和分裂算法。GN 算法[23]是一种分裂算法，基本思想是不断从网络中移除介数最大的边，其中边介数定义为网络中经过该边的最短路径的比例。此外还有利用其他度量方式的分裂算法，如采用边聚类系数[26]，布朗微粒[27]，相异性指数[28]或信息中心度[29]等代替介数的算法。模块度(modality)是公认为衡量社团划分质量的标准，Newman 基于 GN 算法提出了一种基于模块度的快速凝聚算法[30]，算法首先初始化网络为若干社团，然后依次合并有边相连的社团对，并计算合并后的模块度的增量，根据贪婪算法原理，每次合并都沿着增量最多的方向进行。Brandes 等人对凝聚算法性能进行了深入研究[31]。

在主流的基于分级聚类的社团发现算法之外，还涌现出许多其他算法，如派系过滤(Clique Percolation，CP)算法[32]用来分析相互重叠的社团结构，Flake 提出通过最大流最小割算法解决社团划分问题[33]；Rosvall 则提出了一种利用最小描述长度和模拟退火方法的社团发现算法[34]；图分割也是一种非常流行的社团划分技术，启发式的方法常被用来寻找好的分割，特征谱算法[35-37]就是非常成功的启发式算法之一；一种基于物理学的震荡器原理的社团划分方法被提出来[38]，该原理曾被广泛用于复杂网络可视化；Raghavan 等提出通过根据邻居节点判断自身标签，能够在线性时间复杂度内实现社团划分[39]。Danon 等对常用社团发现算法的性能

和优势进行了细致的比较[40]。此外还有专门针对加权网络[41，42]，有向网络[43，44]，二部图网络[44-46]和随时间演化的网络[47-49]的社团发现算法，是目前研究的新热点和新方向。同时也有研究关注网络结构与功能的关系问题[3，50]。

1.2.3 网络演化

在网络演化方面，2005 年前研究者主要关注网络的全局统计性质随时间的演化，例如平均出入度和平均最短路径，得到了一些有意义的结论，如真实网络一般来说比较稳定，这得益于其层次性结构，即大的网络是由很多小的趋于稳定的网络按照层次组织起来的[51]；另外和直观相悖的重要发现是，随着网络节点的增加，网络边数与节点数满足幂律关系，而网络直径却不断减小[52]。

2005 年后有研究者开始从社团结构演化这一更加深入的角度研究复杂网络的动态性，分析社区的出现，增长，合并，分裂，缩小和消失等现象，得出了若干重要结果。如 Palla 等人在 Science 发表论文，提出加权网络上的 CP 算法(WCP 算法)以及基于重叠度的时间演化上的社团定位方法，并在对论文合作演化网络和电话呼叫演化网络的研究中发现小的社团可以通过保持成员稳定而长期生存，而大的社团则通过频繁更换成员避免突然崩溃[49]。也有研究者提出将不同时间上的相似子图串起来形成元子图(meta group)，转而研究元子图之间的关系[53]。

亦有众多工作试图建立复杂网络的增长模型。在复杂网络理论兴起之前，人们一直以随机网络[9]作为实际网络标准模型。后来复杂网络理论否定了实际网络与随机网络的相似性，提出了许多指示实际网络新特性的模型，其中经典的模型就是小世界模型[1]和无标度网络模型[2]。在这两个基础的网络模型基础上，人们又提出许多改进模型，例如在加权网络上的众多演化模型[3，6]。

1.3 基于复杂网络理论与方法的语言网络研究

复杂网络理论与方法在语言学乃至自然语言处理研究领域已经产生了深远的影响。对语言复杂网络的研究，使得人们可以通过这一特有视角

进一步了解，把握语言及其计算的性质，进而有可能对人类认知，语言本质和语言演化规律等人类共同关心的基本问题进行“形而上”式的探究[50，54，55]。

语言网络大体上可以分为两类[56]，语义网络(semantic networks)和表层网络(superficial networks)。前者主要根据词与词之间的语义关系建立连接，如基于词典资源的语义网络和词汇联想网络等，后者则主要根据词与词在文本中的相对关系，如邻接关系或句法依存关系建立连接，如同现网络和句法网络等。在图 1.2 我们给出一个语义网络的示意图。其中空心箭头表示两词之间在语义上属于上下位关系，如“花⇒百合花”说明百合花是花的一种；而实心箭头表示两词之间在语义上属于部分-整体关系，如“花萼→花”说明花萼是花的一部分。图 1.3 展示了表层网络的构造方法。其中图 1.3(A)表示同现网络的构造方法， 图 1.3(B)显示了由该句构造的子网在同现网络中的位置；图 1.3(C)表示依存句法网络的构造方法，图 1.3(D)显示了由该句构造的子网在句法网络中的位置。

当然，还存在一些两种基本类型的混合或者变型， 以及特定角度的语言网络，如标签网络，音节网络等。在语言复杂网络的统计性质，网络结构，演化方面，以及这些研究在 NLP 的应用方面，已经开展了一系列研究工作。

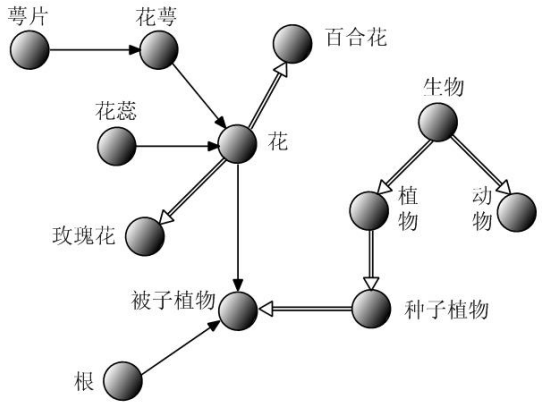

图 1.2 语义网络示意图

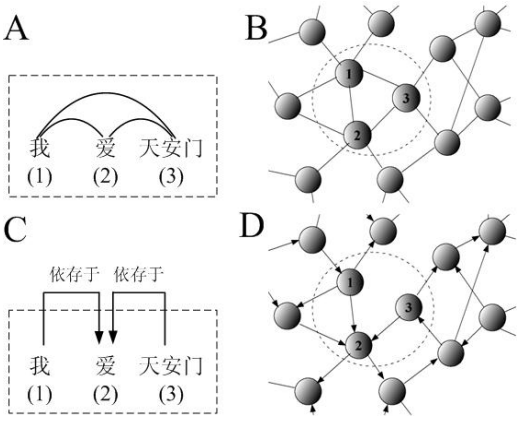

图 1.3 表层网络示意图

1.3.1 网络性质

研究者在众多语言网络上验证了复杂网络的小世界现象和无标度特性等统计性质，这些语言网络包括在 Roget’s Thesaurus[3，25，57]， Moby

Thesaurus II[58]，Merriam-Webster’s Thesaurus[6]，WordNet[57，59]，HowNet[60，61]，Wikipedia[62，63]等语义词典资源上建立的各类语义网络，通过心理学实验得到的词汇联想网络[57，64]，针对各种语言的(英语，汉语，法语，德语，罗马尼亚语，俄语，捷克语等)的同现网络[21，65-67]和句法网络[68-70]。

还有工作关注了音节网络[71，72]等语言网络的统计性质，也有研究者试图探索人类思考机制与 PageRank 算法的相似性[73]。而随着 Web2.0 的兴起，标签系统(tagging system)，又称为大众分类体系(Folksonomy)[74-82]，受到越来越多的关注，作为一种新型的语言网络 (可被视作语义网络和表层网络的某种混合)，标签复杂网络上的研究也方兴未艾[79，83-85]。

在汉语网络上的工作已经有一些工作， 主要包括:Zhou 等人通过北京大学《人民日报》1998 年 1 月份语料库构造汉语同现网络，考察其小世界现象，无标度性质及其它统计性质[86]；刘知远等人在北京大学《人民日报》1998 年上半年 1300 万字左右的人工分词语料库和国家语委 5000 万字左右的人工分词平衡语料库上建立了汉语词同现网络，考察了网络的小世界现象和无标度性质，并得到汉语上的核心词典规模[87]；Liu 主要探讨了如何通过汉语句法树库构造依存句法网络，并在两个小型句法树库(规模为 16，654 词和 19，060 词)上构造汉语句法网络并考察其统计性质[88]；Zlatic 等对 Wikipedia 复杂网络的研究包括了对由汉语词条构成的汉语复杂网络的考察，发现汉语与其它语言复杂网络的统计性质具有很大相似性[63]；也有一些研究考察了我国著名学者董振东教授发明的 HowNet 复杂网络的统计性质[60，61]；Li 等考察了汉语词法网络的统计性质并按照字频进行选字构词构造物理模型，与实际网络符合较好[89，90]；Li 等发现汉语偏旁网络也具有小世界现象和无标度性质[91]。

总的来讲，目前汉语网络上的相关工作虽然已有不少，但还是有一些重要的汉语语言网络未被关注，如汉语语义构词网络，标签网络等，并且已有工作也或多或少地存在观察视角不够全，网络规模太小等问题[86，88]。

1.3.2 网络结构

在语言网络结构方面，Milo 等人在 Science 撰文表示，与生物网络，社会网络和 WWW 网络等相比，多种语言的同现网络(包括英语，法语，日语和西班牙语)在 13 个可能的三元组模体表现出相似的统计性质，构成一个超家族(Superfamily)[24]。也有研究者将社团发现算法和特征谱算法应用在语言网络上，试图通过语言网络的社团结构分析语言的性质[32，36]。 总体上看，针对英语等的网络结构方面的研究还处于刚刚起步阶段，据我们了解，在汉语网络上还没有相应研究，应该说是一个空白。

1.3.3 网络演化

语言演化是探索语言复杂性的重要方面，需要诸多学科的共同努力[92，93]。其中 Nowak 等人将进化博弈论(Evolutionary Game Theory)思想成功地应用到语言演化中来，通过对人类种群中的交流进行建模，研究语音信号与现实对象联系在一起，构造成词和连缀成句的过程，在 Science，Nature 和 PNAS 等期刊发表了许多重要结论[94-104]，如发现当词汇量足够大时，用一定句法结构表述信息的适应度大于没有句法结构的适应度，否则， 没有句法结构的适应度会大于带有句法结构的适应度，这说明了句法结构是自然选择的结果[100]；又如在模型中添加子代对父代的语言的继承因素来考察语言习得机制，发现在假设普遍语法(Universal Grammar)中各种语言完全对等的条件下，当儿童的学习精确度较小时，他们会等可能地选择普遍语法中的任何一门语言，而当学习精确度足够高时， 儿童选择母语的概率最高[97]。也有研究者采用其他定量方法成功地研究语言演化[105-108]。

语言演化网络的研究实际上包括两个方向。一个方向是在静态网络上研究网络的增长模式，如在对语言网络统计性质研究的基础上，有研究者建立了静态语言增长模型[57，67，109]，并推测出语言核心词典的规模[109]。另一个方向则研究随时间变化的语言网络的演化和动态，最近 Nowak 等在 Nature 上讨论了将进化动力学应用到图上的可行性和研究前景[110]。据我们了解，针对汉语的随时间变化的语言演化网络的研究，应该说几乎还没有。

1.4 语言网络在自然语言处理中的应用

近年来基于语言网络的自然语言处理受到越来越多研究者的关注。在语言学领域，语言学家 Hudson 认为语言是一个概念网络，并围绕这一核心思想在句法分析，语义理解和人类认知进行了深入的探讨[111]。有研究者将 PageRank[112]，HITS[113]等排序算法引入到语言网络，把文本中的词或句子作为网络节点，按照相对位置或者相似度将他们连接起来，然后在网络上使用 PageRank 或 HITS 算法对词或者句子进行排序，形成 LexRank[114]和 TextRank[115]等基于文本的排序算法，用来进行关键词语提取[115，116]，文档摘要[114，115，117-119]，词义消歧[120]和异常检测[121]。此外还有基于图的其他方法的应用[122-125]。

复杂网络理论的引入为 NLP 提供了新的研究手段，目前已有不少工作将复杂网络统计性质应用到 NLP 任务中，比较典型的包括关键词语提取[126-130]，文本聚类[131]，文档摘要[132-134]，情感分析[135]，作文水平自动评价[133，136，137]，文风识别[138]，等等。而针对汉语的已有工作主要针对关键词语提取[126，128-130]和文本聚类[131]。其中关键词语提取是目前复杂网络理论应用最多的，较为典型的 NLP 任务，主要做法有二:(1)直接利用网络节点(词或短语)的统计性质对词语排序得到关键词语；(2)将任务转化为是否为关键词语的二分类问题，将网络节点统计性质作为词语的特征向量构造训练集，训练分类器。

对语言网络的分析也有助于文本内容的理解和元信息的自动生成。前文提到的标签系统可以让人快速了解大量文本的内容，主题和类别分布，是近年新兴的一种语义标注方法[74，76，78，139-141]。标签自动生成开始受到研究者的关注。标签生成和关键词提取有密切联系，但也有所不同:关键词提取是从文档所使用的词汇集合中选取，而标签文本的来源则不限于原文档。现有的标签生成方法包括利用已标注文本为每个标签训练一个二分类器，并对每篇新文档都判断是否使用某个标签[142]，以相似文档的标签作为目标文档的标签[143]，以内容关键词作为标签[77]和利用用户计算机上的已有信息关键词和目标文档的关键词来共同生成个性化的标签[144]。这些方法要么受制于已标注的文本[142，143]，要么无法得到除文本关键词以外的其他分类和总结性信息[77，144]，还无法在真正意义上得到待标注文档的标签。基于复杂网络结构分析的方法能对文档进行结构性

的分析，得到重要的词语和由重要词语构成的社区，从而对标签生成提供帮助。同时，标签系统具有很强的时效性。我们认为，标签自动生成是综合运用并检验复杂网络方法在 NLP 中有效性的非常合适的应用任务。

由上可见，针对汉语复杂网络的研究，无论是在理论研究方面，还是在关键应用技术研究方面，都非常薄弱，总体上处于起步阶段，给我们留下了极具前瞻性的探索空间。

1.5 汉语语言网络研究的若干建议

本文概述了目前复杂网络理论在自然语言的网络性质，网络结构和网络演化上的研究及其应用。总的来讲，复杂网络理论在自然语言上的研究刚刚起步，尤其是在汉语上的研究与应用，仍然存在诸多前沿问题亟待系统地研究和解决。主要包括以下几个方面:

1. 全面考察各种类型的汉语语言网络的基本统计性质并进行比较分析。汉语语言网络主要包括: (1)汉语词同现网络； (2)汉语句法依存网络; (3)汉语语义构词网络； (4)汉语语义格关系网络； (5)汉语标签网络。我们需要从宏观角度对汉语语言网络基本面貌的一个完整刻画，在词法、句法、语义各个层次上大大丰富对汉语的科学认识，另一方面也为进一步的研究和分析打下必要基础。

2. 主要基于模体和社区特征,发现上述各种汉语语言网络在不同粒度下的基本结构， 并对其功能进行语言学或者语用学方面的必要解释。这将形成从微观角度对汉语语言网络基本面貌的完整刻画，在词法、句法、语义各个层次上进一步深化对汉语的科学认识，很可能会揭示出新的语言现象和规律；同时，通过对不同语言网络中的模体和社区特征之间的对比分析，或会有助于揭示汉语不同层次之间的某些本质联系。

3. 通过考察带有时间戳的演化语料库(如中小学语文课本)形成的汉语语义格关系网络， 可以研究汉语语义网络的演化过程并从多个角度，如核心语义与外围语义在空间，时间上的动态稳定性，变化率等，对其进行定量描述与分析。我们可以从这个近似于儿童语言习得的过程中， 揭示隐含于其中的语言学和认知语言学方面的规律。

4. 综合运用以上的研究成果，在大规模汉语语料库的基础上，提出并

实现基于复杂网络的自然语言处理的应用，如关键词语提取，文本聚类，文档摘要，情感分析，作文水平自动评价，文风识别，中文博客标签自动生成，等等。

总之，复杂网络理论是研究人类语言的一个崭新角度和工具，对于研究人类语言本质和进化，具有极大的启发意义。这也为如何解决人工智能和自然语言理解的任务带来新的契机和可能。

1.3 本文内容及主要贡献

本文在引言部分首先介绍了复杂网络理论及其在语言网络上的研究和应用，尤其详细介绍了在汉语网络上的研究与应用。

接下来本文将主要介绍作者在汉语网络上的探索成果，主要包括词同现网络、依存句法网络以及汉语博客标签网络上的研究成果，以及基于语言网络的典型应用：(1) 在北京大学《人民日报》1998 年上半年 1300 万字左右的人工分词语料库和国家语委 5000 万字左右的人工分词平衡语料库上建立了汉语词同现网络，考察了网络的小世界现象和无标度性质，并得到汉语上的核心词典规模；(2) 基于清华大学 100 万词句法标注树库，建立了汉语依存句法网络，考察了其复杂网络性质；(3) 基于抓取的博客标签建立了博客汉语标签同现网络，从复杂网络的角度考察了其统计性质；(4)根据二部图的资源分配(resource allocation)思想，在搜狐大规模搜索引擎日志的基础上，进行查询词推荐，实验表明该方法取得了与商用搜索引擎相当的效果，但比传统的基于子串匹配的推荐方法拥有更大的推荐选择空间。

第2章　汉语词同现网络

2.1 汉语词同现网络的构造及相关概念

汉语词法网络的构造主要基于一个静态的基本词语集。而词同现网络的构造则应基于动态的大规模语料库。对汉语而言，这个语料库显然需经过分词处理。

词同现网络的构造算法十分简单：语料库所包含的每一个词型（word type），对应着词同现网络中的一个节点（每一个节点在人脑中可映射为独立的认知实体，这样去考察节点之间的同现关系，才更有意义）。如果在一个句子中，两个词之间在 n 阶 Markov 链的条件下存在同现关系，则认为网络中相应的两个节点之间存在一个连接。对语料库中的所有句子进行上述处理，便可构造出词同现网络。

语言工程的实践表明，n 阶 Markov 链中的 n 取 2 比较合适，因为句子中两个词的邻接同现是最常见的，如“香港回归”的“香港”和“回归”、“清华大学”的“清华”和“大学”。同时存在大量的间隔 1 个词的同现，如“在书桌上”的“在”和“上”，“我的家”的“我”和“家”等）。虽然也存在一些间隔大于 1 的相关词对，但如果在模型中考虑此种远距离关联，则会引入大量的无关词对，降低词同现网络对真实情况反映的准确性。采取这个策略，一方面可较充分地反映词与词之间的上下文制约关系，另一方面，又可使模型的复杂性得到较好的控制。图 2.1 给出了一个根据上述算法由两句话生成的汉语词同现网络的简单示例。

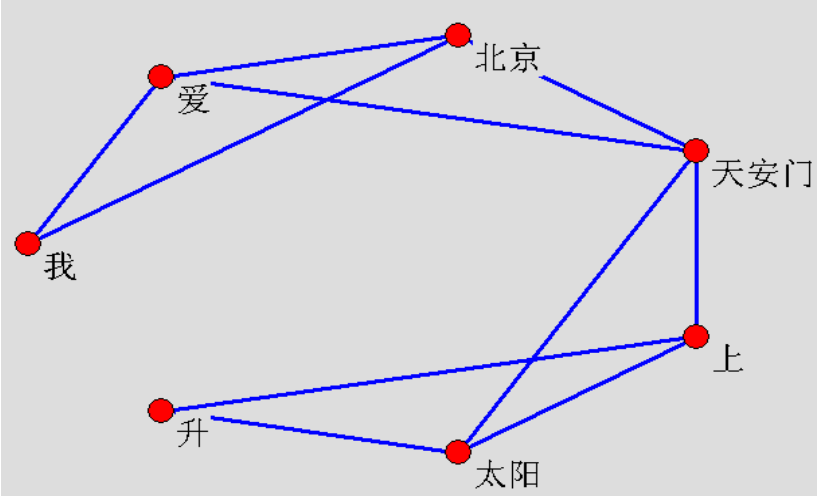

图 2.1 一个由“我爱北京天安门”和“天安门上太阳升”两句话生成的词同现网络。

一个词同现网络可以抽象为由词集 V 和边集 E 组成的无向图 $G=(V,E)$，其中词数 $N=|V|$，边数 $M=|E|$。网络中两个词 i 和 j 的距离 d_{ij} 定义为连接这两个词的最短路径长度。词 i 的度 $<k_i>$ 定义为与该词连接的其他词的数目。网络的平均最短路径 d 定义为任意两个词之间距离的平均值：

$$d=\frac{1}{N(N-1)}\sum_{i,j\in N,i\neq j}d_{ij} \quad (1)$$

如果词 i 与 j 之间存在连接，则设 $\xi_{ij}=1$。假设词 i 有 k_i 条边与其他词相连，那么这 k_i 个词定义为 i 的最近邻，其集合为 $\Gamma_i=\{j\,|\,\xi_{ij}=1\}$。词 i 的最近邻集合中词间的连接数为：

$$L_i=\sum_{j=1}^{N}\xi_{ij}\left[\sum_{k\in\Gamma_i,j<k}\xi_{jk}\right] \quad (2)$$

而在这 k_i 个词之间至多有 $k_i(k_i-1)/2$ 条边。所以词 i 的聚合系数 c_i 定义为：

$$c_i=\frac{L_i}{k_i(k_i-1)/2} \quad (3)$$

则该网络的聚合系数 C 为：

$$C = \frac{1}{N}\sum_{i=1}^{N} c_i \quad (4)$$

2.2 实验及其分析

本文实验利用了北京大学《人民日报（1998 年上半年）》1300 万字左右的人工分词语料库①和国家语委 5000 万字左右的人工分词语料库②（它们也是目前世界上规模最大、质量最高的汉语分词语料库）。前者可以按月份分割，用于在不同规模的汉语词同现网络上考察复杂网络性质及其平稳性；后者包含了各种题材、各个领域的文本，是较好的平衡语料库，可以更全面地考察汉语词同现网络的复杂网络性质。

本文设计了 4 组实验，用来生成词同现网络的语料库分别取北京大学《人民日报(1998 年上半年)》分词语料库的 1~2 月份、1~4 月份、1~6 月份和国家语委分词语料库，记作 CPD12，CPD14，CPD16 和 CYW。实验采用复杂网络分析软件 Pajek③进行数据分析。

实验结果见表 2.1。其中 ENG 是英语词同现网络上的实验数据，它采用了与本文相同的实验方法，在这里作为对照。ENG 实验从 750 万词的语料库得到含 460,902 个节点的词同现网络。而汉语 CPD16 实验从 730 万词语料库最终只得到 124,334 个节点的词同现网络。从相同规模的语料库得到的英语词同现网络节点数明显多于汉语的词同现网络，主要原因是英语是一种屈折语言，其名词、动词等有各种屈折变化，造成了网络中节点数目的激增。而汉语是一种孤立语言，缺少严格意义上的形态变化。

表 2.1 中 C_{random} 和 d_{random} 分别是相同参数下(N 和 $\langle k \rangle$ 相同)的随机网络中的聚合系数与平均最短路径。可以看到 $C \gg C_{random}$，而 $d \approx d_{random}$ 。汉语词同现网络与英语词同现网络一样，平均最短路径远小于网络规模而聚合系数非常高，具有明显的小世界效应。这说明，虽然大量的词(几万甚至以十万计)储存在人脑中，但是人们在这个词网络中，可以只用很短的路径

① http://icl.pku.edu.cn/
② http://219.238.40.213:8080/
③ http://vlado.fmf.uni-lj.si/pub/networks/pajek/

就能从一个词到达另一个词。也就是说在交流中，当使用了某个词，可以通过有限步很快地跳转到另外一个词。这样，语言网络一方面很好地保证人们在交流时的速度，另一方面能够从规模上保证人们在交流时用词的丰富性。

从 CPD12，CPD14 和 CPD16 可以明显看出平均度$\langle k\rangle$随着语料库规模增大。这可以被看作某种语言进化的过程：随着时间的推移和社会的发展，新产生的词被加入到语言中或者原来较少使用的词逐渐被关注，从而为人们所习用，例如从 CPD12 到 CPD14 新增加的词有“楼兰”、“报关员”、“彩色棉”等，从 CPD14 到 CPD16 新增的词有“帕格尼尼”、“公务车”、“核竞赛”等。伴随着这个过程，原有词的连接也会增加，如“市场”的度在各语料库中的变化为 1803(CPD12) -> 2842(CPD14) -> 3607(CPD16)，从一个侧面反映了 Barabasi 和 Albert 所阐述的复杂网络的增长性。

表 2.1 词同现网络的基本数据。其中 ENG 是英语词同现网络的实验数据，引自[65]。

实验	N	E	$\langle k\rangle$	C	C_{random}	d	d_{random}
ENG	4.61×10^{5}	1.61×10^{7}	70.13	0.437	1.55×10^{-4}	2.67	3.06
CPD12	0.71×10^{5}	0.10×10^{7}	28.44	0.535	3.99×10^{-4}	2.75	3.34
CPD14	1.03×10^{5}	0.18×10^{7}	34.18	0.556	3.32×10^{-4}	2.73	3.27
CPD16	1.24×10^{5}	0.24×10^{7}	38.99	0.569	3.14×10^{-4}	2.71	3.20
CYW	1.57×10^{5}	0.83×10^{7}	64.35	0.619	2.50×10^{-4}	2.63	2.99

图 2.2 列出了最短路径的分布。语料库中有部分孤立词与其他词没有连接（当这些词组成“独词句”时），会造成不可达词对，因此图中所列的节点对比例的和小于 1。平均最短路径的分布比较有规律，$d=2$和$d=3$的节点对比例占了绝大多数，CPD12、CPD14 和 CPD16 中都超过 80%，CYW 中也占 78.5%之高。除不可达词对外，存在连接的词对的最短路径都比较小。路径中两个词的距离越短，说明它们之间的跳转越直接，也越容易，

在人们交流的过程中越比较经常地一起使用，如路径“缉拿-凶犯”、“主隧-全长-公里”及“凶手-缉拿-凶犯”（这里“凶手”通过“缉拿”与“凶犯”产生同义关联）；反之，联系越松散，如路径“联系簿-警民-关系-群众”中的“联系簿”与“群众”。

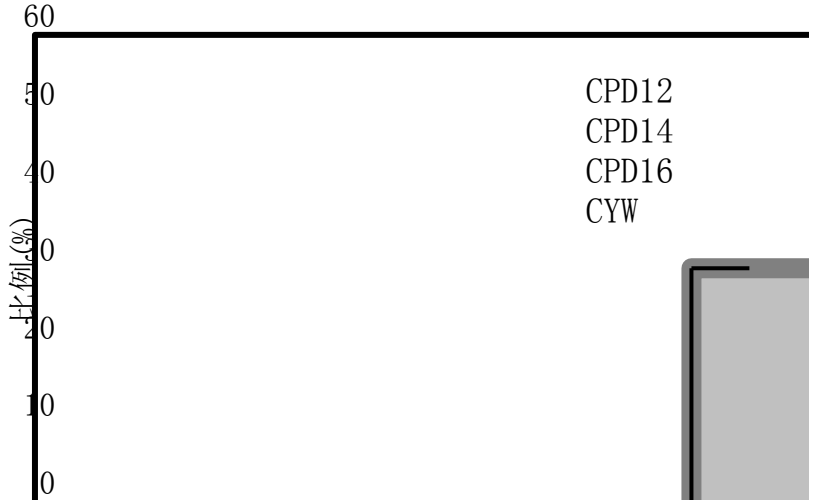

图 2.2 汉语词同现网络最短路径分布，即路径为 d 的数目及其比例。设词同现网络节点数为 N，路径长度为 d 的数目为 m_d，则比例为 $p_d = 2m_d / (N^2 - N)$。

网络节点的累积度分布曲线见图 2.3。累计度分布是度不小于 k 的节点的分布概率:

$$P(k) = \sum_{j=k}^{\infty} \Pr(j) \quad (5)$$

当度分布曲线呈幂律分布时，其累积度分布曲线也呈指数值相差 1 的幂律分布。根据式(3)可得：

$$P(k) = \sum_{j=k}^{\infty} j^{-\gamma} \propto k^{-(\gamma-1)} \quad (6)$$

可以看到四组实验结果都大体呈幂律分布，显示了无标度特性。

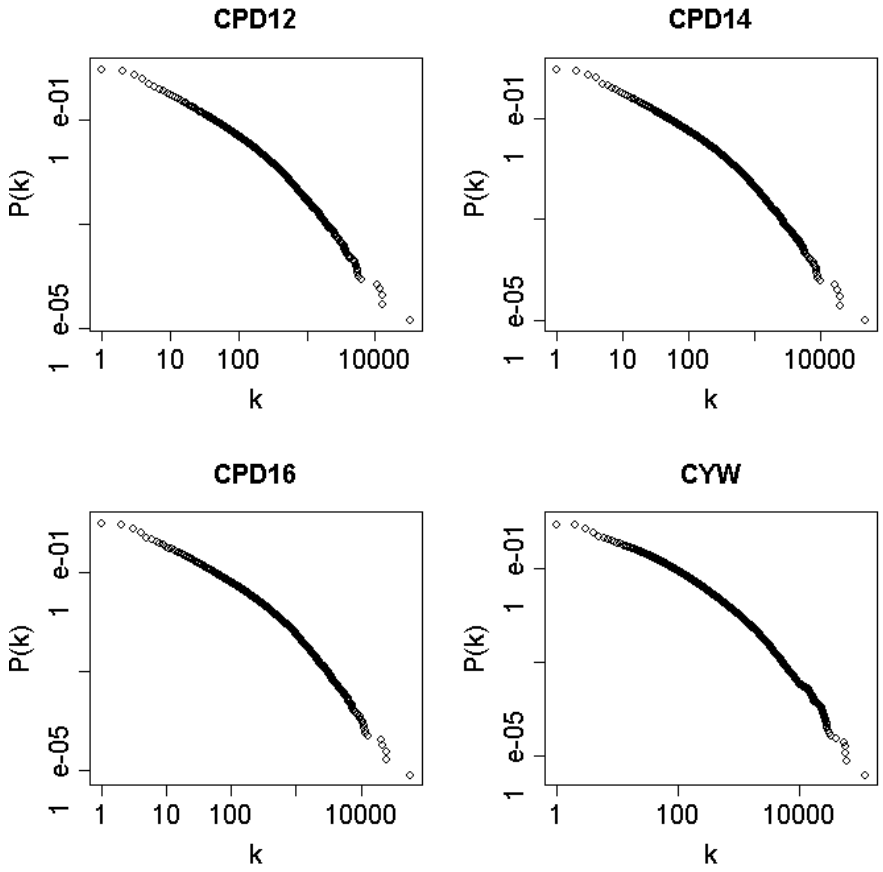

图 2.3 CPD12，CPD14，CPD16 和 CYW 的累积度分布曲线(log-log)。

如果对这些曲线进行更为细致的观察，则会发现其度分布并非一条直线，而是可以划分为两个斜率明显不同的线段（英语词同现网络也存在类似的现象）。图 2.4 显示了对 CYW 累积度分布进行线性拟合的情况：以 $(k, P(k)) = (802, 0.0133511)$ 处为转折点，第一段斜率为 –0.51，第二段斜率为 –2.51，并根据式(9)可得第一段指数 $\gamma_1 = 1.51$，第二段指数 $\gamma_2 = 3.51$；在 CYW 生成的词同现网络中，度大于 802 的词数目为 $P(k) \times N \approx 3434$。其他三组实验的度分布也明显分为两个不同斜率的线段，实验数据见表 2.2。

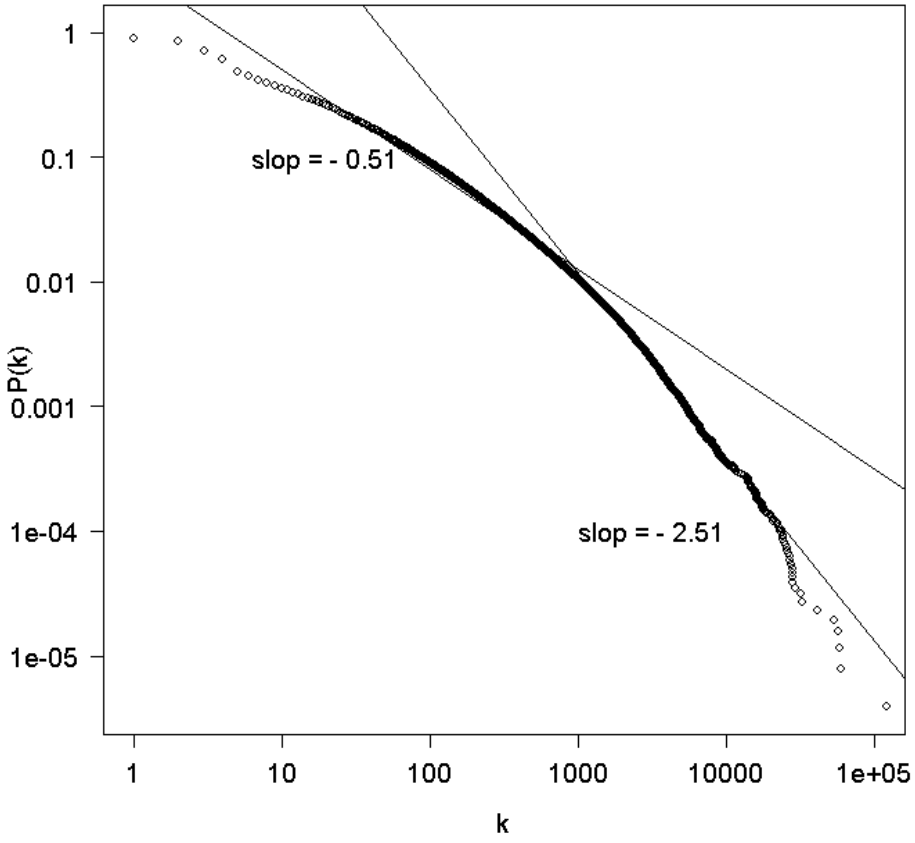

图 2.4 对 CYW 词同现网络累积度分布 (log10-log10) 的拟合曲线。

表 2.2 四组实验的累积度分布曲线通过两个不同线段拟合的结果。其中 k 为转折点的横坐标。$N \times P(k)$ 为度大于 k 的词数。

实验	k	$P(k)$	γ_1	γ_2	$N \times P(k)$
CPD12	177	0.028	0.266	1.921	951
CPD14	240	0.025	0.267	2.066	1553
CPD16	318	0.022	0.301	2.214	1873
CYW	802	0.013	0.511	2.511	3434

心理学实验表明，一个词在交流中出现频度越高，其语言产生的能力越强，即人脑能够更容易地使用这个词表达思想。图 2.5 显示 CYW 组实验中词频 f 与其度 k 之间存在相当强的相关性，即 $f \propto k^{\gamma}(\gamma > 0)$。这表明一个词的度越高，一般来说，其语言产生的能力也就越强。

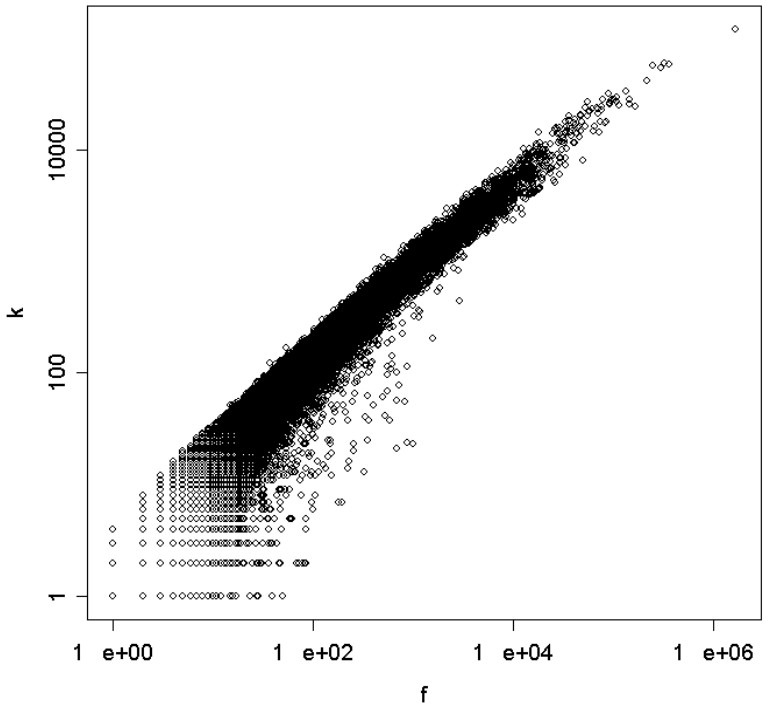

图 2.5 CYW 中词频 f 与其度 k 的相关性分析。

由表 2.2 还可见，$N \times P(k)$ 即核心词典的规模约为 10^3 量级，基本符合 DM 语言模型对核心词典规模的推论，英语的核心词典约含 5000 词。

表 2.3 给出了四组实验所得的核心词典(KLi，i 对应 CPD12，CPD14，CPD16 和 CYW)与从相应语料库产生的规模相同的词频表（FLi）、核心词典相互之间以及与人工建立的《普通话三千常用词表》(PTL)的比较结果。表中数字为相同词条的数目。不同语料库下的核心词典与相同规模的

词频表的比较，符合前文词频 f 越高则度 k 越趋高的结论。CPD12，CPD14和 CPD16 核心词典之间的比较表明绝大部分词条是相同的，说明核心词典具有一定的稳定性；由于 CPD12、CPD14、CPD16 与 CYW 语料库来源不同，它们之间的核心词典存在较大差异。各核心词典与《普通话三千常用词表》进行比较，大部分词条出现在该表中。而它们之间存在一定差别的主要原因是：(1)《普通话三千常用词表》是人工整理的词表，以人的主观感觉为主要判断依据，与词频的定量分析有一定出入。表 2.4 显示了四组实验中核心词典、《普通话三千常用词表》及两者的交集对语料库的词次(word token)覆盖率。此外，CPD16 核心词典对 CYW 语料库的词次覆盖率为 61。7306%，CYW 核心词典对 CPD16 语料库的词次覆盖率为 71。5804%，两核心词典交集对 CPD16 语料库的词次覆盖率为 66。2198%，对 CYW 语料库的词次覆盖率为 61。2254%，各核心词典对语料库的覆盖率明显高于《普通话常用三千词表》。在这一点上，核心词典显示了其定量分析的长处。(2)CYW 和 CPD16 的核心词典依赖于该语料库的来源、规模和分词标准等因素，因此只能是一定意义下的“汉语核心词典”。

表 2.3 四组实验所得的核心词典(KLi)与从相应语料库产生的规模相同的词频表（FLi）、核心词典相互之间以及与《普通话三千常用词表》的比较。表中数字为相同词条的数目。

比较对象	KL:CPD12	KL:CPD14	KL:CPD16	KL:CYW
	951	1553	1873	3434
FLi	851	1364	1658	3151
KL:CPD12	-	951	951	906
KL:CPD14	951	-	1151	1388
KL:CPD16	951	1151	-	1597
KL:CYW	906	1388	1597	-
PTL	610	892	1008	1719

表 2.4 四组实验中核心词典(KLi)、相同规模词频表(KLi)、《普通话三千常用词表》(PTL)及其交集对语料库的词次覆盖率。

覆盖率	CPD12(%)	CPD14(%)	CPD16(%)	CYW(%)
KLi	60.02	67.87	70.49	75.38
FLi	60.93	68.73	71.30	75.73
PTL	56.71	56.89	56.95	62.01
KLi∩PTL	48.06	51.73	52.73	59.46

在核心词典中，“的”、“和”、“在”、“了”、“是”、“为”、“有”、“这”、“他”、“我”和“人”等词的度最高。这些词或者是虚词，用以粘着成句，或者是具有强烈语法作用的实词。。它们中的相当一部分对于句子的理解似乎没有太大的直接贡献，而一旦缺失这些词，句子将变得支离破碎。这也反映了小世界网络的一个特性：如果随机地去掉网络中的节点，该网络仍然可以保持较好的连接性，而如果一旦去除的是高连接度的节点，整个网络将破裂成为若干孤立的网。某些失语症患者就表现为功能词缺失、不能正确组合语句、语句不完整、缺少长句和复杂句。

2.3 结论

本文基于大规模语料库，通过实验揭示了汉语在词同现网络上的小世界效应和无标度特性，并对汉语核心词典进行了初步研究。英语与汉语虽然存在显著差异（前者为印欧语系，后者为汉藏语系），但在词同现网络上表现出了类似的复杂网络性质。这一方面验证了 DM 语言模型对汉语的有效性，另一方面也从一个侧面印证了复杂网络的普适性。

第3章　汉语依存句法网络

3.1 依存句法网络

本文根据依存句法的定义构造句法网络，称为依存句法网络。依存句法是法国语言学家 Lucien Tesniere 提出的。他认为句子的主要动词是该句的中心，支配着其他成分，而它本身不受任何其他成分支配。后来，Robinson 提出了依存句法的四大公理：(1)一个句子中只有一个独立成分；(2)其他成分直接依存于某一成分；(3)任何一个成分都不能依存于两个或以上的成分；(4)如果 A 成分能直接依存于 B 成分，而 C 成分在句子中位于 A 或 B 之间的话，那么 C 或者直接依存于 A，或者直接依存于 B，或者直接依存于 A 和 B 之间的某一成分。

依存句法描述了句子中词与词之间的句法关系，这种句法关系绝大部分是有向的，我们假定这个方向由修饰语指向中心词。如“我爱天安门”这一个简单句中，“爱”是句子的中心词，而“我”和“天安门”是“爱”的修饰语并与之相连。如图 3.1(A)所示，链接由修饰语“我”和“天安门”指向中心词“爱”。

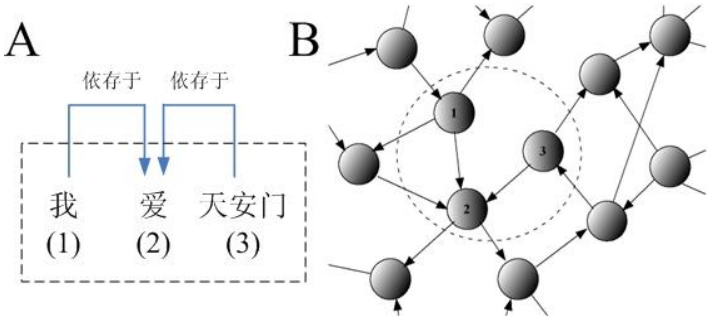

图 3.1 (A)句子“我爱天安门”的依存句法结构，其中动词“爱”是句子中心词，主语“我”和宾语“天安门”作为修饰语都依存于“爱”。(B)依存句法网络中，这三个词对应节点的链接情况。

由于依存句法关系的有向性，这种关系可以很自然地用有向网络表示。依存句法网络中的节点构成词集合 $V=\{s_i\}(i=1,...,n)$；网络中的链接集

合为 $E=\{a_{ij}\}(i,j=1,..n,$ 表示，即如果词 s_i 是词 s_j 的修饰语，那么 $a_{ij}=1$，否则 $a_{ij}=0$。一个句子的依存句法结构可以看作是整个句法关系的子集。如图 3.1 (B)所示，虚线内的部分就是“我爱天安门”这句话在句法网络中的子图。

在句法网络上，可以计算各种参数来考察其统计特性，如小世界效应，无标度性等。本文以清华大学周强的 100 万词句法标注树库作为语料库[145]，根据周明、黄昌宁在 1994 年提出的汉语依存句法规则[146]构造汉语依存句法网络，对其进行复杂网络特性方面的研究。

3.2 网络性质

在依存句法网络上，可以通过以下几个参数考察其复杂网络的统计特性。

1) 小世界效应。有两个关键变量刻画复杂网络的小世界效应。第一个变量是网络的平均路径长度，用 d 表示。定义 $d_{\min}(i,j)$ 是网络中词 s_i 和 s_j 之间的最短路径，那么词 s_i 与其他所有词的平均最短路径为：

$$d(i)=\frac{1}{n}\sum_{j=1}^{n}d_{\min}(i,j) \qquad (7)$$

这样，网络的平均路径长度 d 为：

$$d=\frac{1}{n}\sum_{i=1}^{n}d(i) \quad (8)$$

其中 n 为网络节点数目。另一个变量是网络的聚合系数，物理意义是网络中某词的任意两个邻居节点互为邻居节点的概率，用 C 表示：

$$C=\frac{1}{n}\sum_{i=1}^{n}C_i \qquad (9)$$

其中 C_i 表示词 s_i 的聚合系数，定义为词 s_i 的邻居节点之间的边数与可能的总边数之比[25]：

$$C_i = \frac{2}{k_i(k_i-1)}\sum_{j=1}^{n} a_{ij}\left(\sum_{l=j+1} a_{il}a_{jl}\right) \quad (10)$$

其中 k_i 是 s_i 的度。对于 Erdos-Renyi(ER)随机图网络，设网络的平均度为 $\bar{k}$，那么该网络的聚合系数为 $C_{random} \approx \bar{k}/n$，平均路径长度为 $d_{random} \approx \ln n / \ln \bar{k}$。当网络的平均路径长度 $d \approx d_{random}$ 的时候，我们称该网络具有小世界现象。而实际网络与 ER 随机图网络的主要区别在于而前者聚合系数 $C \gg C_{random}$。

2) 无标度性。网络的度分布 $P(k)$ 是刻画网络统计性质的另一个重要参数。ER 随机图网络的度分布近似为 Poisson 分布。而大部分复杂网络的连接度分布具有幂律形式，即满足：

$$P(k) \propto k^{-\gamma} \quad (11)$$

其中 $P(k)$ 表示度为 k 的节点出现在网络中的概率，这种性质也称为无标度性。网络的无标度性一般按照所有度、出度和入度分别考虑。

3) 层次性。研究表明，许多复杂网络同时存在模块性、局部聚类和无标度性，这些模块会按照等级组织起来。一般通过观察 k 与聚合系数 $C(k)$ 的分布来研究层次性，它表示网络中节点的度 k 与 $C(k)$ 的对应关系。某些网络如演员网、同义词网的 $k-C(k)$ 分布图明显呈幂律分布：

$$C(k) \propto k^{-\theta} \quad (12)$$

且 $\theta \approx 1$。这表明度很小的节点具有较高的聚合系数而且属于高度连接的模块；而度很高的节点具有较低的聚合系数，其作用只是把不同的模块连接起来。

4) 居间中心性。节点 v 的居间中心值 $g(v)$ 定义如下：设 $G_v(i,j)$ 表示节点 s_i 和 s_j 之间通过节点 v 的最短路径的条数，$G(i,j)$ 表示节点 s_i 和 s_j 之间所有的

最短路径条数，即有 $G(i,j)=\sum_v G_v(i,j)$。那么 $g_v(i,j)=G_v(i,j)/G(i,j)$ 可以表示节点 v 在节点 i 和 j 的联系中的重要性。而 $g(v)$ 则是所有节点对 i 和 j 的 $g_v(i,j)$ 之和，表示为：

$$g(v)=\sum_{i\neq j} g_v(i,j)=\sum_{i\neq j}\frac{G_v(i,j)}{G(i,j)} \quad (13)$$

许多实际复杂网络满足：

$$P(g)\propto g^{-\eta} \quad (14)$$

虽然大多数复杂网络的度分布都遵循幂律分布，但是 γ 值却不尽相同。而不同复杂网络的 η 却变化很小。

5) 匹配度。如果网络中高连接度的节点倾向于与高连接度的节点相连，就称该网络具有同配性，如果高连接度的节点倾向于与低连接度的节点相连，就称该网络具有异配性。网络匹配度可以通过计算匹配度系数 Γ 测量：

$$\Gamma=\frac{c\sum_i j_i k_i-\left[c\sum_i\frac{1}{2}(j_i+k_i)\right]^2}{c\sum_i\frac{1}{2}(j_i^2+k_i^2)-\left[c\sum_i\frac{1}{2}(j_i+k_i)\right]^2} \quad (15)$$

其中 j_i 和 k_i 是第 i 条边两端点的度数。设 m 是网络中边的条数，则 $c=1/m$。

若 $\Gamma>0$ 则网络是同配的，若 $\Gamma<0$ 则是异配的。研究表明在实际网络中，Internet、WWW、蛋白质交互网、神经网络和食物网满足 $\Gamma<0$，而各种社会关系网络满足 $\Gamma>0$。

3.3 实验及分析

本文以清华大学周强的 100 万词句法标注树库作为语料库，构建汉语依存句法网络的基本数据如表 3.1 所示。从语料库规模来讲，远大于捷克语(562820 词)、德语(21275 词)和罗马尼亚语(153007 词)。在此基础上构建的汉语依存句法网络更能反映句法网络的本质。

数据表明网络的平均路径长度 $d=3.8$，而聚合系数 $C=0.13$。该网络的 $d \approx d_{random}$ 而 $C \approx C_{random}$，说明汉语依存句法网络具有小世界效应。这种效应表明虽然网络的规模巨大，但是人们可以只用很短的路径从一个词到达另一个词。这样的句法网络既能够保证人类交流的速度，又能保证交流中用语的丰富性。图 3.2 显示句法网络路径长度分布情况，可以看到 $d=3$ 和 $d=4$ 的路径占了绝大多数。

表 3.1 汉语依存句法网络基本数据，作为对照列出捷克语、德语和罗马尼亚语数据[69]。

参数	捷克语	德语	罗马尼亚语	汉语
n	33,336	6,789	5,563	56,232
$\bar{k}$	13.4	4.6	5.1	16.7
C	0.1	0.02	0.09	0.13
C_{random}	4×10^{-4}	6×10^{-6}	9.2×10^{-4}	3.0×10^{-4}
d	3.5	3.8	3.4	3.8
d_{random}	4	5.7	5.2	3.9
γ	2.29	2.23	2.19	1.90/2.44
γ_{in}	1.99	2.37	2.20	1.99/2.27
γ_{out}	1.98	2.09	2.20	1.98/2.68
η	1.91	2.10	2.10	1.10/1.86
ζ	1.03	1.18	1.06	1.25
Γ	-0.6	-0.18	-0.2	-0.13

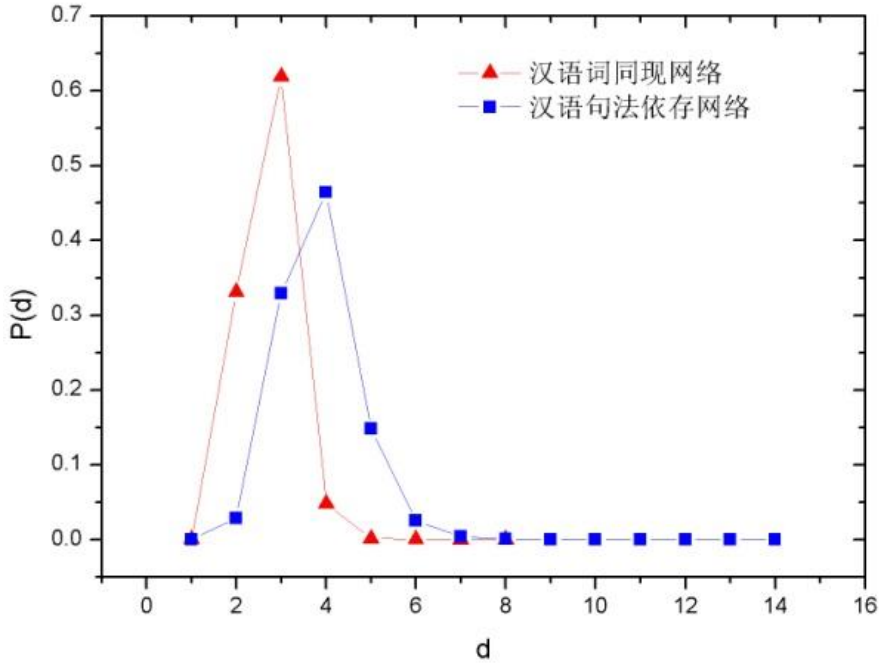

图 3.2 汉语依存句法网络的路径长度分布(正方形)。作为比较，同时画出汉语词同现网络的路径长度分布(三角形)。

图 3.3、图 3.4 和图 3.5 是汉语依存句法网络的所有度、入度和出度的累积分布曲线。度的累积度分布 $P_{\geq}(k)$ 是度不小于 k 的节点的分布概率，它与度分布的关系可以表示为：

$$P_{\geq}(k)=\sum_{j\geq k}P(j) \qquad (16)$$

当度分布曲线呈幂律分布时，其累积度分布曲线也呈指数值相差 1 的幂律分布，根据式(11)可得：

$$P_{\geq}(k)=\sum_{j\geq k}j^{-\gamma}\propto k^{-(\gamma-1)} \qquad (17)$$

可以看到三个累积度分布曲线都大体呈幂律分布，显示了汉语依存句法网络的无标度特性。如果对这些曲线进行更为细致的观察，则会发现其度分布并非一条直线，而是可以划分为两个斜率不同的线段，图中箭头所指即为转折点。两段曲线的斜率见表 3.1。相比词同现网络，句法网络的两段曲线斜率相差较小。在词同现网络或句法网络上的这种度分布斜率划

分为两段的现象，暗示了人类语言核心词典的存在。核心词典的词汇为该语言的使用者所共用，其规模不随语言的进化而显著变化，约为10^3量级。核心词典在词同现网络中表现为两个斜率不同的度分布线段。其中属于第二段的词汇度较高，构成了核心词典；而第一段则为特定领域所使用的词。

汉语依存句法网络的度-聚合系数[$k-C(k)$]分布图如图 3.6 所示。该分布呈明显下降趋势，可以看出网络中既存在很多度虽小却聚合系数很高的节点，也存在很多度非常高而聚合系数较低的节点。但汉语与捷克语、德语和罗马尼亚语类似，其依存句法网络的$k-C(k)$分布并不那么明显遵循$\theta \approx 1$的幂律分布。各语言句法网络的这一相似之处表明它们具有相似的层次结构。

汉语依存句法网络的居间中心性的累计分布曲线如图 3.7 所示。居间中心性的分布曲线类似于连接度，因此：

$$P_{\geq}(g)=\sum_{j\geq g} j^{-\eta} \propto g^{-(\eta-1)} \quad (18)$$

图示表明，汉语依存句法网络的g累计分布曲线比较明显地分为两段，两段的η值如表格 1 所示。这与其他三种语言的分布曲线有明显的不同，可能的原因是汉语网络中缺少更多的关键词(hub word)。而图 3.8 是度-居间中心性[$k-g(k)$]分布图，可以看到两者之间有比较强的正相关关系。

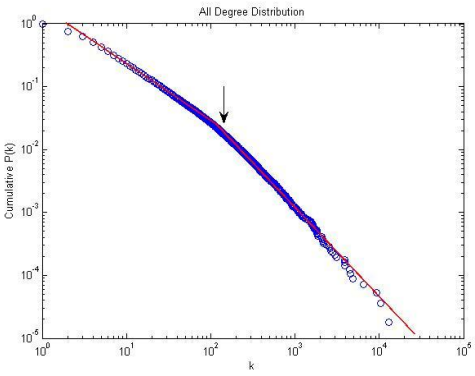

图 3.3 所有度的累积分布曲线。

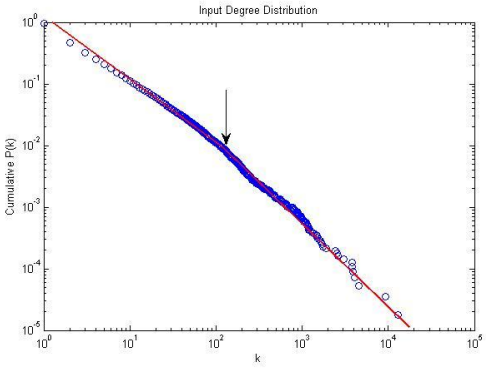

图 3.4 入度的累积分布曲线。

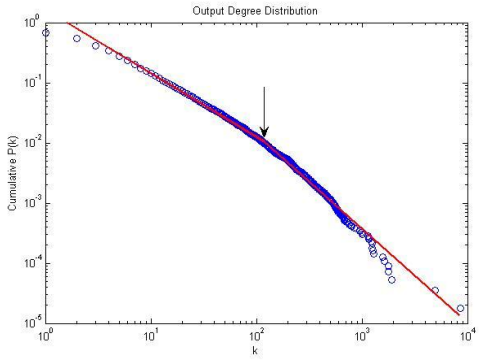

图 3.5 出度的累积分布曲线。

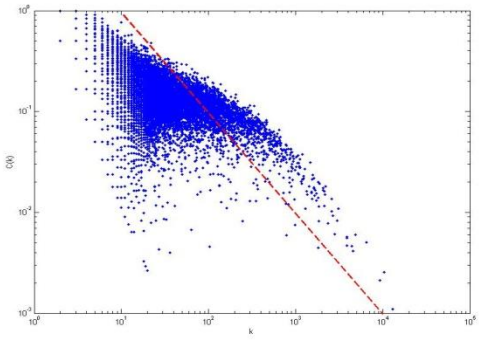

图 3.6 度-聚合系数分布图。其中虚线斜率为-1。

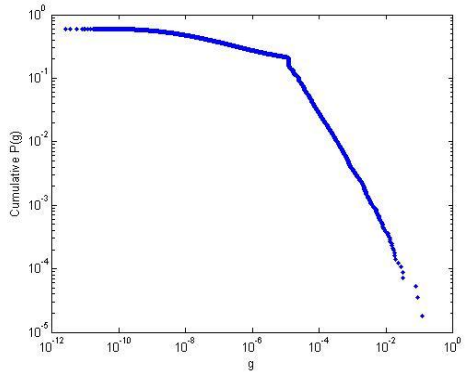

图 3.7 居间中心性的累积分布曲线。

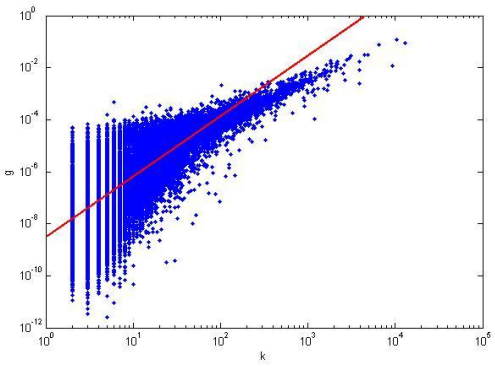

图 3.8 度-居间中心性值分布图。

图 3.9 是汉语依存句法网络中节点(词)的连接度及其在语料库中的词频的分布图，计算表明该分布呈明显的幂律分布：

$$f \propto k^{\zeta} \quad (19)$$

其中$\zeta \approx 1$(见**错误！未找到引用源。**)。汉语依存句法网络的ζ略高于其他三种语言，说明汉语网络更加稀疏，这是语料库的规模较大却并不足够大造成的。其中语料库规模变大使得网络节点增大，而不够大则使得网络中的链接较人类语言的实际句法网络而言不够充分。根据齐夫定律：

$$P(f) \propto f^{-\beta} \quad (20)$$

一般其中 $\beta \approx 2$。而 $\zeta \approx 1$，则 $P(k) \sim k^{-2}$，即 $\gamma \approx 2$，符合表 3.1 中数据。

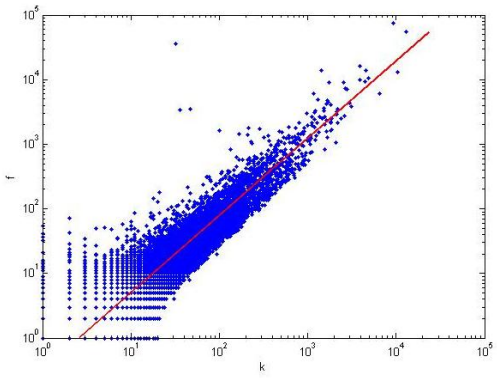

图 3.9 度-词频分布图。

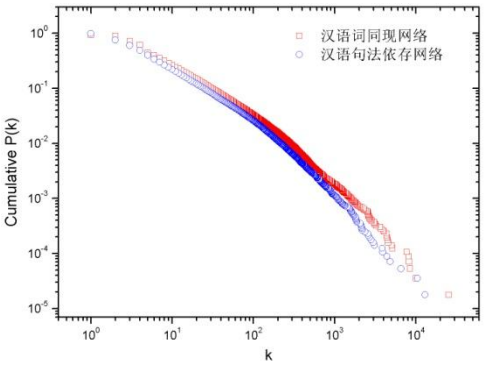

图 3.10 汉语词同现网络与句法网络度分布示意图。

度-词频分布图表明了度与词频的统计关系，即词频越高，其度也越高。一般而言，这些高连接度和高词频的节点都是功能词，如“的”，“了”，“在”，“和”等。计算可得汉依存句法网络与其他三种语言类似，其匹配度系数 $\Gamma < 0$，说明该网络是异配的，即网络中高连接度的节点倾向于与低连接度的节点相连。实际上，根据句法规则，功能词之间一般不构成句法依存关系，在句法网络中就表现为互相之间没有链接。而在 Roget 辞典网络中 $\Gamma = 0.157$，说明该网络是同配的。

3.4 句法网络与词同现网络的比较

根据相同的语料库，我们构造了汉语词同现网络。词同现网络的构造算法十分简单：语料库所包含的每一个词型（word type），对应着词同现网络中的一个节点。如果在一个句子中，两个词之间在 n 阶 Markov 链的条件下存在同现关系，则认为网络中相应的两个节点之间存在一个连接。对语料库中的所有句子进行上述处理，便可构造出词同现网络。一般而言 n 取 2。

首先在度分布上，通过图 3.10 可以看到汉语句法网络与词同现网络都大致呈幂律分布，并在尾部曲线斜率变大。虽然两者在边数上相差近

1/5，但在度分布上差别并不明显，只是在连接度较高的部分有比较大的偏离。表 3.2 可以看到一些主要参数的对比。可以看到，在平均度$\bar{k}$和聚合系数C上两者差别较大。由于平均度$\bar{k}$与网络边数有直接联系，两者差别较大并不难理解。对于聚合系数，句法网络相比词同现网络小很多，主要原因是，句法网络中，词是按照句法依存关系建立链接的，与某个词连接的其他词大部分在词性和句法功能上相似(例如都属于动词或名词等)，这些词较少倾向于互相连接，从而导致网络聚合系数较小。例如“文明”分别与“热爱”和“提倡”存在连接，但“热爱”和“提倡”之间存在句法依存关系的概率非常小。图 3.2 对两种路径长度分布进行了比较，两曲线形状大致相同。如图 3.11 是“文明”节点在同现网络和依存句法网络中邻居节点的间连接情况的比较，可以明显看出，同现网络的邻居节点间的连接紧密程度高于依存句法网络的邻居节点的紧密程度。

从本质上讲，词同现网络是句法网络的一种近似，可以揭示出人类语言的性质。通过比较，可以看到词同现网络与句法网络存在一定的差别，究竟这些差别是依存句法的什么特性造成的，是一个值得定量分析的问题。

表 3.2 汉语词同现网络与句法网络参数比较。

参数	汉语句法网络	汉语同现网络
n	56,326	56,326
E	447,519	589,654
$\bar{k}$	15.9	20.9
C	0.1	0.6
C_{random}	2.8×10^{-4}	3.7×10^{-4}
d	3.8	2.7
d_{random}	3.9	3.6
Γ	-0.12	-0.11

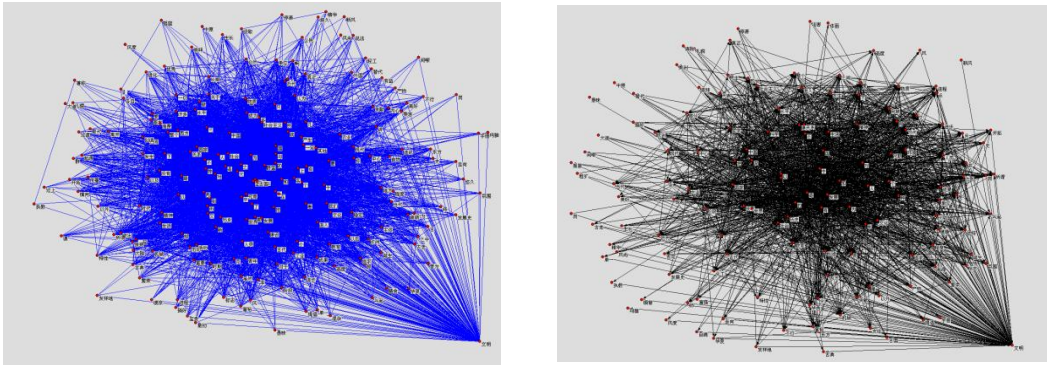

图 3.11 节点“文明”在同现网络(左)和依存网络(右)中邻居节点之间的连接情况比较。两幅图的右下角节点是“文明”。

3.5 结论

本文展示了汉语依存句法网络的小世界效应、无标度特性以及其他复杂网络性质，与捷克语、德语和罗马尼亚语等做了定量和定性的比较，此外还将汉语词同现网络与句法网络进行了初步比较。本文首次以句法网络的形式研究汉语的句法性质，并与其他语言进行了比较，有利于更全面的把握和分析人类语言的本质。在未来我们可以进行以下工作：(1)词同现网络实际上是句法网络的一种近似，那么如何更深层次挖掘这两者之间的异同，将对于理解人类语言具有重要意义；(2)目前我们还仅在无权的句法网络上进行了研究，而网络连接的权重是非常重要的信息，那么进一步研究带权句法网络将有助于更深入了解语言的性质；(3)目前对于语言网络的研究还停留在整体统计性质上的分析，进一步分析和研究语言网络的局部结构的性质将帮助我们更全面认识语言的本质。

第4章　汉语博客标签网络

4.1 介绍

博客是一种日志性质的网站，主要由按新旧顺序排列的带有日期的文章及对应的评论组成。不同的博客之间通过链接、评论和反向链接互相联系，带有明显的社区性质。博客反映了作者群体的观点和生活，并且对网络和现实的世界都有一定的影响，虽然说博客的作者群体存在偏向性，但是其中还是蕴含了重要的信息。经过对大量博客的索引和挖掘，可以得到例如博客群体对商品和公司的意见、广告投放的效果、博客群体对某些事件的关注程度和看法等等对商业和社会有价值的信息。这里值得特别注意的一个要点是 “大众分类法”(Folksonomy，意谓 Folk 的 Taxonomy)在博客中的应用。这种主要出现在博客、网络相册和网络书签系统上的、依靠大量用户使用自由选择的词汇作为标签(Tag)来对事物进行标记的人工分类方法，被称为“大众分类法”。这种分类方法显然和传统的基于少数专家的分类体系不同，每个分类者的自由度很大，通常也没有层次关系。现在比较流行的英文自由分类法有 del.icio.us 网络书签和 Flickr 网络相册中使用的体系，中文自由分类法的代表有豆瓣网(www.douban.com)等。

与大众分类法在网络上的大行其道相比，对其结构性质、动态演化过程和机理的研究仍然处在刚刚起步阶段。本文将对中文博客网站的标签体系的统计性质和演化模型进行研究，以期对中文博客的标签标注机理有更为深入的了解。

4.2 统计性质

4.2.1 数据集

本文使用的博客数据集是通过我们设计的一个聚焦抓取系统(focused crawling)来实现的，该抓取系统可收集并定期查看已经发现的博客所发布的 RSS 列表，下载其中出现的新文章，并且收集博客文章中附带的标签信

息。整个系统在无人值守的情况下连续运行，并且只关注中文博客。我们已利用此系统发现了 48,647 个博客站点，包括 362,889 篇文章，其中标签出现的总次数为 890,935，互不相同的标签 129,001 个。

4.2.2 统计信息

图 4.1 显示了数据集中标签按照所含字数的分布情况。可以看到，标签多数为 2 字词；从频度最高的十个标签全为 2 字词也可窥豹一斑，这与现代汉语以多字词为主的现象相符。在该数据集中，频度排名前 10 位的标签为：爱情，日记，生活，情感，女人，心情，男人，希望，妈妈，公司。可以发现，这些经常被用来作为标签的词，其语义具有一般意义，这种词汇具有更强的概括能力，因此会被更频繁地使用。图 4.2 显示了数据集中每次标注按照标签数目的分布情况。每次标注的标签数目，以 1 个最多，但标注 2~5 个的情况也较多，这使研究标签同现网络成为可能。表 4.1 显示了“爱情”，“日记”，“生活”，“情感”等四个指定标签的同现统计信息。

表 4.1 指定标签的同现统计信息

标签	使用该标签的文章数	与该标签同现的标签次数	与该标签同现的标签种类
爱情	9,066	11,481	3,220
日记	7,839	6,816	1,147
生活	7,769	12,726	2,494
情感	7,569	5,635	1,249

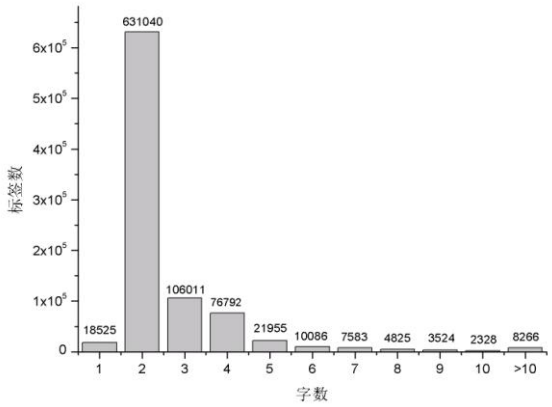

图 4.1 数据集中标签按照标签字数的分布图

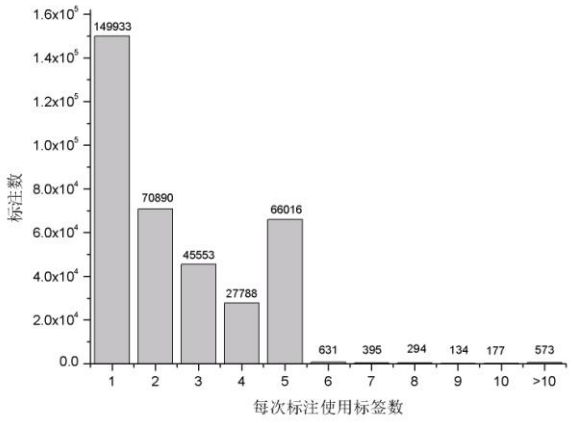

图 4.2 数据集中标注按照标注中的标签数目的分布图

对标签按照频度由高到低排序，前 N 个标签及其频度覆盖率如表 4.2 所示。如前所述，互不相同的标签数为 129，001 个，而标签频度高于等于 6 的 15，845 个标签已经覆盖了标签使用总频度的 80%以上。如图 4.3 所示，是频度高于等于 6 的 15,845 个标签按照所含字数的分布情况，可以发现，与图 4.1 相比，在覆盖率较高的这些标签中，虽然 2 字词仍然占主要地位，但是 3 字词和 4 字词的比例明显增大。

表 4.2 按频度由高到低排序的标签数目及其频度覆盖率对应关系

标签频度	标签数目(N)	覆盖率(%)
≥ 6	15,845	81.93
≥ 5	18,482	83.41
≥ 4	22,274	85.11
≥ 3	28,689	87.27
≥ 2	41,797	90.21

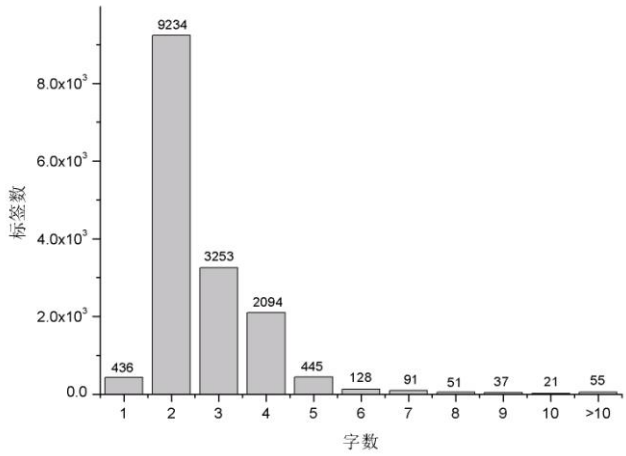

图 4.3 频度高于等于 6 的标签按照标签字数的分布图

4.2.3 齐夫定律

如图 4.4 所示，是所有标签的频度统计的排名分布图；图 4.5 则显示了指定标签的同现标签的频度统计排名分布图。可以明显看出，分布基本符合齐夫定律。然而，也可以发现在排名较高的部分相对平坦，这主要是有两个原因：(1)语义相近或重叠的常用词语会在使用上存在竞争关系，如“情感”和“心情”之间就存在这种关系。(2)标签在语义上存在潜在的层次结构，对于更为通用的标签如“爱情”、“生活”相对于“杜鹃”等会更多地与其他标签搭配出现。

每次标注中的任意两个标签之间存在同现关系。图 4.6 是对同现的两个标签(可以称为 BiTag)按照频度统计得到的排名分布图。可以看到也明

显基本符合齐夫定律。图中还标注了频度最高的几个同现标签。

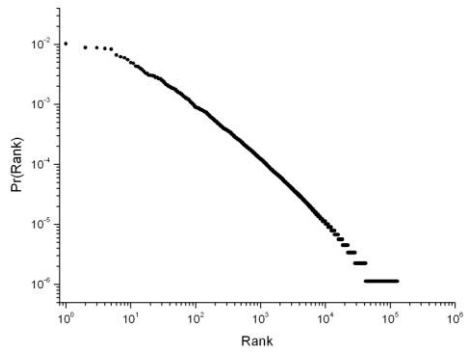

图 4.4 所有标签的频度排名分布图

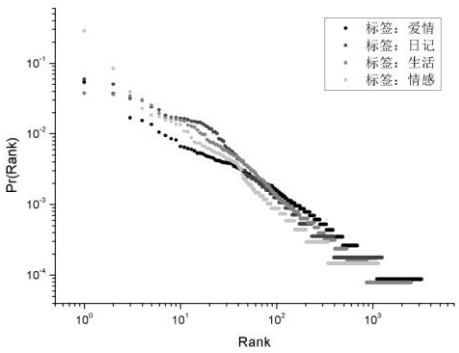

图 4.5 选定标签的同现标签频度排名分布图

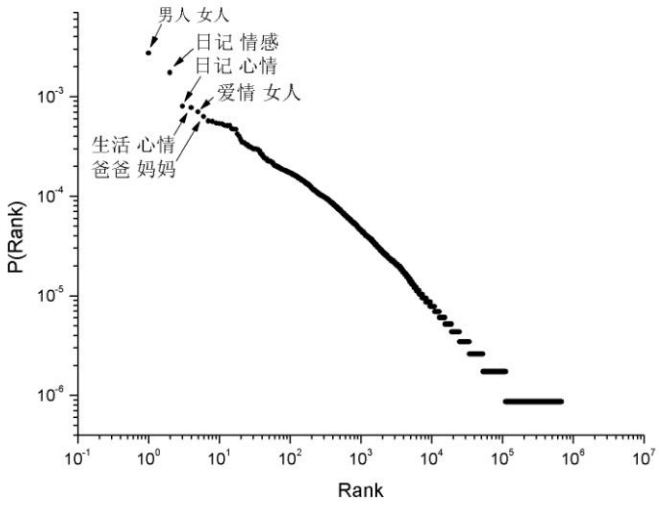

图 4.6 同现两个标签的频度排名分布图

4.2.4 复杂网络性质

最近复杂网络的研究取得了引人注目的成果。研究发现各领域中有着不同拓扑结构的复杂系统，如 Internet、食物网以及社会网络等，都表现出相似的统计规律。首先是复杂网络的小世界效应，即网络的聚集程度相对随机网络明显较高，同时平均最短路径比较小。此外，复杂网络表现出

无标度特性，即节点连接度呈幂律分布。

大众分类体系由于各种关系(如同现关系等)也构成规模巨大的复杂网络。然而目前对于大众分类体系复杂网络性质的研究还刚起步，其中对非协同的大众分类体系的研究，就作者了解，到目前还未相关工作发表。本文将在中文博客上的标签同现网络进行这方面的尝试。

设网络节点个数为 N，边数为 E。以下是复杂网络的三个重要参数。

平均最短路径长度。网络中两节点之间的平均距离。具有小世界性质的网络的平均最短路径会很短，远小于网络规模（这也是“小世界”命名的原因）。设平均最短路径为 d，网络节点平均度为 $\bar{k}$，对“小世界”网络，则有 $d \approx \ln(N)/\ln\bar{k}$。

聚合系数。一个节点的聚合系数反映了其相邻节点所构成集合的聚集程度。整个网络的聚合系数 C 是每个节点 i 的聚合系数 C_i 的平均值 $(0 \le C \le 1)$。对一个包含 N 个节点的 ER 随机图网络，当 N 很大时，有 $C \approx \bar{k}/N$，即其聚合系数远小于 1。而大规模的实际复杂网络表现出显著的聚合效应。

节点连接度分布。大量研究表明，实际复杂网络的度分布明显不同于 Poisson 分布，而更接近于幂律分布（无标度分布），即 $\Pr(k) \propto k^{-\gamma}$，其中 $\Pr(k)$ 是度为 k 的节点出现在网络中的概率，γ 为常数。

设标签同现网络节点数为 N，网络边的条数为 E，网络节点的平均连接度为 $\bar{k}$，平均最短路径长度为 d，同等规模随机网络的平均最短路径长度为 d_{randon}，聚合系数为 C，同等规模随机网络的平均最短路径长度为 C_{random}，这些统计参数值列在表 4.3 中。可以发现标签同现网络的 d 很小，而聚合系数 C 相对于 C_{randoi} 较大，表现出复杂网络的小世界性质。如图 4.8 所示，标签同现网络的最短路径长度集中分布在 3、4 数值上，也就是说，从一个标签到任意另外一个标签，平均只需要 3 到 4 跳，这对研究用户进行标签标注的行为模式具有一定的启发意义。

表 4.3 标签同现网络各参数值

参数	标签同现网络
N	129,001
E	680,367
$\bar{k}$	10.
d	3.5
d_{random}	5.0
C	0.4
C_{random}	8.1×10^{-5}

如图 4.7 所示，标签同现网络的累积度分布大致呈幂律分布，即 $\Pr(k)\propto k^{-\gamma}$，对曲线拟合得到 $\gamma\approx2.28$。表现出复杂网络的无标度性质。

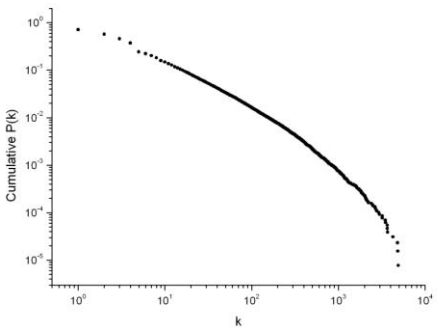

图 4.7 标签同现网的累积度分布曲线

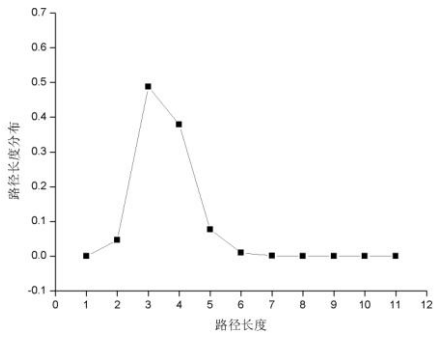

图 4.8 标签同现网的最短路径长度分布曲线

4.3 结论与展望

本文从齐夫定律、复杂网络等几个方面对中文博客标签数据集进行考查，初步揭示了非协同标注的“大众分类法”的一些有意义的性质。这些性质为接下来更深入的工作提供实证基础。在此基础上，我们可以开展以下工作：(1)分析用户标注的行为模式，为“大众分类法”的演化建模，用于预

测未来趋势；(2)通过标签在标签网络中表现的性质，对其进行语义分类；(3) 分析 “大众分类法”与“专家分类法”的异同，并尝试两者的融合，等等。

第5章 基于二部图的查询词推荐

5.1 介绍

随着 WWW 的发展，数以亿计的网页通过 Google、百度等搜索引擎进行检索。虽然基于关键词的网页搜索不是一个完美的检索方式，但是目前所有主流的搜索引擎都是通过计算关键词与网页之间的相似度决定返回结果的。因此，作为唯一的用户输入的方式，查询词的处理是影响搜索引擎效果的主要因素之一。

但是，搜索引擎返回的结果往往不是用户期望得到的。一个调查发现[147]，在 40,000 个被调查者中，有 76%的用户会在搜索失败后通过修改查询词重新搜索而不是换到另外一个搜索引擎。因此，对于搜索引擎而言，如何找到更好地表达用户搜索意图的查询词是一个非常重要的工作。

实际上，大部分用户的搜索行为，包括查询词输入和在搜索引擎返回结果中的 URL 点击选择，都是有意义的。用户输入的查询词可以被看做是用户给 URL 标记的标签[148]，因此用户查询日志包含了大量的用户协同标注的信息，可以帮助我们提高搜索引擎的查询效果。各种基于用户查询日志的任务，如查询词聚类、分类、推荐和扩展等纷纷被研究者提出，以期从各个角度提高搜索引擎的效果。而这些任务的基础就是如何度量查询词之间的语义关系。狭义上语义关系是指概念或者意义上的关系。查询词，作为用户给 URL 或者网页标记的标签，他们之间蕴含着丰富的语义关系，暗示了用户使用这些关键词进行信息搜索的分类[148]。之前的大部分工作都是通过子串匹配的方式定义查询词的相似度。其主要的缺点在于，两个查询词之间的相似度是对称的。但是，在许多情况下，他们之间的关系是非对称的。例如，对于“ipod”来讲，作为 apple 公司的产品，该查询词与“apple”的相似度非常大；但是对于“apple”来讲，用户可能不止是希望查询这家 IT 公司的产品，还有可能是了解“苹果”这种水果的相关信息，因此，对于“apple”而言，它与“ipod”的相似度就相对较小。因此，我们提出了一种基于二部图的查询词推荐算法，能够较好地量化这种非对称的查询词之间的相似度，并用于查询词推荐中。该算法最

早被提出用来进行电影个性化推荐[155]，取得了较好的效果。

5.2 前人工作

已经有很多工作进行查询词语义关系的度量，它们被广泛用于查询词聚类、分类、推荐和扩展中。各种方法包括：基于返回文档的相似度[157，158]，基于选择文档的相似度[159，152]或者返回文档的片段[160]。它们中的大部分取得了不错的效果，但是由于涉及到大规模的文档之间相似度的计算，因此并不适合海量查询日志的处理。

Beeferman 和 Berger[147]提出了一种“与内容无关”的方法进行查询词聚类和推荐。其主要思想是对“查询词-URL”二部图中的两部分别迭代地进行合并最相似的两个查询词或 URL。Wen 等[149]提出了一种综合考虑多种因素的查询词聚类的方法，包括查询词中的关键词、查询词的子串匹配、选择 URL 集合和选择文档的内容相似度等。此外还有研究者提出了利用查询日志的关联规则方法进行查询词推荐。查询词的关系还可以通过将查询词映射到预先定义的类别中得到，例如 informational/ navigational/ transactional 的类别[162，163]、基于地理位置的类别[164]和其他人工定义的类别[150，151]等。但是这种方法的局限性较大，不能够识别和量化查询词之间多样化的语义关系。Baeza-Yates[165]提出了基于不同信息资源的查询词间的多种语义关系，例如基于选择 URL 集合的重叠程度、选择的文档中的链接关系。研究发现基于选择 URL 集合的重叠程度的方法能够得到明显的语义关系。

以上的方法都只能度量查询词之间的对称的相似度关系。与本章工作最相近的研究是 Baeza-Yates[148]提出了基于选择 URL 集合覆盖关系提取非对称语义关系，但是这种方法得到的非对称语义关系限制较大，需要集合完全覆盖才能满足被抽取的条件。而本章提出的基于二部图的资源传递的方法则能够非常灵活的抽取查询词间的所有非对称语义关系。

5.3 基于二部图的资源传递算法

我们首先需要构建一个带权的查询词-URL 二部图。记查询词集合为 $Q=\{q_1,q_2,q_3,...,q_n\}$，URL 集合为 $U=\{u_1,u_2,u_3,...,u_m\}$，那么二部图可以被一

个 $n\times m$ 的矩阵 $A=\{a_{ij}\}$ 表示，其中 $a_{ij}>0$ 表示 u_j 在查询词 q_i 返回的结果中被点击，而 a_{ij} 表示被点击的次数。

5.3.1 方法描述

基于二部图的资源传递方法可以做如下描述：如果希望找到查询词 q_i 的相关查询词，首先在二部图中为 q_i 赋资源值 f_i。资源传递通过两步完成，首先由 q_i 按照链接权重将资源分配给 q_i 的所有邻居节点；然后再由这些带有资源值的 URL 节点将资源按照链接权重分配给它们的所有邻居节点。最后，所有的资源都分布在查询词集合的某个子集中，记为 R_i。从 q_i 到 $q_j\in R_i$ 的关系强度 r_{ij} 可以记作

$$r_{ij}=f_i\times s_{ij} \quad (1)$$

$$s_{ij}=\frac{1}{k(q_i)}\sum_{l=1}^{m}\frac{a_{il}a_{jl}}{k(u_l)} \quad (2)$$

其中 $k(q_i)=\sum_{j=1}^{n}a_{ij}$，$k(u_l)=\sum_{j=1}^{m}a_{jl}$，分别是查询词 q_i 和 URL u_l 的连接度。设关系强度矩阵表示为 $S=(s_{ij})_{n\times n}$，最初的资源分布为 $\mathbf{f}^{(0)}=(f_1,f_2,...,f_n)$，那么资源传递之后的资源分布为 $\mathbf{f}^{(1)}=\mathbf{f}^{(0)}\bullet S$，在 S 中，第 i 行表示来自 q_i 的资源经过资源传递之后在查询词集合中的分布。

5.3.2 计算复杂度

设二部图中的边数为 e，而查询词或者 URL 的最大连接度为 $k_{\max}$，那么基于二部图的资源传递算法复杂度为 $O(nk_{\max}^2)$，需要 $n+m$ 内存储存二部图。而传统的聚类算法[147]则需要至少 $O((n+m)k_{\max}^2+e(4k_{\max}))$ 的操作和

$n+m+n^2+m^2$的内存用来保存二部图信息与查询词对、URL 对的相似度信息。而基于选择 URL 集合的方法[148]则也需要$O(n^2k_{\max})$的操作。因此基于二部图的资源传递算法无论在空间复杂度还是时间复杂度上都非常高效。

5.3.3 递归算法及其收敛性

一个非常自然的推论就是资源传递的过程可以是递归的，即$\mathbf{f}^{(t+1)}=\mathbf{f}^{(t)}\bullet S=\mathbf{f}^{(0)}\bullet S^t$，其中$\mathbf{f}^{(t)}$代表第$t$次资源传递之后在查询词集合中的资源分布情况。递归将会收敛至一个稳定解，即$\mathbf{f}^*=\mathbf{f}^*\bullet S$，这一点在[166]中被提起但未作深入讨论，本文将在这方面给予深入分析。

如果我们我们把所有的查询词看做可以互相转换的状态，那么S就可以被看做状态转移的概率矩阵。那么递归的资源传递过程就成为一个马尔科夫过程[167]。根据马尔科夫过程的理论，如果马尔科夫链是不可规约且非周期的，那么将会有一个唯一的稳定分布$\mathbf{f}^*$，即

$$\lim_{t\to\infty}S^t=\mathbf{1}\bullet\mathbf{f}^* \quad (3)$$

其中$\mathbf{1}$是列向量，每个值都为 1。在真实数据中，一个查询词-URL 二部图一般由一个较大的连通图和若干小的小联通子图组成。在每个连通子图中对应一个不可规约的马尔科夫链。同时由于图中的链接是由千万个用户共同产生的，因此这个马尔科夫链也是非周期的。所以如公式 3 所示，无论初始的资源如何设置，最终经过递归得到的稳定的资源分布$\mathbf{f}^*$将是唯一的，只取决于图的拓扑结构。这一特点说明，随着资源分配的递归进行，查询词之间的关系将越来越多的受到所有查询词在全局中的流行度的影响。这个过程也是查询词个体的特异性和全局流行度之间的动态平衡的过程。

5.3.4 算法参数

基于二部图的资源分配算法有两个参数：(1)上节叙述的递归次数；和

(2)资源分配的策略。资源分配策略的最简单方法是根据点击频度进行，如公式(1)(2)所示，一个更为复杂的方法是：

$$s_{ij}=\frac{1}{k(q_i)}\sum_{l=1}^{m}\frac{a_{il}^{\alpha}a_{jl}^{\alpha}}{k(u_l)} \quad (4)$$

$$k(q_i)=\sum_{j=1}^{n}a_{ij}^{\alpha},k(u_l)=\sum_{j=1}^{m}a_{jl}^{\alpha} \quad (5)$$

其中 α 是一个可调的参数，控制点击频率对资源分配的影响。公式(1)(2)是 α=1时的特殊情况。

5.4 实验和评价

5.4.1 数据集合

试验中，我们采用搜狗实验室提供的 2007 年 3 月份第一周的用户查询日志作为数据集，其中包括 10,046,246 次独立查询，1,310,135 个不同的查询词，980,395 个不同的关键词和 4,055,171 个不同的 URL。其中大部分的查询词包括 1~3 个关键词，绝大部分关键词包含 2~6 个汉字。

去除其中只出现一次的查询，建立二部图，其中包含 834,107 个查询词和 886,702 个 URL。对于每个查询词，我们赋资源值 $f_i=100$，分别运行基于二部图的资源分配算法，得到推荐查询词。如表 5.1 所示，是该方法的查询词推荐示例。可以看到，对于“小说网”，“小说”以 28.96 的强度被推荐，而对于“小说”，“小说网”只以 1.44 的强度被推荐。

表 5.1 推荐查询词示例。

查询词	推荐查询词
小说网	**小说**，玄幻小说，小说阅读网，xiaoshuo，言情小说，小說，免费小说，中文小说网,言情
小说	玄幻小说，起点，小说阅读网，言情小说，潇湘书院，起点中文网，**小说网**，幻剑书盟，起点中文

5.4.2 评价

大部分商业搜索引擎仍然通过子串匹配的方式进行查询词推荐，如表 5.2 所示，是本方法、Google 和百度对“文学”推荐的查询词。可以看到，Google 和百度推荐的查询词都含有子串“文学”，而本方法则可以推荐得到“榕树下”这样没有共同子串的查询词。而“榕树下”是国内最大的原创文学网站，被作为“文学”的推荐查询词非常合适。

表 5.2 三种方法对于“文学”的推荐查询词。

方法	推荐查询词
本方法	**榕树下**,玄幻小说,世纪文学,成人小说,小说,原创文学,文学小说,读书,天地文学
Google	世纪文学,文学屋,天翼文学,吾爱文学网,起点文学,文学家,成人文学,晋江文学网,晋江文学
百度	世纪文学,吾爱文学网,星辰变世纪文学,79 文学网,天地文学,文学城,艳情文学,极品家丁世纪文学,文学殿堂,文学屋

本文还采用人工标注的方法进行评价。本实验随机寻找了 180 个推荐查询词，邀请 4 名用户对其相关程度进行打分，从 5 到 0 分分别表示不同层次的相关程度。图 5.1 显示了 4 个用户对百度和不同递归次数得到的 9 个推荐查询词的打分情况。图 5.2 显示了不同个数的推荐查询词下的平均分。可以看到，本方法有着与商用搜索引擎相当的用户体验。

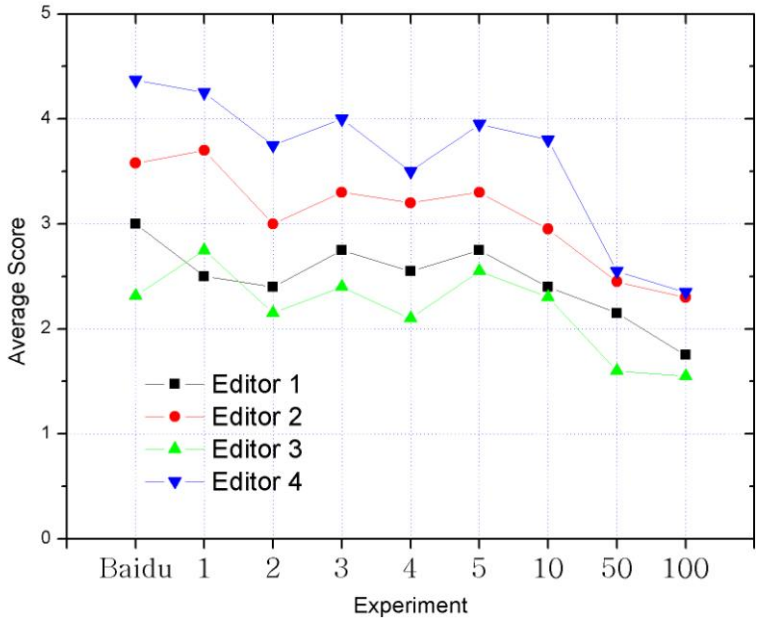

图 5.1 人工评价平均分

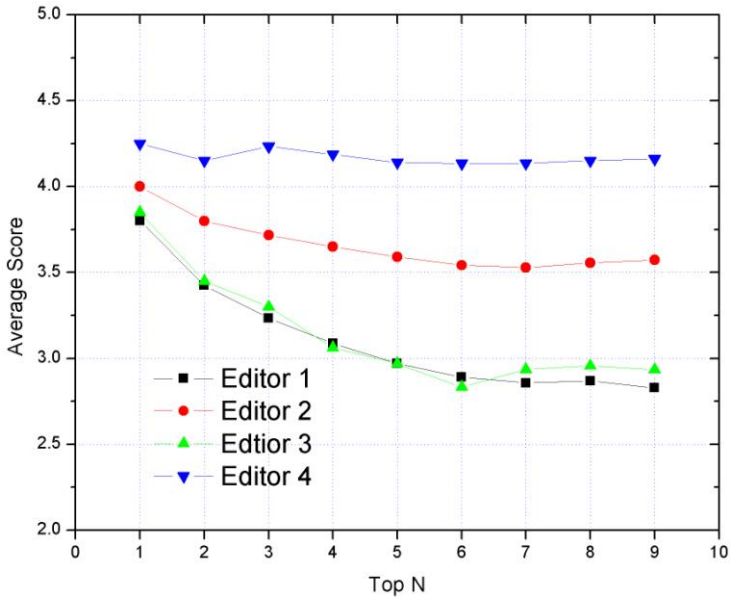

图 5.2 不同推荐个数的评价平均分

5.4.3 参数影响

下面考察不同参数对该方法的影响。图 5.3 显示了不同递归次数下的相关查询词集合的大小。我们也在表 5.3 中显示了不同递归次数下对“小说”的推荐查询词及其强度，可以看到排名最高的几个推荐查询词没有太大的变化。

表 5.3 不同递归次数下的推荐查询词及其强度。

递归次数	推荐查询词
1	玄幻小说 (23.8), 起点 (6.5), 小说阅读网 (5.2), 言情小说 (4.9), 潇湘书院 (1.9)
2	玄幻小说 (28.6), 起点 (6.7), 言情小说 (4.8), 小说阅读网 (3.3), 起点中文网 (2.2)
3	玄幻小说 (30.1), 起点 (6.4), 言情小说 (4.6), 小说阅读网 (2.4), 起点中文网 (2.2)
4	玄幻小说 (30.4), 起点 (5.9), 言情小说 (4.4), 起点中文网 (2.1), 小说阅读网 (2.0)

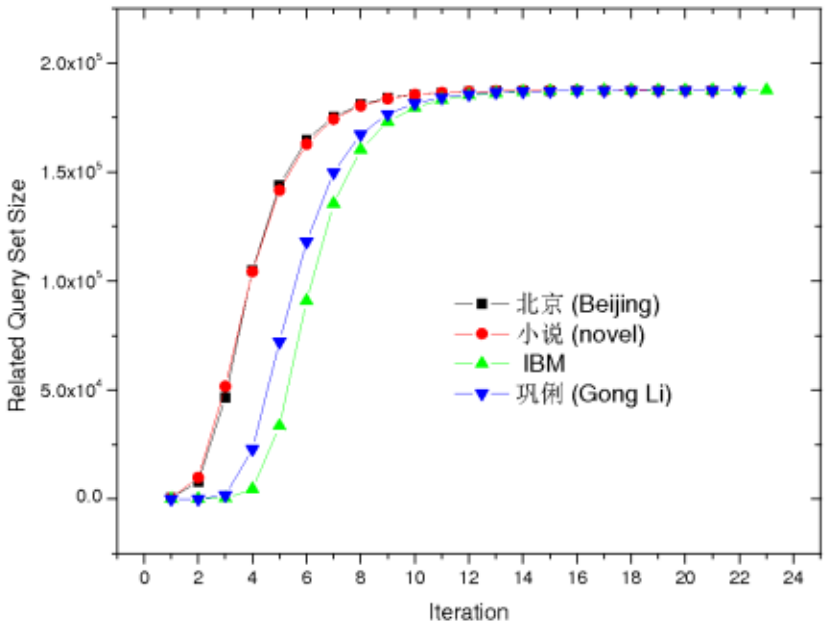

图 5.3 不同递归次数下的相关查询词集合大小

为了追踪递归过程中的变化，我们采用欧式距离来度量不同资源分布之间的差异。在图 5.4 中我们展示了 4 个不同查询词随着资源分配递归的差异的变化，均在差异小于 0.1 时停止迭代。在图 5.5 中我们也显示了不同的查询词之间随着递归次数的资源分布的差异。这些都表明了迭代收敛的趋势。

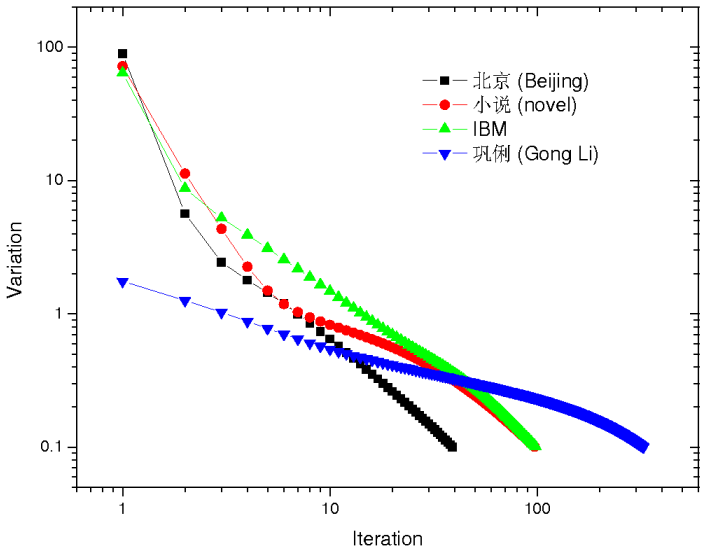

图 5.4 四个查询词在资源分配递归过程中的变化

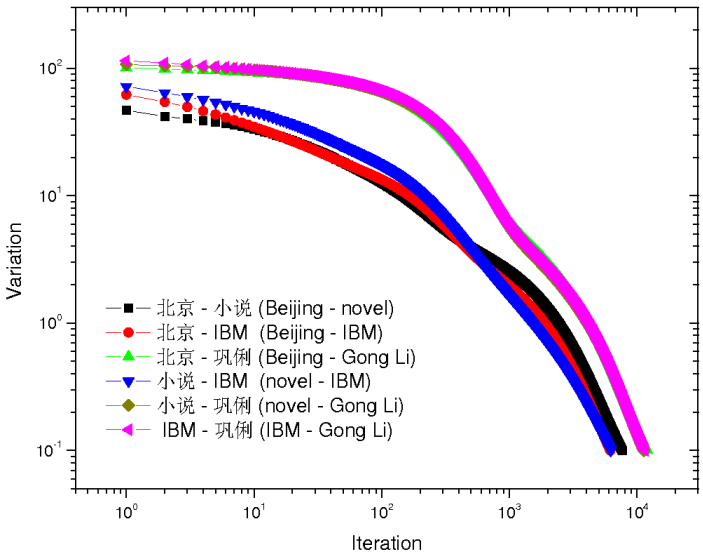

图 5.5 查询词之间的差异随着资源分配递归过程的变化

另外一个参数是资源分配策略。在图 5.6 我们显示了不同策略下，查

询词“北京”经过一次资源分配后的资源分布情况。当α=0时资源等分，当α=1时完全按照点击频度分配。

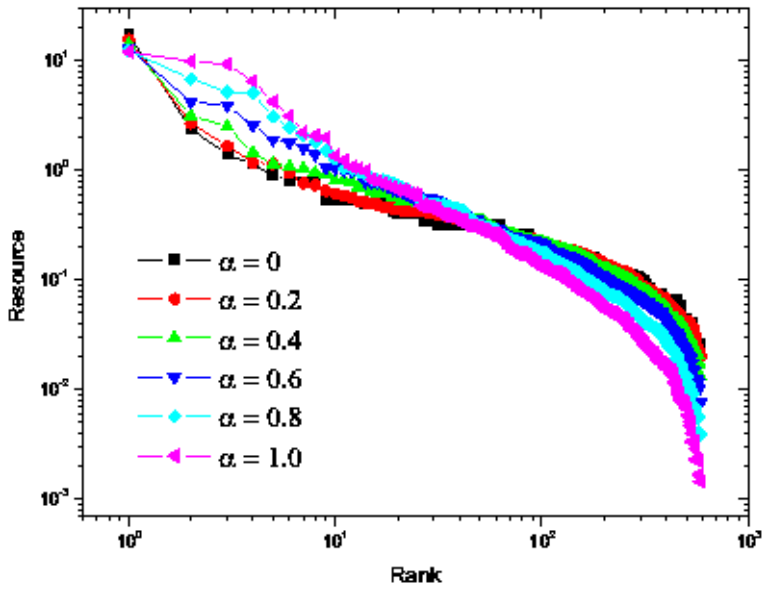

图 5.6 不同分配策略下的资源分布情况

5.5 查询词语义网络

实际上可以通过资源分配方法得到的查询词的语义关系构建查询词语义网络。我们保留所有语义关系强度超过 0.1 的边，构建语义网络。得到如表 5.4 的若干性质，图 5.7 显示了不同参数情况下的度分布情况。以上均说明该语义网络表现出比较明显的小世界性质和无标度特性。

查询词的语义网络包含了丰富的语义信息，表 5.5 显示了在网络中的语义关系路径，可以明显地看出其中的语义关系和主题漂移。

表 5.4 查询词语义网络的若干性质。

性质	值
节点数	834,107
边数	4,735,880
平均度	11.355
平均出/入度	5.678

平均最短路径(有向)	7.609
平均最短路径(无向)	7.231
聚类系数	0.527
联通子图数	556,900
γ	0.915/7.867

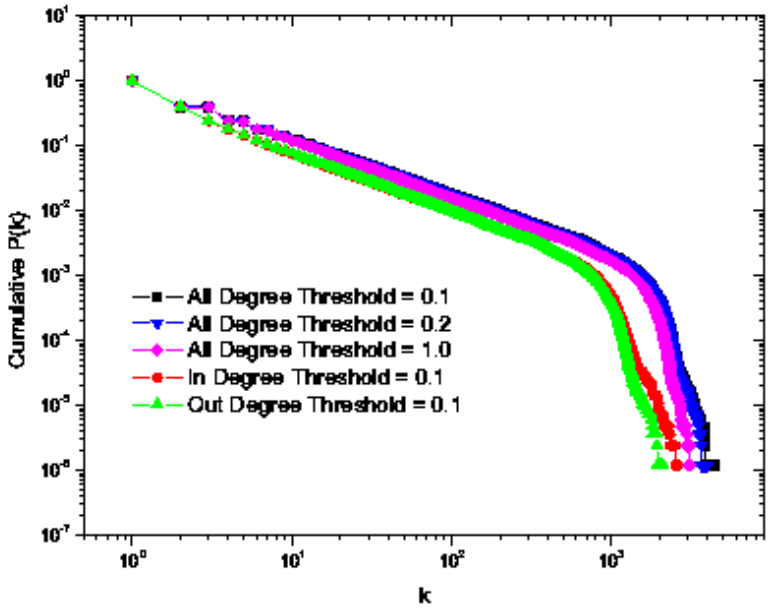

图 5.7 不同条件下的查询词网络的度分布情况

表 5.5 相关查询词路径示例。

相关查询词路径
巩俐→张艺谋→章子怡→艺妓回忆录
雅虎→Yahoo→www.yahoo.com.cn→雅虎中国
百事可乐→可乐→可口可乐→ www.icoke.cn →icoke

5.6 结论

本章提出了一种基于二部图的资源分配算法进行查询词推荐。这种方

法无论在时间复杂度还是在空间复杂度上都非常高效，而且不同于以往的基于字串匹配的方法，本方法可以量化非对称的查询词之间的关系，更加接近真实的查询词的语义关系。此外，基于该方法，本章还探讨了查询词语义网络的相关性质。

第6章 结论

本文主要对以下两个方面进行了研究：

首先是在若干汉语资源上考察了不同类型的汉语语言网络的包括小世界现象、无标度性质等复杂网络特性。它们包括：(1) 在北京大学《人民日报》1998 年上半年 1300 万字左右的人工分词语料库和国家语委 5000 万字左右的人工分词平衡语料库上建立了汉语词同现网络，考察了网络的小世界现象和无标度性质，并得到汉语上的核心词典规模；(2) 基于清华大学 100 万词句法标注树库,建立了汉语依存句法网络，考察了其复杂网络性质；(3) 基于抓取的大规模博客标签建立了博客汉语标签同现网络，从复杂网络的角度考察了其统计性质。总的来讲本文证实了虽然汉语与英语等分属于不同的语系，但是它的各种语言网络都表现出了与英语等一致的复杂网络性质，为进一步研究汉语语言网络奠定了基础。

其实是根据二部图的资源分配的思想，在搜狐大规模搜索引擎日志的基础上，进行查询词推荐的研究。实验表明该方法取得了与商用搜索引擎相当的效果，但比传统的基于子串匹配的推荐方法拥有更大的推荐选择空间。这也是语言网络的典型应用之一。

总之，语言网络无论是在实证还是在应用方面都具有巨大的研究价值。

参考文献

[1]Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks[J]. Nature. 1998, 393: 440-442.

[2] Barab A L, Albert R. Emergence of Scaling in Random Networks[J]. Science. 1999, 286(5439): 509-512.

[3] Newman M E. The structure and function of complex networks[J]. SIAM Review. 2003, 45(2): 167-256.

[4] Boccaletti S, Latora V, Moreno Y, et al. Complex networks: Structure and dynamics[J]. Physics Reports. 2006, 424(4-5): 175-308.

[5] Luciano, Rodrigues F, Travieso G, et al. Characterization of complex networks: A survey of measurements[J]. 2005.

[6] AlbertRéka, Barab A L. Statistical mechanics of complex networks[J]. Reviews of Modern Physics. 2002, 74(1): 47-97.

[7] Guo L, Xu XM. Complex Networks[M]. Shanghai: Shanghai Scientific and Technological Education Publishing House, 2006.

[8] Wang XF, Li X, Chen GR. Complex Networks: Theory and Applications[M]. Beijing: Tsinghua University Press, 2006.

[9] Erdos P, Renyi A. On the Evolution of Random Graphs[J]. Publications of the Mathematical Institute of the Hungarian Academy of Science. 1960, 5: 17-60.

[10] Milgram S. The small-world problem[J]. Psychology Today. 1967, 2: 60-67.

[11] Newman M E, Watts D J. Renormalization group analysis of the small-world network model[J]. Physics Letters A. 1999, 263(4-6): 341-346.

[12] Barabasi A L, Albert R, Jeong H. Mean-field theory for scale-free random networks[J]. Physica A: Statistical Mechanics and its Applications. 1999, 272(1-2): 173-187.

[13] Ravasz E, Barab A L. Hierarchical organization in complex networks[J]. Physical Review E. 2003, 67(2): 026112.

[14] Song C, Havlin S, Makse H A. Self-similarity of complex networks[J]. Nature. 2005, 433(7024): 392-395.

[15] Newman M E. Assortative Mixing in Networks[J]. Physical Review Letters. 2002, 89(20): 208701.

[16] Latora V, Marchiori M. Efficient Behavior of Small-World Networks[J]. Physical Review Letters. 2001, 87(19): 198701.

[17] Goh K I, Oh E, Kahng B, et al. Betweenness centrality correlation in social networks[J]. Physical Review E. 2003, 67(1): 017101.

[18] Kleinberg J M. Navigation in a small world[J]. Nature. 2000, 406(6798).

[19] Albert R, Jeong H, Barabasi A L. Error and attack tolerance of complex networks[J]. Nature. 2000, 406(6794): 378-382.

[20] Milo R, Shen-orr S, Itzkovitz S, et al. Network motifs: simple building blocks of complex networks.[J]. Science. 2002, 298(5594): 824-827.

[21] Masucci A P, Rodgers G J. Network properties of written human language[J]. Physical Review E. 2006, 74(2): 26102-26108.

[22] Kashtan N, Itzkovitz S, Milo R, et al. Topological generalizations of network motifs[J]. Physical Review E. 2004, 70(3).

[23] Girvan M, Newman M E. Community structure in social and biological networks[J]. Proceedings Of The National Academy Of Sciences Of The United States Of America. 2002, 99(12): 7821-7826.

[24] Milo R, Itzkovitz S, Kashtan N, et al. Superfamilies of evolved and designed networks[J]. Science. 2004, 303(5663): 1538-1542.

[25] Motter A E, De M A, Lai Y C, et al. Topology of the conceptual network of language[J]. Physical Review E. 2002, 65(6): 065102.

[26] Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks[J]. Proceedings of the National Academy of Sciences. 2004, 101(9): 2658-2663.

[27] Tsuchiura H, Ogata M, Tanaka Y, et al. Electronic states around a vortex core in high-T_{c} superconductors based on the t-J model[J]. Physical Review B. 2003, 68(1): 012509.

[28] Zhou H. Distance, dissimilarity index, and network community structure[J]. Physical Review E. 2003, 67(6): 061901.

[29] Fortunato S, Latora V, Marchiori M. Method to find community structures based on information centrality[J]. Physical Review E. 2004, 70(5): 056104.

[30] Newman M E. Modularity and community structure in networks[J]. Proceedings of the National Academy of Sciences of the United States of America. 2006, 103(23): 8577-8582.

[31] Brandes U, Delling D, Gaertler M, et al. On Modularity Clustering[J]. IEEE Transactions on Knowledge and Data Engineering. 2008, 20(2): 172-188.

[32] Palla G, Derenyi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature. 2005, 435(7043): 814-818.

[33] Flake G W, Lawrence S, Giles C L. Efficient identification of Web communities[C]. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston, Massachusetts, United States: ACM Press, 2000. 150-160.

[34] Rosvall M, Bergstrom C T. An information-theoretic framework for resolving community structure in complex networks[J]. Proceedings Of The National Academy Of Sciences Of The United States Of America. 2007.

[35] Kannan R, Vempala S, Vetta A. On clusterings: Good, bad and spectral[J]. Journal of the ACM. 2004, 51(3): 497-515.

[36] Spielman D A, Teng S H. Spectral partitioning works: Planar graphs and finite element meshes[J]. Linear Algebra And Its Applications. 2007, 421(2-3): 284-305.

[37] Alon N. Spectral techniques in graph algorithms[J]. LATIN '98: Theoretical Informatics. 1998, 1380: 206-215.

[38] Boccaletti S, Ivanchenko M, Latora V, et al. Detecting complex network modularity by dynamical clustering[J]. Physical Review E. 2007, 75(4): 045102.

[39] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E. 2007, 76(3): 36106-36111.

[40] Danon L, Duch J, Diaz-guilera A, et al. Comparing community structure identification[J]. 2005.

[41] Alves N A. Unveiling community structures in weighted networks[J]. Physical Review E. 2007, 76(3): 036101.

[42] Fan Y, Li M, Zhang P, et al. Accuracy and precision of methods for community identification in weighted networks[J]. Physica A: Statistical Mechanics and its Applications. 2007, 377(1): 363-372.

[43] Leicht E A, Newman M E. Community structure in directed networks[J]. 2007.

[44] Guimera R, Pardo M S, Nunes L A. Module identification in bipartite and directed networks[J]. Physical Review E. 2007, 76(3): 036102.

[45] Gibson D, Kleinberg J, Raghavan P. Inferring Web communities from link topology[C]. In: Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia. Pittsburgh, Pennsylvania, United States: ACM Press, 1998. 225-234.

[46] Barber M J. Modularity and community detection in bipartite networks[J]. Physical Review E. 2007, 76(6): 066102.

[47] Tantipathananandh C, Wolf T B, Kempe D. A framework for community identification in dynamic social networks[C]. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, California, USA: ACM, 2007. 717-726.

[48] Falkowski T, Spiliopoulou M. Data Mining for Community Dynamics[J]. Kunstliche Intelligenz. 2007, 3: 23-29.

[49] Palla G, Barabasi A L, Vicsek T. Quantifying social group evolution[J]. Nature. 2007, 446(7136): 664-667.

[50] Sole R V, Murtra B C, Valverde S, et al. Language Networks: their structure, function and evolution[Z]. 2005.

[51] Variano E A, Mccoy J H, Lipson H. Networks, Dynamics, and Modularity[J]. Physical Review Letters. 2004, 92(18): 188701.

[52] Leskovec J, Kleinberg J, Faloutsos C. Graph evolution: Densification and shrinking diameters[J]. ACM Transactions on Knowledge Discovery from Data. 2007, 1(1): 2.

[53] Berger-wolf T, Saia J. A framework for analysis of dynamic social networks[C]. In: the Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006. 523-528.

[54] Bales M E, Johnson S B. Graph theoretic modeling of large-scale semantic networks[J]. J. of Biomedical Informatics. 2006, 39(4): 451--464.

[55] Mehler A. Large Text Networks as an Object of Corpus Linguistic Studies[M]. Corpus Linguistics. An International Handbook., L A, Kyt M, Berlin/New York:de Gruyter, 2007.

[56] Luciano, Oliveira O, Travieso G, et al. Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications[J]. 2007.

[57] Steyvers M, Tenenbaum J B. The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth[J]. Cognitive Science: A Multidisciplinary Journal. 2005, 29(1): 41-78.

[58] De J H, Torres P I, Kinouchi O, et al. Thesaurus as a complex network[J]. Physica A: Statistical Mechanics and its Applications. 2004, 344(3-4): 530-536.

[59] Sigman M, Cecchi G A. Global organization of the Wordnet lexicon[J]. Proceedings of the National Academy of Sciences of the United States of America. 2002, 99(3): 1742-1747.

[60] Tang L, Zhang Y G, Fu X. Structures of Semantic Networks: Similarities between Semantic Networks and Brain Networks[C]. In: 5th IEEE International Conference on Cognitive Informatics. 2006. 356-361.

[61] Lu T, Lu T, Zhang Y G, et al. The statistic properties of Chinese semantic network in HowNet[C]. In: Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE '05). 2005. 58-61.

[62] Capocci A, Servedio V D, Colaiori F, et al. Preferential attachment in the growth of social networks: the case of Wikipedia[C]. In: 1st Workshop on Wikipedia Research. 2006.

[63] Zlatic V, Bozicevic M, Dataformat H S, et al. Wikipedias: Collaborative web-based encyclopedias as complex networks[J]. Physical Review E. 2006, 74(1): 016115.

[64] Ferreira A A, Corso G, Piuvezam G, et al. A scale-free network of evoked words[J]. Brazilian Journal of Physics. 2006, 36: 755-758.

[65] Cancho R F, Sole R V. The small world of human language[J]. Proceedings of the Royal Society of London Series B-Biological Sciences. 2001, 268(1482): 2261-2265.

[66] Kapustin V, Jamsen A. Vertex Degree Distribution for the Graph of Word Co-Occurrences in Russian[C]. In: Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing. Rochester, NY, USA: Association for Computational Linguistics, 2007. 89--92.

[67] Caldeira S M, Petit L T, Andrade R F, et al. The network of concepts in written texts[J]. The European Physical Journal B - Condensed Matter and Complex Systems. 2006, 49(4): 523-529.

[68] Cancho R F. The structure of syntactic dependency networks: insights from recent advances in network theory[M]. Problems of quantitative linguistics, V L, G A, 2005, 60-75.

[69] Cancho R F, Sole R V, Kohler R. Patterns in syntactic dependency networks[J]. Physical Review E. 2004, 69(5): 051915.

[70] Ferrer C R, Mehler A, Pustylnikov O, et al. Correlations in the Organization of Large-Scale Syntactic Dependency Networks[C]. In: Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing. Rochester, NY, USA: Association for Computational Linguistics, 2007. 65--72.

[71] Choudhury M, Mukherjee A, Basu A, et al. Analysis and synthesis of the distribution of consonants over languages: a complex network approach[C]. In:

Proceedings of the COLING/ACL on Main conference poster sessions. Sydney, Australia: Association for Computational Linguistics, 2006. 128-135.

[72] Medeiros S M, Corso G, Lucena L S. The network of syllables in Portuguese[J]. Physica A: Statistical Mechanics and its Applications. 2005, 355(2-4): 678-684.

[73] Griffiths T, Steyvers M, Firl A. Google and the Mind: Predicting Fluency With PageRank[J]. Psychological Science. 2007, 18(12): 1069-1076.

[74] Brooks C H, Montanez N. An Analysis of the Effectiveness of Tagging in Blogs[C]. In: the 2005 AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs. 2005.

[75] Hayes C, Avesani P, Veeramachaneni S. An Analysis of the Use of Tags in a Blog Recommender System[C]. In: the 2007 International Joint Conference on Artificial Intelligence. 2007.

[76] Marlow C, Naaman M, Boyd D, et al. HT06, tagging paper, taxonomy, Flickr, academic article, to read[C]. In: Proceedings of the seventeenth conference on Hypertext and hypermedia. Odense, Denmark: ACM Press, 2006. 31-40.

[77] Brooks C H, Montanez N. Improved annotation of the blogosphere via autotagging and hierarchical clustering[C]. In: Proceedings of the 15th international conference on World Wide Web. Edinburgh, Scotland: ACM Press, 2006. 625-632.

[78] Cattuto C, Loreto V, Pietronero L. Semiotic dynamics and collaborative tagging[J]. Proceedings of the National Academy of Sciences (PNAS). 2007, 104(5): 1461-1464.

[79] Halpin H, Robu V, Shepherd H. The complex dynamics of collaborative tagging[C]. In: Proceedings of the 16th international conference on World Wide Web. Banff, Alberta, Canada: ACM Press, 2007. 211-220.

[80] Golder S, Huberman B A. Usage patterns of collaborative tagging systems[J]. Journal Of Information Science. 2006, 32(2): 198-208.

[81] Cattuto C, Baldassarri A, P V D, et al. Vocabulary growth in collaborative tagging systems[J]. oai:arXiv.org:0704.3316. 2007.

[82] Zhang D, Lee W S. Web taxonomy integration using support vector machines[C]. In: Proceedings of the 13th international conference on World Wide Web. New York, NY, USA: ACM Press, 2004. 472-481.

[83] Lambiotte R, Ausloos M. Collaborative Tagging as a Tripartite Network[J]. Computational Science – ICCS 2006. 2006: 1114-1117.

[84] Schmitz C, Grahl M, Hotho A, et al. Network Properties of Folksonomies[C]. In:

Workshop "Tagging and Metadata for Social Information Organization" in the 16th International World Wide Web Conference (WWW2007). Banff, Alberta, Canada: 2007.

[85] Shen K, Wu L. Folksonomy as a Complex Network[J]. ArXiv Computer Science e-prints. 2005.

[86] Zhou S, Hu G, Zhang Z, et al. An empirical study of Chinese language networks[J]. Physica A: Statistical Mechanics and its Applications. 2008, In Press, Accepted Manuscript.

[87] Liu ZY, Sun MS. Chinese Word Co-occurrence Network: Its Small World Effect and Scale-free Property[J]. Journal of Chinese Information Processing. 2007, 21(6): 52-58.

[88] Liu H. The complexity of Chinese syntactic dependency networks[J]. Physica A: Statistical Mechanics and its Applications. 2008, In Press, Accepted Manuscript.

[89] Li Y, Wei L, Li W, et al. Small-world patterns in Chinese phrase networks[J]. Chinese Science Bulletin. 2005, 50(3): 286-288.

[90] Li Y, Wei L, Niu Y, et al. Structural organization and scale-free properties in Chinese Phrase Networks[J]. Chinese Science Bulletin. 2005, 50(13): 1304-1308.

[91] Li J, Zhou J. Chinese character structure analysis based on complex networks[J]. Physica A: Statistical Mechanics and its Applications. 2007, 380: 629-638.

[92] Hauser M D, Chomsky N, Fitch W T. The faculty of language: What is it, who has it, and how did it evolve?[J]. Science. 2002, 298(5598): 1569-1579.

[93] Wang W S, Minett J W. The invasion of language: emergence, change and death[J]. Trends In Ecology & Evolution. 2005, 20(5): 263-269.

[94] Mitchener W G, Nowak M A. Chaos and Language[J]. Proceedings of The Royal Society of London. Series B, Biological Sciences. 2004, 271(1540): 701-704.

[95] Mitchener W G, Nowak M A. Competitive Exclusion and Coexistence of Universal Grammars[J]. Bulletin of Mathematical Biology. 2003, 65(1): 67-93.

[96] Nowak M A, Komarova N L, Niyogi P. Computational and evolutionary aspects of language[J]. Nature. 2002, 417: 611-617.

[97] Nowak M A, Komarova N L, Niyogi P. Evolution of universal grammar[J]. Science. 2001, 291: 114-118.

[98] Plotkin J B, Nowak M A. Language evolution and information theory[J]. Journal of Theoretical Biology. 2000, 205(1): 147-159.

[99] Komarova N L, Nowak M A. Natural selection of the critical period for language acquisition[J]. Proceedings of The Royal Society of London. Series B, Biological

Sciences. 2001, 268(1472): 1189-1196.

[100] Nowak M A, Krakauer D C. The evolution of language[J]. Proceedings of the National Academy of Sciences. 1999, 96(14): 8028-8033.

[101] Nowak M A, Plotkin J B, Jansen V A. The evolution of syntactic communication[J]. Nature. 2000, 404(6777): 495-498.

[102] Komarova N L, Nowak M A. The evolutionary dynamics of the lexical matrix[J]. Bulletin of Mathematical Biology. 2001, 63(3): 451-485.

[103] Nowak M A, Plotkin J B, Krakauer D C. The Evolutionary Language Game[J]. Journal of Theoretical Biology. 1999, 200: 147-162.

[104] Nowak M A, Komarova N L. Towards an evolutionary theory of language[J]. Trends in Cognitive Sciences. 2001, 5(7): 288-295.

[105] Pagel M, Atkinson Q D, Meade A. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history[J]. Nature. 2007, 449(7163): 717-720.

[106] Kirby S, Dowman M, Griffiths T L. Innateness and culture in the evolution of language[J]. Proceedings of the National Academy of Sciences. 2007, 104(12): 5241-5245.

[107] Atkinson Q D, Meade A, Venditti C, et al. Languages Evolve in Punctuational Bursts[J]. Science. 2008, 319(5863): 588-.

[108] Abrams D M, Strogatz S H. Linguistics: Modelling the dynamics of language death[J]. Nature. 2003, 424(6951): 900-900.

[109] Dorogovtsev S N, F J F. Language as an evolving word web[J]. Proceedings of the Royal Society of London. Series B, Biological Sciences. 2001, 268(1485): 2603-2606.

[110]Lieberman E, Hauert C, Nowak M A. Evolutionary dynamics on graphs[J]. Nature. 2005, 433(7023): 312-316.

[111] Hudson R. Language Networks: The New Word Grammar[M]. Oxford University Press, 2006.

[112] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine[C]. In: Proceedings of the Seventh International World Wide Web Conference. 1998. 107-117.

[113] Kleinberg J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM. 1999, 46(5): 604-632.

[114] Erkan G, Radev D. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization.[J]. Journal of Artificial Intelligence Research. 2004, 22: 457-479.

[115] Mihalcea R, Tarau P. TextRank: Bringing Order into Texts[C]. In: Lin D, Wu D. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004). Barcelona, Spain: Association for Computational Linguistics, 2004. 404-411.

[116] Wang J, Liu J, Wang C. Keyword Extraction Based on PageRank[J]. Advances in Knowledge Discovery and Data Mining. 2007: 857-864.

[117] Erkan G, Radev D. LexPageRank: Prestige in Multi-Document Text Summarization[C]. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004). Barcelona, Spain: 2004. 365-371.

[118] Mihalcea R. Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization[C]. In: Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Barcelona, Spain: Association for Computational Linguistics, 2004. 170-173.

[119] Lin Z, Kan M Y. Timestamped Graphs: Evolutionary Models of Text for Multi-Document Summarization[C]. In: Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing. Rochester, NY, USA: Association for Computational Linguistics, 2007. 25--32.

[120] Mihalcea R, Tarau P, Figa E. PageRank on Semantic Networks, with Application to Word Sense Disambiguation[C]. In: Proceedings of the 20th international conference on Computational Linguistics. Geneva, Switzerland: 2004. 1126-1132.

[121] Gamon M. Graph-Based Text Representation for Novelty Detection[C]. In: Proceedings of TextGraphs: the Second Workshop on Graph Based Methods for Natural Language Processing. New York City: Association for Computational Linguistics, 2006. 17--24.

[122] Nastase V, Szpakowicz S. A Study of Two Graph Algorithms in Topic-driven Summarization[C]. In: Proceedings of TextGraphs: the Second Workshop on Graph Based Methods for Natural Language Processing. New York City: Association for Computational Linguistics, 2006. 29--32.

[123] Hassan H, Hassan A, Noeman S. Graph Based Semi-Supervised Approach for Information Extraction[C]. In: Proceedings of TextGraphs: the Second Workshop on Graph Based Methods for Natural Language Processing. New York City: Association for Computational Linguistics, 2006. 9-16.

[124] Molla D. Learning of Graph-based Question Answering Rules[C]. In: Proceedings of TextGraphs: the Second Workshop on Graph Based Methods for Natural Language Processing. New York City: Association for Computational Linguistics, 2006. 37-44.

[125] Muller P, Hathout N, Gaume B. Synonym Extraction Using a Semantic Distance on a Dictionary[C]. In: Proceedings of TextGraphs: the Second Workshop on Graph Based Methods for Natural Language Processing. New York City: Association for Computational Linguistics, 2006. 65-72.

[126] Zhu M, Cai Z, Cai Q. Automatic keywords extraction of Chinese document using small world structure[C]. In: International Conference on Natural Language Processing and Knowledge Engineering. 2003.

[127] Huang C, Tian Y, Zhou Z, et al. Keyphrase Extraction Using Semantic Networks Structure Analysis[C]. In: Proceedings of the Sixth International Conference on Data Mining (ICDM '06). Washington, DC, USA: IEEE Computer Society, 2006. 275--284.

[128] Shi J, Hu M, Dai GZ. Topic Analysis of Chinese Text Based on Small World Model[J]. Journal of Chinese Information Processing. 2007(03).

[129] Liu J, Wang J. Keyword Extraction Using Language Network[C]. In: Proceedings of 2007 IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE '07). 2007. 129-134.

[130] Geng HT, Cai QS, Yu K, et al. A Kind of Automatic Text Keyphrase Extraction Method Based on Word Co-occurrence[J]. Journal of Nanjing University. 2006(02).

[131] Dong LB, Ma L, Jiao LC. Research on Text Cluster Method Based on Small World Model and Similarity Principle[J]. Journal of Information. 2006(2): 52-54.

[132] Mihalcea R. Random Walks on Text Structures[J]. Computational Linguistics and Intelligent Text Processing. 2006: 249-262.

[133] Salgueiro P T, Antiqueira L, Das G V, et al. Using Complex Networks for Language Processing: The Case of Summary Evaluation[C]. In: International Conference on Communications, Circuits and Systems Proceedings. 2006. 2678-2682.

[134] Geng HT, Cai QS, Zhao P, et al. Research on Document Automatic Summarization Based on Word Co-occurrence[J]. Journal of the China Society for Scientific andTechnical Information. 2005(06).

[135] Pang B, Lee L. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts[C]. In: Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL '04). Barcelona, Spain: 2004. 271--278.

[136] Pardo T, Antiqueira L, Nunes M, et al. Modeling and Evaluating Summaries Using Complex Networks[J]. Computational Processing of the Portuguese Language.

2006: 1-10.

[137] Antiqueira L, Nunes M G, Oliveira J O, et al. Strong correlations between text quality and complex networks features[J]. Physica A: Statistical and Theoretical Physics. 2007, 373: 811-820.

[138] Antiqueira L, Pardo T A S, Nunes M G V, et al. Some issues on complex networks for author characterization[C]. In: the Proceedings of the 4th Workshop on Information and Human Language Technology. Ribeiro Preto-SP, Brazil: 2006.

[139] Berendt B, Hanser C. Tags are not Metadata, but “Just More Content” to Some People[C]. In: Proceedings of International Conference on Weblogs and Social Media. 2007.

[140] Golder S, Huberman B A. The Structure of Collaborative Tagging Systems[J]. oai:arXiv.org:cs/0508082. 2005.

[141] Chi E, Mytkowicz T. Understanding Navigability of Social Tagging Systems[C]. In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems. ACM Press, New York, NY, 2007.

[142] Ohkura T, Kiyota Y, Nakagawa H. Browsing System for Weblog Articles based on Automated Folksonomy[C]. In: WWW-2006 Workshop on the Webloging Ecosystem. 2006.

[143] Mishne G. AutoTag: a collaborative approach to automated tag assignment for weblog posts[C]. In: Proceedings of the 15th international conference on World Wide Web. Edinburgh, Scotland: ACM Press, 2006. 953-954.

[144] Chirita P - A, Costache S, Handschuh S, et al. P-TAG: Large Scale Automatic Generation of Personalized Annotation TAGs for the Web[C]. In: Proceedings of the 16th international conference on World Wide Web. 2007.

[145]周强. 汉语句法树库标注体系. 中文信息学报. 2004, 18(4): 1-8.

[146]周明, 黄昌宁. 面向语料库标注的汉语依存体系的探讨. 中文信息学报. 1994, 8(3): 35-52.

[147] Beeferman, D., Berger, A.: Agglomerative clustering of a search engine query log. In: Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining. (2000).

[148] Baeza-Yates, R., Tiberi, A.: Extracting semantic relations from query logs. In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining. (2007).

[149] Wen, J.R., Jian-Yun, N., Hong-Jiang, Z.: Query clustering using user logs. ACM Transactions on Information Systems 20(1) (2002).

[150] Shen, D., Pan, R., Sun, J.T., Pan, J.J.,Wu, K., Yin, J., Yang, Q.: Query enrichment for web-query classi ¯ cation. ACM Transactions on Information Systems 24(3) (2006) 320-352.

[151] Beitzel, S.M., Jensen, E.C., Lewis, D.D., Chowdhury, A., Frieder, O.: Automatic classification of web queries using very large unlabeled query logs. ACM Transactions on Information Systems 25(2) (2007) 9.

[152] Baeza-Yates, R., Hurtado, C., Mendoza, M.: Query recommendation using query logs in search engines. Workshops on current trends in database technology of 9th international conference on extending database technology (2004).

[153] Chirita, P.A., Firan, C.S., Nejdl, W.: Personalized query expansion for the web. In: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval. (2007) 7-14.

[154] He, X.F., Yan, J., Ma, J.W., Liu, N., Chen, Z.: Query topic detection for reformulation. In: Proceedings of the 16th international conference on World Wide Web. (2007) 1187-1188.

[155] Zhou, T., Ren, J., Medo, M., Zhang, Y.C.: Bipartite network projection and personal recommendation. Physical Review E 76(4) (2007).

[156] Deshpande, M., Karypis, G.: Item-based top-n recommendation algorithms. ACM Transactions on Information Systems 22(1) (2004).

[157] Raghavan, V.V., Sever, H.: On the reuse of past optimal queries. In: Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval. (1995) 344-350.

[158] Fitzpatrick, L., Dent, M.: Automatic feedback using past queries: social searching? In: Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval. (1997) 306-313.

[159] Baeza-Yates, R., Hurtado, C., Mendoza, M.: Query clustering for boosting web page ranking. Advances in Web Intelligence (2004) 164-175.

[160] Sahami, M., Heilman, T.D.: A web-based kernel function for measuring the similarity of short text snippets. In: Proceedings of the 15th international conference on World Wide Web. (2006) 377-386.

[161] Fonseca, B.M., Golgher, P.B., de Moura, E.S., Ziviani, N.: Using association rules to discover search engines related queries. In: Proceedings of the first conference on Latin American Web Congress. (2003) 66-71.

[162] Broder, A.: A taxonomy of web search. ACM SIGIR Forum 36(2) (2002) 3-10.

[163] Kang, I.H., Kim, G.C.: Query type classi ¯ cation for web document retrieval. In:

Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval. (2003) 64-71.

[164] Gravano, L., Hatzivassiloglou, V., Lichtenstein, R.: Categorizing web queries according to geographical locality. In: Proceedings of the 12th international conference on information and knowledge management. (2003) 325-333.

[165] Baeza-Yates, R.: Graphs from search engine queries. Proceedings of the 33rd conference on current trends in theory and practice of computer science (2007) 1-8.

[166] Zhou, T., Jiang, L.L., Su, R.Q., Zhang, Y.C.: E?ect of initial conguration on network-based recommendation. Europhysics Letters 81(5) (2008) 58004.

[167] Ross, S.M.: Introduction to Probability Models, Ninth Edition. Academic Press, Inc., Orlando, FL, USA (2006).

[168] Kapp, A.V., Tibshirani, R.: Are clusters found in one dataset present in another dataset? Biostatistics 8(1) (2007) 9-31.

致　谢

衷心感谢导师孙茂松教授对本人的精心指导。他的言传身教将使我终生受益。感谢实验室的全体老师和同窗们学的热情帮助和支持!

本研究受到国家自然科学基金（项目号：60573187）和国家 863 计划（项目号：2007AA01Z148）的支持，特此致谢。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____日　期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1984 年 11 月 1 日出生于山东省新泰市。

2002 年 9 月考入清华大学计算机科学与技术系计算机科学与技术专业，2006 年 7 月本科毕业并获得工学学士学位。

2006 年 9 月免试进入清华大学计算机科学与技术系攻读计算机应用技术硕士至今。

发表的学术论文

[1] Zhiyuan Liu, Maosong Sun. Asymmetrical Query Recommendation Method Based on Network-resource-allocation Dynamics. WWW 2008 poster.

[2] 刘知远, 郑亚斌, 孙茂松. 汉语依存句法网络的复杂网络性质. 复杂系统与复杂性科学, Vol. 5, No. 2, pp. 37-45, 2008.

[3] 刘知远, 孙茂松.汉语词同现网络的小世界效应和无标度特性. 中文信息学报, Vol. 21, No. 6, pp. 52-57, 2007.

[4] 刘知远, 司宪策, 郑亚斌, 孙茂松.中文博客标签的若干统计性质. International Conference on Chinese Computing (ICCC'07), 2007.

[5] 刘知远, 孙茂松.基于 WEB 的计算机领域新术语的自动检测. 第九届全国计算语言学学术会议 (CNCCL'07), 2007.

[6] 郑亚斌, 刘知远, 孙茂松.中文歌词的统计特征及其检索应用.中文信息学报, Vol. 21, No. 5, pp. 61-67, 2007.

[7] Yabin Zheng, Shaohua Teng, Zhiyuan Liu, Maosong Sun. Text Classification Based on Transfer Learning and Self-Training. The 4th International Conference on Natural Computation (ICNC'08), 2008.

[8] 郑亚斌, 曹嘉伟, 刘知远. 基于最大匹配和马尔科夫模型的对联系统. 第四届学生全国计算语言学学术研讨会 (SWCL'08), 2008.