申请上海交通大学硕士学位论文

基于社会网络分析的Blog社区发现

学　　校： 上海交通大学

院　　系： 软件学院

班　　级： Z0503791班

学　　号： 1050379061

工程硕士生： 张　　浩

工程　领域： 软件工程

指导　老师： 吴刚（副教授）

导　　师： 姜丽红（副教授）

上海交通大学软件学院

2008年1月20日

A Dissertation Submitted to Shanghai Jiao Tong University for the Degree of Master

Blog Community Discovering Based on Social Network Analysis

Author： Zhang Hao

Specialty: Software Engineering

Advisor I: Associate Prof. Wu Gang

Advisor II: Associate Prof. Jiang Li hong

School of Software

Shanghai Jiao Tong University

Shanghai，P.R.China

January 20，2008

上海交通大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期：　　年　　月　　日

基于社会网络分析的 Blog 社区发现

摘　要

Blog 网络是一个由复杂超文本所组成的巨大信息源，而且以很快的速度在不断的扩大。针对这样一个不断变化的信息源，如何利用和发现 Blog 网络中的有用信息变得越来越具有挑战性。在 Blog 发展的过程中产生和演化了大量的社区，这些社区是 web 中非常重要的组织结构，也包含了大量有用信息。Blog 社区可以为用户提供有价值的、可靠的、及时的信息,并且代表着 Blog 网络中的社会活动。对 Blog 社区的深入研究有助于了解 Blog 的知识信息及其组织结构的发展状况。

本文对目前较为流行的一些 Blog 社区发现技术进行了分类和回顾，在此基础上提出了结合文本内容分析和社会网络分析的 Blog 社区发现方法。并且通过对比实验来和仅使用结构分析进行社区发现的方法进行比较，验证本文方法的可行性和有效性。本文着眼于提高社区挖掘的效率以及社区的质量，同时尝试进一步地挖掘社区中的信息。本文在使用社会网络分析理论，对链接结构特征进行分析的基础上，加入了 Blog 文章和评论内容分析的环节，使得找到的社区的凝聚性更强，稳定度更高。另外，将概率分布和统计的思想引入到社区的主题发现方法中，给出了利用主题词频率、文章时空分布以及评论反馈等统计结果和行为特征挖掘社区主题的方法。该方法既考虑了关键字的频率等文本内容，又充分利用了评论的内容与文章内容的联系以及人们发表文章的行为特征来挖掘社区中的有用信息。实验证明，本文给出的内容与结构分析相结合的社区发现方法在一定程度上提高了社区发现过程的执行效率和性能，并

且所得到的 Blog 社区的质量较高。另外，本文给出的社区主题信息发掘方法，为进一步对 Blog 社区进行深入的研究和数据挖掘提供了基础和保证。

关键词： Blog，SNA，社区发现，主题挖掘

Blog Community Discovering Based on Social Network Analysis

Abstract

Blog is a complicated collection of hypertext and expands with tremendous speed. Discovering and applying useful information of blogsphere is a challenging job. There are a lot of communities which have very important information in the Internet. Knowing these communities is helpful to understand the whole blogsphere and the whole web. Organizing Blog into communities has many advantages. With communities，users can navigate their interesting information，Internet service providers can arrange efficient ports，and manufacturers can find right consumers. Community also reflects sociality of Blog, because blogsphere is a social network.

So far many approaches have been proposed for detecting and mining blog communities. One of them is finding and maintaining some communities by human effort. It is costly and difficult to update. Nevertheless, there are still many unknown and newly emerged communities. Therefore some approaches and technologies are raised to find blog communities automatically or semi-automatically. The method of community extraction consists of two categories, one is structure oriented, and the other is content oriented. They have different data processing and analyzing methods. The former uses links and relations between the nodes of the network and the latter analyzes the text content of the web pages to find communities. But both of the approaches are not good enough and miss some important information in the processing step. This field is still new and there remain still many problems. In this paper we try to combine the structure analysis and content analysis together to improve the crawling efficiency and analyzing performance. We also designed and

conducted feasible experiments to support our method and the result turned out to be good.
After discovering a community we proposed a method to detect frequently discussed topics in the Blog community. We get the topics by analyzing the keywords frequency, spatiotemporal distribution of posts and comments response features of the blog community. This is very useful and provides important information for further mining in the community.

Keywords： Blog ，SNA，Community Discovering，Topic Mining

目　录

第一章 绪论

1.1 研究背景

一、 Blog 的发展

1. 什么是 Blog

互联网是世界上最庞大的信息资源库，随着互联网的不断发展，Blog 作为一种新兴的网络媒体，其技术、规范和标准正日益成熟与完善，规模也在迅速壮大。博客是“一种表达个人思想的网络链接，其内容按照时间顺序排列且不断更新”。Blog 一词由环球网 Web 和航海日志 Log 结合而成，意即“网络日志”。中文“博客”一词既指博客网站，又指博客作者，而英文博客作者的称谓则是 Blogger。

博客的结构包括文本、链接、图片、音频、视频等等，内容多为对个人生活或网络事件的记录、评论，是一种介于私人日记和网络向导之间的个人发布平台。博客的发展历史可以划分为萌芽期、初始期、成长期。博客发展的萌芽阶段年代为 20 世纪 90 年代早期至 90 年代末期，也被称为启蒙期。在萌芽阶段，博客最早可以追溯到被人垢病的“Justinps Home Page”，它主要记录的是个人对吸毒、性爱的体验，吸引了很多人。1998 年的旅行记成为博客发展的导火索，由此开始博客成为了一种潮流。随后 Pyra 工作室推出 Blogger 软件，博客发展完成萌芽阶段进入发展的初期阶段。从 2000 年到 2006 年是博客发展的初期阶段，在“911”时期和伊拉克战争时期的战争博客，使博客进入主流媒体视野，被全世界人所关注。2006 年以后，是博客的飞速发展时期，它成为一种全新的信息组织和传播方式，发挥越来越大的作用。据统计，目前全球约有 7000 万个 Blog 站点，而且正在飞速增长。

2. Blog 的分类

博客的分类方式多种多样，可以按照内容、技术，甚至博客作者的社会属性分类。按照博客作者的社会属性分类，博客可以分为成人博客、青少年博客、个人博客、群体博客等等。

按照应用技术的不同，博客可以分为文字博客、视频和移动博客等。视频博客与音频博客又被合译为“播客”。由于文字博客产生时间较长、技术较完善、普及程度

较高，而且本文的内容分析方法仅限于对文本内容进行，因此本文的研究主要针对文字博客。

按照内容分类又有多种标准。方兴东在《博客——时代的盗火者》一书中就将博客分作“战争博客、日记博客、知识博客、新闻博客、专家博客、技术博客、法律博客、文摘博客”等多种类型。这种分类方式尽管细致，但不成系统，难以应用于定量研究。论文《传播学角度的博客研究》作者庞大力则按照内容将博客分为三类：“一类是以时效性内容为主的博客，比如各类新闻的博客；一类是以专业性的知识为主的博客，专注于某一特定领域，进行知识过滤和知识积累；一类是以个人性的交流为主的博客，比如真正记录个人生活的日记，或者有着共同兴趣的人形成一个博客社区。”这一分类标准基本能够与美国博客研究接轨，同时被众多博客研究者引用，是当前较为主流的分类方式。

3. Blog 的特性

《新媒体观察》作者孙坚华将博客的特征概括为两个方面：内容的个性化表达和日记体方式。除此之外，还有更新频繁、充分利用链接、拓展文章内容、知识范围以及与其它博客相互联系等特点。这种高度精炼的概括可以扩展为以下内容：

- 私有性

一个博客只能由一个或一群特定的作者进行更新，这也是博客与论坛和新闻组的最主要区别。博客本身就是一种简易的个人信息发布方式，任何人都可以将自己的意见、看法、感悟、经历上传到自己的博客上，也可以通过转载相关信息表明自己的立场。从这一点来看，博客是真正意义上的个人传媒工具。

- 链接性

网络媒体相对于传统媒体的最大优势在于前者有“超链接”技术。所谓的超链接是指从一个网页指向一个目标的连接关系，这个目标可以是另一个网页，也可以是相同网页上的不同位置，还可以是一个图片、一个电子邮件地址、一个文件、甚至是一个应用程序。通过超链接技术，小小一个网页就可以成为通向网络上海量资源的窗口。对于博客来说，使用超链接可以有效的节省网络资源例如博客网页中的图片、动画、音乐等内容，如果通过本地上传到自己的博客网页就会占用自己的网页内存，而通过超链接就可以直接利用网络上已有的资源。同样的，文字类的博客也可以利用超链接直接引用网络上原作品的地址，而不必整篇引用。

- 便捷性

博客让用户能便捷地更新网页内容。用户只需要写作文章的题目、日期、正文等

非常简单的内容，并提交确认，博客模板就会自动将文章以适当格式更新至博客主页上，生成新的文章页面，对其赋予一个链接，并将文章添加至相应的日历存档当中。访问者可以根据日期、作者或其他特征来排列、挑选特定的文章进行阅读。用户不需要拥有任何网页编辑或网站建设知识，就可以编辑博客的内容，真正实现了“零门槛”。

➢ 开放性

如果把电子论坛比喻为开放的广场，那么博客就是一个人开放的私人房间，这就是博客与传统日记的区别。只要博客写作者们把博客作为一个信息传播工具，而不仅仅是个人信息的载体，这就意味着自己的博客已经成为公共领域的一部分。传统意义上个人与公共之间的门槛已经消失于无形。

➢ 交互性

博客是交互性的，与传统的单向媒体完全不同。博客的读者和编者可以实现真正意义上的实时互动，甚至读者和编者的身份也模糊了，二者之间成为真正意义上的对话者。硅谷专栏作家丹·吉尔默认为，博客作者是博客网站的核心，而围绕着博客与博客，博客与读者，读者与读者间多重交互的沟通是关键。没有交互就没有生命。

4. 社会化的 Blog

从上面的特征中可以看到，在某种意义上，Blog 也是一种社会化的软件。传统软件是把人们连接到计算机或网络上的人机交互工具。因此，技术悲观者认为互联网是一种终极的隔离技术，将使人们逐渐减少社会参与和人际交往，最终导致人们与社会脱离，由此也引发了人们对互联网“去社会性”趋势的忧虑。而“社会性软件”则按照每个人的思想、兴趣、观点、需求等把人们联系起来，是一种“人——工具——人”的交互方式。工具是为了最大程度地帮助人们自由地建构个人的社会关系和结构，既包括对传统社会关系如家人、亲戚、朋友、同事之间的维系和加强，也包括一种全新的社会关系如志趣相投的网友、协同工作的网络合作者等的形成。

社会化软件让人们能够通过电脑通讯媒介而聚集、联系或合作，并形成在线交流。最初的电子邮件、论坛系统、网络游戏和即时通讯软件等，均已具备一定的社会关系功能，这些社会化软件功能比较简单，主要是“一对一”模式，以实现通讯交流的功能为目的，因此我们称上述几类软件为“第一代社会性软件”。随着“六度分隔”理论与法则的渗透，社会化软件正在发生着颠覆性的变化。上世纪，专注于社会关系的互联网软件诞生，典型的代表是博客、维基、社交网络服务等。这类软件社会性功能趋于完善，涵盖从个人导向到群体导向，从简单的通讯到群体的网络协同工作，以及计算机支持协同工作等众多方面的功能，极大地提升了互联网用户的人际交流和社会

交往，这类软件被称为“第二代社会性软件”。可以预见的是随着互联网技术的发展，社会性软件将变得更加高效和人性化，人们对以此为纽带的社会网络也将变得更加依赖。

社会化软件具有促进网络社会圈形成与构建的功能。不仅因为社会化软件长期写作积累所形成的个性化身份特征，以及留言评论、友情链接等引发的社会联系，还由于社会化软件在技术层面上具有促进社会网络构建的功能，这些关键技术主要有：

- 引用通告（TrackBack & Ping）

引用(TrackBack)与通告(Ping)是相对应的一组功能。通告是指当作者引用其他用户贡献内容时，系统能够发出一个信号给对方。引用是指透过这项功能可以得知某篇内容究竟被多少其他用户所引用。引用通告一方面能够帮助读者追溯到信息源头，另一方面也有助于作者之间建立联系。

- 输出标准格式RSS FEED

使得用户贡献更新信息能够被订阅，从而有利于用户贡献内容的传播。通过订阅各种社会化软件的，读者可以很快地掌握像是作者、发表时间、标题、描述等属性数据，也可以用这些属性数据来检索、比对、排序、重组等。

- 自定义标签Tag

该技术主要是让作者给自己贡献内容进行标签定义，以实现对贡献内容的分类。由于这种分类内容具有由用户设定，因此称为“由下至上”的分类法。Tag在英文中是比较口语化的词，表示一群人、一伙人的意思。是指群众自发性定义的平面非等级标签分类。

可以看到，Blog 的发展方兴未艾，并且在深刻地影响和改变着我们的生活。越来越多的研究人员投身于 Blog 相关技术的研究，越来越多的普通人也加入到 Blogger 的行列，发布文章和评论。通常，一个 Blog 站点对应一个主人，因此，我们可以用 Blog 站点的 url 作为这个人的唯一标识。而每个 Blog 站点会包括很多针对作者文章发表的评论，每条评论也对应一个唯一的作者，包含着评论者 Blog 站点的 URL。由这些评论的相互联系，无数的 Blog 站点相互联结，形成了许多庞大和错综复杂的 Blog 网络。这些 Blog 网络虽然是无形的，但是从某种程度上说，它反映了现实生活中的社会网络。Blog 之间的联系也正是现实社会中人与人的联系.

二、虚拟社区

在网络世界中，人们利用互联网络相互沟通，通过互动形成虚拟社区，它是人际关系、共享经验的累积与凝聚。由互联网络架构出来的虚拟社区，不仅提供了信息流通的通道，同时也累积了这些信息中所蕴含的知识，形成一种巨大的知识仓库。随着

信息技术的发展，互联网络上的虚拟社区已成为一种重要的知识共享平台。互联网络技术的发展同时使得人与人之间知识和情感的来源和表现形式更加多样化。电脑和网络技术的结合创造了虚拟沟通的可能性，从而扩大了人们在互联网络上建构社会网络的形式与空间。当互联网络连接起一台又一台电脑之时，同时也就联系了这一台又一台电脑的使用者，这样电脑的使用者通过互联网络架构了一个社会关系网络。这个完全通过互联网络所构建的社会网络是虚拟社区的重要基础。虚拟社区中的社会网络与真实社区中的一样，也存在人际关系中的强联系和弱联系等人际网络关系特性，从而能够在虚拟社区中提供信息交换、知识共享和社会支持。简单的说，互联网络的发展突破了人们建构人际关系与社会网络必须通过有限节点的先天限制，使得人们都能轻易地通过互联网络自由的建构起个人的社会联系。互联网络发展之初，使用者便互相分享资料、解答问题、交换意见，共享的精神一直是网络的特色，网络使用者也是从知识的共享开始逐渐发展出情感的联系。

最早的关于虚拟社区的定义由瑞格尔德(Rheingole)做出，他将其定义为“一群主要藉由计算机网络彼此沟通的人们，他们彼此有某种程度的认识、分享某种程度的知识和信息、在很大程度上如同对待朋友般彼此关怀，从而所形成的团体”。虚拟社区至少具有以下四个特性：

1. 虚拟社区通过以计算机、移动电话等高科技通讯技术为媒介的沟通得以存在，从而排除了现实社区；
2. 虚拟社区的互动具有群聚性，从而排除了两两互动的网络服务；
3. 社区成员身份固定，从而排除了由不固定的人群组成的网络公共聊天室；
4. 社区成员进入虚拟社区后，必须能感受到其他成员的存在。

从社会学的角度看，虚拟社区是指由网民在电子网络空间进行频繁的社会互动形成的具有文化认同的共同体及其活动场所。由此可见，虚拟社区与现实社区一样，也包含了一定的场所、一定的人群、相应的组织、社区成员参与和一些相同的兴趣、文化等特质。而最重要的一点是，虚拟社区与现实社区一样，提供各种交流信息的手段，如讨论、通信、聊天等，使社区居民得以互动。但同时，它具有自己独特的属性：

- 虚拟社区的交往具有超时空性，人们之间的交流不受时间和地域的限制
- 人际互动具有匿名性和彻底的符号性
- 人际关系较为松散，社区群体流动频繁
- 自由，平等，民主，自治和共享是虚拟社区的基本准则

显然，Blog 社区属于虚拟社区的范畴，也具有虚拟社区的所有特点。

三、研究意义和价值

在 Blog 世界中，越来越多地通过创立一个网上虚拟社区平台来形成一个有良好知识共享环境的社区。通过社区中的人际互动，个人知识成为社区的共享知识。通过具体的协作，这些知识又被结构化。另外，越来越多的人热衷于通过虚拟社区进行交流，这种交流又受着各种因素的影响。虚拟社区内的人际关系对学习和知识共享的影响越来越受到人们的关注。

虚拟社区和现实社区并不是完全独立的，他们之间的关系就如同物质和意识之间的关系一样。网络社区来源于现实社区，虚拟社区是现实空间在虚拟空间的“投影”。首先，虚拟社区提供的服务版面也是根据人们现实的需要而设定的；现实社区中的生活方式和观念、规范会影响到虚拟社区的构建。其次，虚拟社区所提供的服务是现实社区的服务的延伸和提高。脱离现实，虚拟社区是不可能存在的。网络社区与现实社区是互补互动关系，从根本上是一致的。二者应该各取所长，互相弥补。网络社区使现实社区中的不可能成为可能，网络社区空间开拓了人的思维。从网络社员的观点来看，所谓现实性，无非是从以前的一种可能性发展而来的。二者是互补而非取代的关系。网络社区是一种对现有生活方式的冲击，同时，它也是对现实的社会空间的发展。

无论其规模的大小，对于一个 Blog 社区来说，进行研究和分析都有重要的意义。一个大的社区是由一个个或大或小的社区所组成的。任何一个社区就是一个规模不等的具体的小的社会，是整个大社会的不同程度的缩影。从一定意义上说，社区研究是研究整个社会的起点。同整个大社会相比，社区则显得具体可感，易于把握。一般地说，社会的一切活动都是在一个个具体的社区里进行的。整个社会普遍存在的一些现象必然会在各个社区里有所表现。社区研究是社会研究的具体化，人们通过社区研究对社会进行典型调查，从微知著，研究和探讨社会发展的普遍规律及同类社区的共同特点。对互联网上庞大的 Blog 网络进行社区发现以及深入的信息挖掘是目前国内外研究的热点，是十分有现实意义和价值的。

1. 这些社区为了解互联网用户的兴趣提供了有价值的，甚至是最及时，最可靠的信息。

2. 对 Blog 网络的文章和评论内容进行分析和统计，可以了解人们所关注的话题和领域，监测公众舆论的焦点。

3. 社区发现技术的研究及改进这些社区展现了互联网社会学，研究和发现这些社区可以深入了解互联网的进化过程。

4. 通过对 Blog 社区的结构进行分析，可以挖掘出一些有用的信息，从而发挥其商业价值。比如可以挖掘出社区中最有影响力，最受关注的站点，在其上有针对性投

放商业广告或市场调查，可以取得更好的效果。

1.2 研究现状

1. Blog 社区发现

在Blog的信息挖掘上已经有了一些研究和探索，其中在社区发现方面主要集中在内容分析、Blogger的参与程度分析以及聚类和结构分析三个方向。Nardi et al. [28] 采用了前两种方法的结合，对Blog站点的文章和评论进行文本分析，然后对Blogger组织调查和采访来了解Blog站点的参与程度及社区归属感，并用磁带记录下来，最后根据两方面的结果相结合来发现社区。另外，有人通过聚类和图论算法来发掘社区。因为Blog站点本质上就是网页，因此可以使用Web搜索算法来进行挖掘。Kleinberg就是如此，他使用hub和authorities的概念来发掘社区。其中hubs是包含指向关于同一个主题的其它页面的链接集合的页面，而authorities是指包含关于特定主题相关信息的页面，通过发现相关的authorities来发现社区。Kumar[22]扩展了hubs和authorities的概念，通过使用联合引用的概念来发掘Web上的所有社区，并使用图论算法找出所有代表社区的图结构。Flake [32] 等人使用联通性和图论方法来发现社区。Merelo -Guervos et al 使用模式和数据挖掘算法来对社区进行分组。以上这些方法和途径都取得了一定的成果，为后面的研究奠定了基础。但是同时也存在着一些问题。采用调查采访的形式需要高度的人工干预，主观性非常强，并且是非常耗时的，无法自动完成，也不能进行大规模的扩展。而使用聚类分析的方法则需要进行验证和调整。仅通过结构上的分析来发现社区的方法忽略和遗漏了文章和评论文本内容的重要信息。而单纯地对Blog站点的文章和评论内容进行分析则忽略了站点之间的交互特征以及Blogger的行为特征，都有一定的局限性。最后，以上这些工作都仅仅停留在发现社区上，而没有对发现的社区进行进一步的信息挖掘，没有体现出更大的价值。

2. 社区信息挖掘

与此同时，也有很多学者和研究人员在对单个Blog站点以及Blog网络的数据挖掘方面取得一定的成果。Teng等[7]提出了根据文本内容，时间特性和交互特性来研究Blogger的兴趣取向的方法；Shen等[6]研究了一种两阶段方法对具有相同兴趣的Blogger进行聚类，发现有潜在的朋友；Sekiguchi等[3]提出了基于Blogger兴趣的Blog文章主题提取的方法；Chi等[11]则研究了Blogsphere的趋势分析。而在主题挖掘方面，Qamra等[1]提出了基于社区结构和时间特性的CCT方法，对Blog社区的文章进行按给

定关键字聚类，然后确定各类的标签。最后可以得到最热门的类别，然而过程和参数估计十分复杂。

3. 社会网络分析

通过对网络的结构进行分析可以得到一些十分有用的信息，主要用来研究 web 中的网络关系和结构。社会网络分析理论是被广泛使用的一种网络分析方法。有很多研究者探索了用社会网络分析理论来发现社区的途径。但是在哪些参数和度量方法可以用来最合适地描述社区的哪些特征上没有达成一致。Tyler et al. [31] 使用中介度和趋中度来发现 e-mail 系统中的社区。Efimova et al. [13]用 Pajek 工具把知识管理 Blog 中的 Blog 数据集进行了可视化，但是没有计算出度的分布特征和趋中性。Herring et al. [16]则首先通过排序系统列出前 100 个被引用最多的 Blog，然后用它们作为种子和核心，根据 Blog 之间的链接来扩展 Blog 网络，最后使用 Pajek 工具进行了可视化，并用社会网络分析理论中的重叠性，相关性、出度、入度等参数来分析和发掘其中的模式。

以上这些研究和方法都没有完整地提出一个统一的、正式的社区发现方法。而 Alvin Chin 等[9]则提出了一套完整的通过社会网络分析发现 Blog 社区的方法。他们主要利用 SOC（Sense of Community）以及 SNA（Social Network Analysis）的方法来发现社区。他们通过在起始站点上设置调查问题来获取成员的社区归属感和参与度，并用社会网络分析的方法来对抓取的结果进行分析，最后找出潜在的社区。但是同样地，这种方法需要对成员进行问卷调查，十分费时，可操作性不强，也遗漏了文章和评论内容中的重要信息。

1.3 本文工作

通过以上介绍，我们对 Blog 社区信息挖掘的基本背景以及研究现状有了比较深入的了解。本文的目标是从精心选取的起始站点出发，在庞大的 Blog 网络中准确、高效地寻找出符合一定条件的社区，继而对社区进行进一步的挖掘，得到社区的主题等更有价值的信息。

为了实现这个目标，本文的工作在吸取前人长处的基础上，作了一些调整和改进。传统的社区发现的大多数方法都是完全以链接结构为出发点进行分析来寻找社区，没有对页面内容进行分析，遗漏了一些重要信息。还有一部分人的工作在内容分析的基础上作了一些探索，但没有同时深入考虑结构和关系。而且，这些研究在找到潜在社

区之后，没有对社区的结构和内容进行深入的挖掘和探索以发现有用信息。因此这些研究虽然都取得了较大的成果和进步，但是仍然需要进一步改进和完善。

为了提高挖掘的社区质量以及挖掘的效率和性能，本文提出一种结构分析与文本内容分析相结合的方法来进行 Blog 社区发现。在结构分析方面，仍然以社会网络分析作为社区发现的主要方法和理论基础，但是在度量参数的选取上作了不同的调整。在内容分析方面，主要体现在原始的 Blog 网络采集过程中。通过精心设置一些规则和模式，分析 Blog 文章和评论的文本内容，仅仅抓取和文章内容相关的评论者站点，而丢弃内容不相关的站点。这样，可以很大程度上避免抓取不相关的站点，提高效率和性能。另外，得到的原始网络在理论上内容相关度更高，更趋于聚集和稳定。此外，本文在发现社区之后，对社区进行进一步的信息挖掘，得到社区的主题和其它有价值信息。最后，设计和实现了一个 Blog 社区发现应用系统，并精心设计了相关实验，通过对实验的数据和结果进行分析和讨论，来验证本文方法的可行性并找出缺点与不足。

因此，本文的工作是十分有意义的。首先，在结合以前研究优点的基础上，提出了一个系统性的社区发现及信息挖掘的完整方案；其次，通过结构分析与内容分析相结合，可以更好、更高效地利用 Blog 网络的信息地发现社区和挖掘信息。

1.4 本文结构

本文的结构组织如下：第二章将介绍目前本文将要用到的基本理论和方法，包括社会网络分析和文本内容分析以及其各自在本文中的应用。第三章介绍 Blog 社区发现的整体框架和具体实现步骤。第四章介绍原型系统的设计和实现。包括系统的总体结构和各个具体模块的实现以及相互关系。第五章介绍实验的方法设计，包括起始站点选取，参数估计等，并对实验结果进行分析讨论。第六章对全文进行总结，总结本文的创新和不足，并指出将来进一步研究和改进的方向。

第二章 相关理论与技术分析

2.1 社会网络分析

二十世纪 30 年代，Jacob Moreno 和哈佛大学的一组研究人员分别提出了社会网络模型来分析社会学中的现象和问题。社会学家发现社会实体之间存在着相互的依赖和联系，并且这种联系对于每个社会实体有着重要的影响。基于这样的观察，他们通过网络模型来刻画社会实体之间的关系，并进一步用来分析社会关系之间的模式和隐含规律。和以往社会学研究的方法不同，它提供了一种形式化、概念化的途径来看待“社会”这个研究对象的性质和发展进程，是一种应用性很强的社会学研究手段，有很多的应用。当社会学家建立一个准确一致的社会网络模型之后，就可以通过逻辑推理的方式来研究社会的性质。

由于数据收集方式的限制，早期的社会网络局限于一个小的团体之内，往往仅包含几十个结点。借助于图论和概率统计的知识，人工处理可以从中分析出一些简单的性质和模式。但是，随着现代的通信技术的发展，越来越多的数据被收集和整合在一起，建立一个大的社会网络成为可能。例如，可以通过电子邮件的日志来建立使用者之间的联系网络，或者通过网络日志及网络通讯录等方式将用户提交的联系人信息建立社会网络。所以，现在的社会网络规模比早期网络庞大，通常包含几千或者几万的结点，甚至有多达百万个节点的网络。面对这样庞大复杂的网络，简单的数学知识和原始的人工处理已经不可能进行有效的分析。社会网络分析是一种应用性很强的社会学研究方法，成功地解决了一些社会学问题上，得到了广泛的关注。随着信息技术的发展，越来越多的社会关系数据被收集。如果能够有效地对它们进行分析，必将加深人们对社会学的理解，促进社会学的发展。

从本体论的角度看，社会网络分析坚持一种实在论的本体论。认为社会结构是客观的存在各个行动者之间的关系，可以作为外在物对行动者产生作用。社会网络分析提供的就是对这种结构的分析，利用量化的语言对网络数据的结构进行描述。从认识论的角度看，社会网络分析认为世界是由网络而不是群体组成的，它把世界看成是网络的结构，把行动者之间的关系看成是资源流动物质的或者非物质的渠道，从而可以通过分析发现复杂的资源流动网络而不是简单的分层结构。基于这种认识论，我们就

可以根据行动者之间的关系模式来理解行动者的属性特征和网络的整体特征。它认为社会网络的结构特征决定了行动者之间关系，发生的环境只有在由各种关系构成的结构脉络中才能理解两个行动者的之间的互动关系。从方法论的角度看，社会网络分析用图论工具代数模型技术描述关系模式，认为从社会关系视角进行的社会学解释要优越于从个人属性的视角进行的解释。根据周涛，张际平[33]的总结分析，社会网络分析主要有以下内容和方法。

2.1.1 社会网络的涵义

社会网络指的是社会行动者及其间的关系的集合。换句话说，一个社会网络是有多个点社会行动者和各点之间的连线行动者之间的关系组成的集合。用点和线来表达网络，这是社会网络的形式化界定。社会网络这个概念强调每个行动者都与其它行动者有或多或少的关系。社会网络分析者建立这些关系的模型，力图描述群体关系的结构，研究这种结构对群体功能或者群体内部个体的影响。下面对社会网络这个概念做进一步说明：

1. 点　社会网络中的点是指社会行动者，边是行动者之间的各种社会关系。在社会网络研究领域，任何一个社会单位或者社会实体都可以看成点，或者行动者。例如行动者可以是个体、或集体性的社会单位，也可以是一个教研室、系、学院、学校、更可以是一个村落、组织、社区、城市、国家等，当然也包括网络上每一个虚拟社区的成员或社区本身。

2. 关系　每个行动者是通过各种关系联系在一起。在社会网络分析中，一些得到广泛研究的关系有，

- 个人之间的评价关系如喜欢、尊重等
- 物质资本的传递如商业往来、物资交流
- 非物质资源的转换关系如行动者之间的交往，信息的交换等
- 隶属关系如属于某一个组织
- 行为上的互动关系如行动者之间的自然交往，如谈话、拜访等
- 正式关系权威关系正式角色也是有关系性的，如教师学生、医生病人、老板职员关系等
- 生物意义上的关系如遗传关系、亲属关系以及继承关系等

社会网络分析者还重点关注行动者之间的“多元关系”，也就是联系。例如，两个学生之间可能同时存在同学关系、友谊关系、恋爱关系等。按联系的强弱可分为强

联系和弱联系。行动者与其较为紧密、经常联络的社会关系之间形成的是强联系。与之相对应，个人与其不紧密联络或是间接联络的社会关系之间形成的是弱联系。

但是在传递资源、信息、知识过程中，一般认为弱联系更具重要性。强联系之间由于彼此很了解，知识结构、经验、背景等相似之处颇多，并不能带来进一步的新的资源信息和知识，所增加的部分大多是冗余的。而弱联系所提供的资源信息或知识会比较具有差异性，如果在弱联系之间搭起某种形式的桥梁，就可以传递多种多样的资源信息和知识。网络中的虚拟社区就起到了这样的桥梁作用。

2.1.2 社会网络分析

社会网络分析主要是研究社会实体的关系连结以及这些连结关系的模式、结构和功能。社会网络分析同时也可用来探讨社区众个体间的关系以及由个体关系所形成的结构及其内涵。换句话说，社会网络分析的主要目标是从社会网络的潜在结构中分析发掘其中次团体之间的关系动态。

社会网络分析主要研究行动者以及彼此之间的关系。通过对行动者之间关系与联系的连结情况进行研究与分析，将能显露出行动者的社会网络信息，甚至进一步观察并了解行动者的社会网络特征。而通过社会网络，除了能显示个人社会网络特征外，还能够了解许多社会现象。因为社会网络在组织中扮演着相当重要的无形角色，当人们在解决问题或是寻找合作伙伴时通常都是依循所拥有的社会网络来寻找最可能提供帮助和协作的对象。

社会网络分析是社会科学中的一个独特视角，它是建立在如下假设基础上的：在互动的单位之间存在的关系非常重要。社会网络理论、模型以及应用都是建立在数据基础上的，关系是网络分析理论的基础。网络模型把结构社会结构、经济结构等概念转化为各个行动者之间的关系模型。

2.1.3 社会网络分析的主要内容

社会网络分析用于描述和测量行动者之间的关系或通过这些关系流动的各种有形或无形的东西，如信息、资源等。自人类学家 Barnes 首次使用“社会网络”的概念来分析挪威某渔村的社会结构以来，社会网络分析被视为是研究社会结构的最简单明朗、最具有说服力的研究视角之一。20 世纪 70 年代以来，除了纯粹方法论及方法本身的讨论外，社会网络分析还探讨了小群体(clique) 、同位群(block) 、社会圈(social

circle) 以及组织内部的网络、市场网络等特殊的网络形式。这些讨论逐渐形成了网络分析的主要内容。

根据分析的着眼点不同，社会网络分析可以分为两种基本视角：关系取向和位置取向。关系取向关注行动者之间的社会性粘着关系，通过社会连结本身如密度、强度、对称性、规模等来说明特定的行为和过程。按照这种观点，那些强联系的且相对孤立的社会网络可以促进集体认同和亚文化的形成。与此同时，位置取向则关注存在于行动者之间的、且在结构上处于相等地位的社会关系的模式化。它讨论的是两个或两个以上的行动者和第二方之间的关系所折射出来的社会结构，强调用“结构等效”来理解人类行为。

1. 关系取向中的主要分析内容

由于社会网络分析是以网络中的关系或通过关系流动的信息、资源等为主要研究对象的，这种取向中的主要分析内容也大多集中在网络“关系”上。其中几项重要内容如下：

- 规模(range) 社会网络中的行动者都与其他行动者有着或多或少、或强或弱的关系。规模测量的是行动者与其他行动者之间关系的数量。如果把研究的焦点集中在某一特定行动者节点上时，对关系数量的考察就变成了对网络集中性的考察。所谓的“集中性”是指特定行动者身上凝聚的关系的数量。一般说来，特定行动者凝聚的关系数量越多，他她在网络中就越重要。不过，关系的数量多少并不是行动者重要性的惟一指标，有时候行动者在网络中所处的位置比集中性更为重要。特别地，当行动者的位置处于网络边缘时，数量的多少就远不如桥梁性位置重要。
- 强度(strength) 格兰诺维特认为测量关系强度的变量包括关系的时间量（包括频度和持续时间）、情感紧密性、熟识程度相互信任以及互惠服务。如果花在关系上的时间越多、情感越紧密、相互间的信任和服务越多，这种关系就越强，反之则越弱。
- 密度(density) 网络中一组行动者之间关系的实际数量和其最大可能数量之间的比率称为密度。当实际的关系数量越接近于网络中的所有可能关系的总量，网络的整体密度就越大，反之则越小。与格兰诺维特的“情感密度”不同的是，网络密度只用来表示网络中关系的稠密程度，测量的是联系本身，而“情感密度”则是指联系的特定内容在情感上的亲密程度。
- 内容(content) 即使在相同的网络中，行动者之间的关系也会具有不同的内容。所谓网络关系的内容，主要是指网络中各行为者之间联系的特定性质或类型。任何可能将行动者联系起来的东西都能使行动者之间产生关系，因此内容的表现形式也

是多种多样的，交换关系、亲属关系、信息交流关系、感情关系、工具关系、权力关系等都可以成为具体的内容。

- 不对称关系(asymmetric ties)与对称关系(symmetric ties)　在不对称关系中，相关行动者的关系在规模、强度、密度和内容方面是不同的，而在对称关系中，行动者的关系在这些方面的表现都是相同的。例如，当信息只从行动者流向行动者，而行动者不向行动者提供信息时，两者之间的关系就是不对称关系。
- 直接性(direct)与间接性(indirect)　网络关系的另一个内容就是直接性或间接性，前者指行动者之间直接发生的关系，后者则指必须通过第三者才能发生的关系。一般说来，直接关系连结的往往是相同或相似的行动者，他们往往彼此认同，具有相同的价值观，因此其关系通常为强联系而间接关系中由于有中间人的存在，相互联系的行动者之间关系的强度受距离中间人的数量的影响很大，经历的中间人越多，关系越弱，反之则可能越强。

2. 位置取向中的主要分析内容

与关系取向不同的是，位置取向强调的是网络中位置的结构性特征。如果说关系取向是以社会粘着为研究基点，以关系的各种特征为表现的话，那么位置取向则以结构上的相似为基点，以关系的相似性为基本特征。在位置取向看来，位置所反映出来的结构性特征更加稳定和持久，更具有普遍性，因而对现实也更有解释力，且需要分析的内容也更为简单明了。其主要内容有：

- 结构等效性(structural equivalence)　当两组或两组以上的行动者（他们之间不一定具有关系）与第二个行动者具有相同的关系时，即为结构等效。这里强调的是在同一社会网络中所谓的等效点必须与同一个点保持相同的关系。网络中等效点的数量和质量将对网络的驱动力产生很大的影响。
- 位置(position)　作为位置取向的核心概念，位置在这里指的是在结构上处于相同地位的一组行动者或节点，是被剥落了行动者而剩下的结构性特征，哪个行动者处在这个位置上并不重要，重要的是这个位置在网络本身中的处境。
- 角色(role)　与位置密切相关的另一项内容是角色，它是在结构上处于相同地位的行动者在面对其他行动者时表现出来的相对固定的行为模式。反过来说，具有相同社会角色的往往在社会网络结构或地位网络结构中处于相同的位置。因此，角色在某种程度上是位置的行为规范。

2.1.4 社会网络分析研究方法

从上世纪六十年代至今，基于对以往的人际方法研究弊端的思考，一些研究者借助于快速发展的计算机技术，采用社会网络分析法研究人际关系。从不同的学术背景出发，社会网络研究沿着两个不同特色的研究取向平行发展，目前这两种分析方法都得到了广泛的应用。

1. 整体网络分析方法

此方法关注的焦点是整体网络，即一个社会体系中角色关系的综合结构或群体中不同角色的关系结构。该方向继承了社会计量学的研究传统，代表人物是林顿·弗里曼。目前的研究集中于小群体内部关系研究，探讨网络结构随时间变迁和网络中成员的直接或间接的联系方式。主要概念有桥梁性、紧密性、结合性。

从数据收集上来看，整体网络分析方法主要使用提名选择法、参数选择法与循环选择法等各种选择方法。从数据整理上来看，整体网络分析主要采用社会矩阵方法与社群图示法。社会矩阵是一个 N 阶矩阵。N 代表群体的人数，横行代表选择者，纵行代表被选择者，在选择者与被选择者交叉的地方标出选择结果，最后就可以得到该群体的整体网络矩阵。社群图示法则在一张图上标出所有的群体成员，然后使用箭头表示群体成员的相互选择情况，最为研究者所熟知的即为由同心圆组成的箭靶图。从数据分析上来看，主要采用矩阵解析、社群图分析方法以及使用有关指数，如声望指数、中心指数加以标示。具体而言，矩阵解析方法将社矩阵作为初级矩阵予以分析，平方之后的矩阵则代表群体成员之间的二级关系，立方之后则代表三级关系，以此类推。社群图分析方法则通过解剖社群图的基本结构，掌握群体中的社会网络分布情况，区分网络中不同地位的角色，如明星，联络人，孤立者等。声望指数的计算则包括相对声望指数与内部声望指数两种，指的是行为主体在关系中被选择为客体的比重或者在整体网络中的绝对人数。中心指数则通过计算行为主体介入的关系占据网络中所有关系的比重来获得。从数据处理采用的软件来看，整体网络分析主要采用独特设计的社会网络分析软件与社会测量软件。

2. 自我中心网络分析

这种研究方法集中于个体间的自我中心网络，从个体的角度来界定社会网络。该方法是沿着英国人类学家的传统发展的，它所关心的问题是个体行为如何受到其人际网络的影响。这个领域的著名代表性人物是马克·格拉诺维特、哈里森·怀特和林南

等。核心概念则主要是网络的范围、密度、强弱联系等。

从数据收集与数据整理上来看，根据荷兰学者范德普尔的总结，自我中心网络分析主要有以下几种方法互动方法：角色关系和情感方法以及社会交换法。情感方法要求被试者指出与其关系最为密切的人，如最好朋友提问法或者十项提名法。这种方法的缺点在于不同的人的评价标准可能不一致。社会交换法以社会交换理论作为基础，认为拥有报偿性互动资源的人在影响被试者的态度和行为的时候相当重要。这种方法目前得到了普遍运用，并且被证明在不同文化背景之中也是适用的。它的优势在于考察的是现实存在的关系，并且由于报偿性互动是相当特殊的，因此保证了所有被试者按照同一标准来回答问题。从数据分析与使用的软件来看，自我中心网络分析主要是运用 SPSS、SAS 等大型统计软件中的线性相关分析、协方差分析等模块来探索影响自我中心网络特征的因素。

2.1.5 社会网络的形式化表达

从数学角度上讲，有两种方法可以描述社会网络：社群图法和矩阵代数方法。社群图法常常应用于结构对等性和块模型的研究。代数学法可用于分析角色和关系。当然，其他统计方法也可以用来描述社会网络。社群图是由莫雷诺最早使用的，现已在社会网络中得到广泛使用。用来表达一种关系的矩阵叫做社区矩阵。社群图主要由点（代表行动者）和线（代表行动者之间的关系）构成。这样，一个群体成员之间的关系就可以用一个由点和线连成的图表示。

如果根据关系（线）的方向，可以分为“有向图”和“无向图”。无向图是从对称图中引申出来的，它仅仅表明重要关系的存在与否。如果关系是有方向的（例如借款关系、权力关系等），也就是说，a 到 b 的关系与 b 到 a 的关系是不同的，那么，就应该用有向图来表示。我们用代表有向线的集合，用代表其中的单条线，用箭头代表关系的方向。行总和与列总和构成一个有向图及其邻接矩阵。

利用社群图表达关系网络的一个优点是比较清晰、明确，并且社会行动者之间的关系一目了然。但是，如果社群图涉及的点很多，例如人，那么图形就相当复杂，很难分析出关系的结构，这是社群图的一个缺点。在这种情况下，我们最好利用矩阵代数法表达关系网络，用来研究多元关系，研究两种关系或者多种关系的“叠加”。这种方法最先由怀特和伯德提出来。

如果行和列都代表来自于一个行动者集合的“社会行动者”，那么矩阵中的要素代表的就是各个行动者之间的“关系”。这种网络是 1-模网络。如果行和列代表来自

两个行动者集合的“社会行动者”，那么矩阵中的元素分别代表的就是两个行动者集合中的各个行动者之间的“关系”，这种网络是2-模网络。如果“行”代表来自一个行动者集合的“社会行动者”，“列”代表行动者所属的“事件”，那么矩阵中的元素就表达行动者隶属于“事件”的情况，这种网络也是2-模网络，具体地说是“隶属关系网络”。

如果没有数学工具图论、矩阵代数的支持，社会网络分析就不可能取得重要进展。在表达关系数据的时候，社会网络分析者主要利用数学领域中的两种工具社群图和矩阵代数。当然，社会网络方法论上的突破也离不开统计技术的发展。拥有了这两种工具，我们就能够计算一些网络测度例如密度、出入度等参数。在社会网络中，与“关联性”密切相关的研究是行动者之间的距离。有的行动者可能与网络中的任何一个人都建立了联系，与其他人的距离都很“近”。有的人可能交往比较少，相对“孤立”一些。如果行动者之间的距离不一样，我们就可能找到这些行动者在网络意义上的社会分层来，也可能有助于我们理解社会群体的“同质性”、“团结性”等特点。

下面将介绍几种距离相关的概念，并用这两种工具阐述它们在社会网络分析中所代表的含义。

- 测地线　在给定的两点之间可能存在长短不一的多条途径。两点之间的长度最短的途径叫做测地线。如果两点之间存在多条最短途径，则这两个点之间存在多条测地线。

- 距离　两点之间的测地线的长度叫做测地线距离，简称为“距离”。也就是说，两点之间的距离指的是连接这两点的最短途径的长度。我们把点和之间的距离标记为：如果两点之间不存在途径即二者之间是不可达的，则称二者之间的距离是无限的或者无定义。如果一个图是不关联图，那么其中至少有一对点的距离是无限的。

- 直径　一个图一般有多条测地线，其长度也不一样。我们把图中最长测地线的长度叫做图的直径。如果一个图是关联图，那么其直径可以测定。如果图不是关联的，那么有的点对之间的距离就没有界定，或者就距离无穷大。在这种情况下，图的直径也是没有定义的。

- 密度　这个概念是为了汇总各个线的总分布，以便测量该分布而与完备图的差距有多大。固定规模的点之间的连线越多，该图的密度就越大。具体地说，密度指的是一个图中各个点之间联络的紧密程度。

- 权力和中心性　“权力”是社会学中的一个重要概念。从社会网络的角度对权力的这种界定可以进一步体现在网络研究者对权力的各种定量表述上。也就是说，网络分析者是从“关系”的角度出发定量地界定权力的，并且给出多种关于社会权力

的具体的形式化定义，即各种中心度和中心势指数。这可以看成是网络分析者的独特贡献，因为网络研究者更倾向于用“中心性”表达权力概念。

“中心性”是社会网络分析中的重点之一。个人或者组织在其社会网络中具有怎样的权力，或者说居于怎样的中心地位，这一思想是社会网络分析者最早探讨的内容之一。这个观点最初体现在社会计量学的一个重要概念——“明星”。所谓明星指的是那个在其群体中最受关注的中心人物。巴乌拉斯最先对中心度的形式特征进行了开创性研究，验证了如下假设：即行动者越处于网络的中心位置，其影响力越大。研究发现，中心度与群体效率有关，也与参与群体的个人的满意度有关。随后的学者用这个概念解释复杂的社会系统。

在社会网络分析中对权力的探讨集中体现在对“中心度”和“中心势”的量化分析上。常用的中心度和中心势指数包括点度中心度、中间中心度、接近中心度、特征值中心度以及伯纳西茨权力指数，还有与它们对应的中心势指数。中心度刻画单个行动者在网络中所处的核心位置，中心势刻画的则是一个网络所具有的中心趋势。假设研究点度中心性，那么对于一个拥有 m 个行动者的网络来说，其中可以计算出来的中心度指数有 m 个，但是计算出来的中心势指数只有一个。

“点度中心度”刻画的是行动者的局部中心指数，测量网络中行动者自身的交易能力，没有考虑到能否控制他人。“中间中心度”研究一个行动者在多大程度上居于其他两个行动者之间，因而是一种“控制能力”指数。“接近中心度”考虑的是行动者在多大程度上不受其他行动者的控制。如果网络中的一个行动者在交易的过程中较少依赖于他人，此人就具有较高的中心度。一个点越是与其他点接近，该点就越不依赖于他者。刻画一个行动者的特征向量中心度是为了在网络总体结构的基础上，找到最居于核心的行动者，而不关注“局部”的模式结构。对中心度的测量不能脱离其他点的中心度。因此，在计算中心度的时候包含着内在的循环。

2.2 文本主题提取

2.2.1 概述

主题提取的概念最早由 Luhn 于 1958 年提出。他的基本思想是：“机器利用词语出现的频率和分布等统计信息计算词语及句子的相对重要程度，提取并输出重要度最高的句子，从而获得“文本主题”。在此后的几十年中，主题提取的研究领域中出现了很多重要的进展。目前，不同的方法主要可以分成两大类：基于统计的机械式主题

提取方法和基于语法语义分析的理解式主题提取方法。

1．机械式主题提取方法

由于避开了语义分析的难点，基于统计技术的机械式主题提取一直是最常用的方法。它的核心思想是：根据特殊的统计特征，计算每个语言单元（通常是句子）的重要度，最后将最重要的句子抽取出来，形成主题。计算句子重要度分数时，需要识别某些统计特征并进行加权，如关键词频率、位置线索、标题词线索、提示词线索和指示词语线索等。最后，对上述所有权值求和，就得到了整个语言单元的值。为了提高抽取的准确度，统计模型中借用了信息检索中的一些标准技术，如 TF-IDF、VSM、RF 及 LCA 等。因为不需要深刻理解文本的语义，机械式抽取主题的方法可以应用到各种题材的文章中。

2．理解式主题提取方法

基于语义分析和理解的主题提取方法需要较成熟的人工智能技术和大型的专家知识库，对文章进行深层的句法和语义分析。典型的理解式方法使用预定制的模板，从原文中提炼重要的信息填入模板中，从而生成主题。采用这种方法的方法的系统包括 FRUMP、TOPIC、SCISOR 和 SUMMON 等。这些系统可以产生连贯的、符合文体要求的主题。它的主要缺点是，由于需要庞大的专家知识库和完善的语言学规则，只能应用到某些特定题材的文体和内容具有相当可预见性的文章中。当前，自动主题提取研究的主要方向是基于统计的机械式方法。但是，越来越多的现象表明，统计并不能完全取代语义分析。不考虑句子的含义和句子间的关系机械抽取，必然导致主题的准确率低，连贯性差，产生一系列问题，如主要内容缺失、指代词悬挂、文摘句过长等。因此，理想的自动主题提取模型应当将两种方法相结合。本文所提出的方法是将语义分析融入统计算法，同时结合文本聚类技术进行主题抽取。其基本的方法仍然是“统计-抽取”模型，因为这一技术已经相对成熟并拥有丰富的研究成果。

2.2.2 文本模型

目前的文本模型主要是 Gerard Salton 和 Mc Gill 于 1969 年提出的向量空间模型。向量空间模型的基本思想是把文档简化为以特征项的权重为分量的向量表示。在 VSM 模型中，文档空间被看作是由一组正交词条矢量所组成的矢量空间，每个文档表示为其中的一个范式化特征矢量：

$$V(d) = (t_1, w_1(d); t_2, w_2(d); ...t_n, w_n(d);) \qquad (2\text{-}1)$$

式中t_i为词条项，$w_i(d)$ 为t_i在d中的加权值。可以将d中出现的所有单词作为t_i，也可以用d中的所有短语作为t_i，从而提高内容表示的准确性。$w_i(d)$一般被定义为在中出现频率$tf_i(d)$的函数，即$w_i(d)=\psi(tf_i(d))$。权重用词频表示，而词频分为绝对词频和相对词频。绝对词频，即用词在文本中出现的频率表示文本；相对词频即为归一化的词频，其计算方法主要运用TF-IDF方法。TF-IDF函数如下：

$$tf_i(d)\times\lg\frac{N}{n_i} \qquad (2\text{-}2)$$

其中，N为所有文档的数目，n_i为含有词条t_i的文档数目。文本之间相似度的计算有多种方法可供选择。最简单的方法是仅考虑两个特征矢量中所包含的词条的重叠程度；而最常用的方法是考虑两个特征矢量之间的夹角余弦，即

$$Sim(d_k,d_i)=\frac{V(d_k)\bullet V(d_i)}{|d_k|\times|d_i|} \qquad (2\text{-}3)$$

通常，为了确定哪些句子可以作为候选主题句，首先需要识别文本的主题概念，即首先需要从文本中抽取出反映文本主题概念的字串。词频统计是定量衡量概念重要性的重要手段，通常采用 TF-IDF 公式计算字串的权重。通过 TF-IDF 公式，根据词长、位置等信息的调整，可以计算出文本中出现的词等字串的权值。显然，具有较高权值的字串将构成文本主题字串的候选集。许多文本自动处理系统和相关的文献都采用了以上方法。这种方法通常假定文本中的一个特定的词对于文本主题的贡献度独立于其他词。做此假定的目的主要是为了简化处理工作，即词间的相互关系被忽略了。而在真实文本中，实际上作为分量的词汇或字串间往往具有很大的相关性。 由于该假定对于文本处理应用的不准确性，将会增加文本自动处理应用的出错率。同一文本中的词、字串之间存在很强的相互关系，如同义关系、共现关系等，对这类关系进行分析将有助于提高文本分析的准确性。例如“自行车”和“单车”、“二轮车”等表达了相同的概念。建立概念间的关系，可有多种方式，如本体论方法。一般情况下，多数作者不愿意在文中多次重复同一词汇，常用多个同义词表示同一概念。同义归并是一种最基本的概念归并。对于文本中出现的同义词，用代表该同义词簇的标准字串统一替换，并在计算主题字串的权重时，对它们统一进行度量。假设T出现在文本D_j中，

T 的同义词 $T_1, T_2, ..., T_n$ 也在文本 D_j 中出现。则可对 $T, T_1, T_2, ..., T_n$ 进行概念归并，并调整相应的权重度量。对于同一文中出现的上下位概念也可进行概念归并，即进行权值调整。若文本涉及多个主题，并且该多个概念有共同的上位概念，则常用上位概念来表示这些主题。 然而如不加区分地使用这种方法表示文本主题，有时会降低文本标引的专指度，使得主题的表示过于空泛。因此，要根据具体的情况，在标引的专指度和概括性之间进行权衡。

第三章 结构分析与内容分析相结合的 Blog 社区发现

3.1 概述

我们知道，Web是一个非常庞大的信息资源库，并在不断地增长。其数据存在无组织、海量等特征，但是Blog作为有一定结构性的页面，仍然存在一些规律。从宏观上看，Blog具有与规模无关的节点连接度幂律分布和小世界效应等统计特性。从微观上，Blog通过链接结构在拓扑结构上的聚团性与Blog网页的内容聚团性在一定程度上相关。Blog社区发现及其关系分析是一个热门话题，但是对Blog社区的严格定义目前还没有一个统一的体系。在Blog网络中某些站点间有着明显的关联性，相互之间存在大量的链接，并且在内容上相似度较高。这种Blog站点的聚类，可被称为Blog社区。它是基于物理连边的聚集概念，即Blog社区可以松散地被定义为基于某个特定主题的、相互链接的Blog站点集。

从Blog网络的拓扑结构分析角度看，一个社区就是指一个Web图的子集，在这个子集内的节点之间连接密度高于子集内部节点与外部节点的连接密度，可见拓扑结构社区是一个层次化的概念，即一个大的社区可能包含了若干的小社区。主题相同或相似的页面和站点常常相互稠密链接在一起，本身呈现出社区现象，反映了Web中普遍存在的、复杂的聚团关系和层次关系。无论具体的社区定义如何，社区结构都反映了Blog中存在的层次现象，所以对Blog社区的层次分析相当于从不同的颗粒度来分析Blog的性质。

从Blog网络的内容相似性角度看，Blog社区是客观存在的一些Web群体，它们在内容上一般都是围绕某一主题具有一定的相关性，或者具有某一相似特性，即Blog社区内容具有初步的自组织性，如何发现这些潜在的Blog社区是近几年来引起众多研究者关注的研究领域。反之，一个活跃度高，稳定的社区的Blog站点所发布的文章和评论在内容上必然倾向于有一定的主题。

Blog站点之间有通过好友链接形成的直接、精确的链接，也有通过发表评论或转载形成的间接链接。我们的目标是通过对这些链接形成的Blog网络进行结构分析和内容分析，找到潜在的社区，并挖掘出社区的主题。下面首先介绍总体的实现思路和结构，然后详细阐述具体步骤。

3.2 总体框架

本文的社区发现方法在理论和技术上，采取基于社会网络理论的链接结构分析与文本内容分析相结合进行。主要分为 3 个步骤：原始网络抓取，潜在社区发现，社区主题挖掘。

1. 基于内容分析的原始网络抓取

本文在原始网络的抓取过程中加入了内容分析的环节，以提高抓取的性能和准确度。我们的 Blog 爬虫从给定的种子站点出发，根据设置的规则和参数抓取相应的 Blog 站点 url 及其文章和评论的内容。

在本文中，不考虑转载和引用文章以及好友链接，仅以文章的评论为线索进行抓取。在抓取的过程中，每抓取一个站点的一篇文章，都要对该文章的所有评论的文本内容进行分析，根据分析的结果，动态地决定添加或丢弃该站点。这样可以有效地缩小抓取和分析的数据量。同时考虑到六度分割理论，即每两个人之间平均通过六个人就可以建立联系。如果不添加限制条件，爬虫可能会无休止地抓取下去，无法结束。因此我们设置抓取的深度为 2，即抓取到评论者的评论者为止，不再向外延伸。抓取完成后，把抓取的结果存放在 XML 文件中。实际上抓取的结果是一个连通图。

2. 基于结构分析的社区发现

把抓取站点及其相互关系提取出来，建立对应的网络模型，并使用网络分析工具可视化。然后去除噪音，并对可视化后的网络结构进行社会网络分析。根据选取的参数，用社会网络分析软件进行分析，得到一个或几个潜在的社区，同时得到社区的核心节点等信息。

3. 社区主题挖掘

对分析得到的社区进行进一步的信息挖掘，通过关键字频率，时空分布、评论反馈特征等因素的分析得到社区的主题。这里，主题表示为关键字的集合。

图 3-1 阐述了社区发现的总体框架和结构。可以看到，内容分析是在根据评论链接抓取相关站点的同时进行的，由于需要在线对文本内容进行分析，可能会受到网络速度，计算机性能等因素的影响。但是由于一个 Blog 站点往往包括较多的文章，通过分析其评论内容的相关性来取舍评论站点，可以节省更多的抓取不相关站点信息的时间和空间。

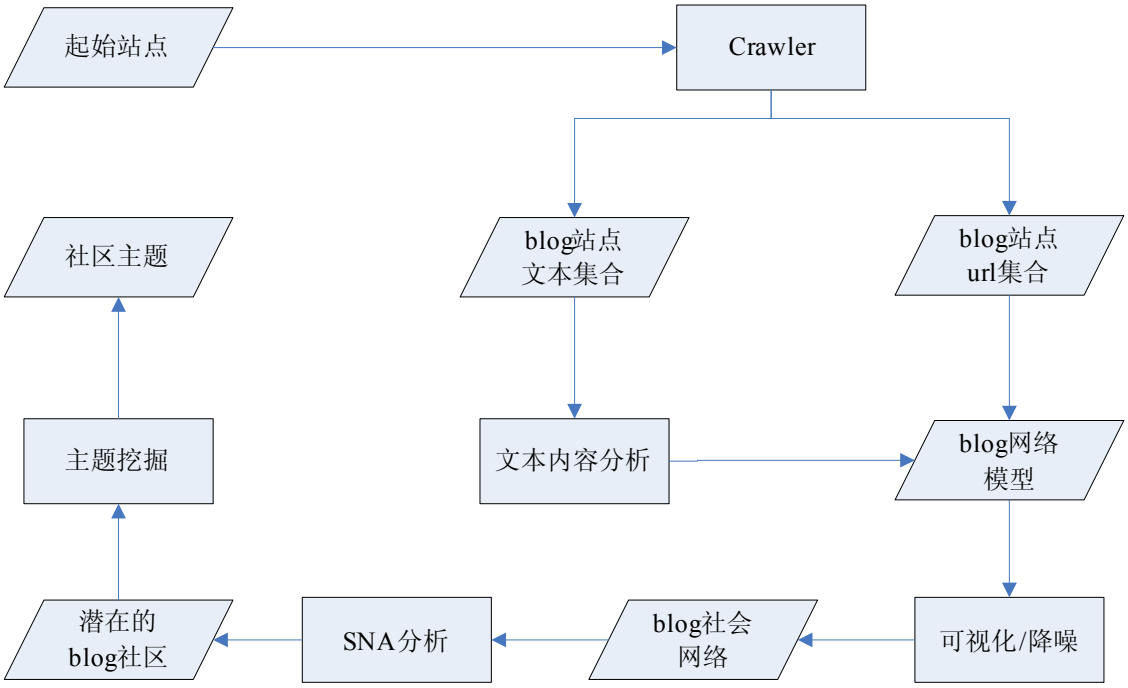

图 3-1 社区发现的总体框架

Fig. 3-1 Framework of Blog Discovering

3.3 具体工作

3.3.1 原始网络采集

1. Blog 爬虫

本节讲述网络爬虫进行原始网络抓取的流程和算法。网络爬虫是一个自动提取网页的程序，它为搜索引擎从万维网上下载网页，是搜索引擎的重要组成。传统爬虫从一个或若干初始网页的 URL 开始，获得初始网页上的 URL，在抓取网页的过程中，不断从当前页面上抽取新的 URL 放入队列，直到满足系统的一定停止条件为止。网络爬虫主要分为聚焦爬虫和非聚焦爬虫。聚焦爬虫的工作流程较为复杂，需要根据一定的网页分析算法过滤与主题无关的链接，保留有用的链接并将其放入等待抓取的 URL 队列。然后，它将根据一定的搜索策略从队列中选择下一步要抓取的网页 URL，并重复上述过程，直到达到系统的某一条件时停止。所有被爬虫抓取的网页将会被系统存贮，进行一定的分析、过滤，并建立索引，以便之后的查询和检索；对于聚焦爬

虫来说，这一过程所得到的分析结果还可能对以后的抓取过程给出反馈和指导。

本文的爬虫需要对抓取的页面进行内容分析，因此属于聚焦爬虫。但是它有别于一般的网络爬虫，因为其抓取的对象不是一般网页，而是 Blog 站点的链接和文本。根据 Blog 站点的特点，可以利用 Rss 文件为线索进行抓取。RSS(Really Simple Syndication) 也叫聚合内容，是一种描述和同步网站内容的格式，是目前使用最广泛的 XML 应用，并且成为了描述 Blog 主题和更新信息的最基本方法，目前已经发展到了 RSS 2.0 版本。Rss 文件包含了 Blog 站点的主题、作者、最新文章及其评论的链接和日期等重要信息。而每个 Blog 站点都有自己的 Rss 文件，用来发布自己的主题、最近更新的文章和评论等内容。由于 Blog 社区会随时间发生变化，而我们根据 Rss 文件抓取的文章内容和 Blog 站点之间的联系都是最近，最新的，因此所抓取到的网络和文章必然是最新的。下图是本文中爬虫的抓取算法流程图。

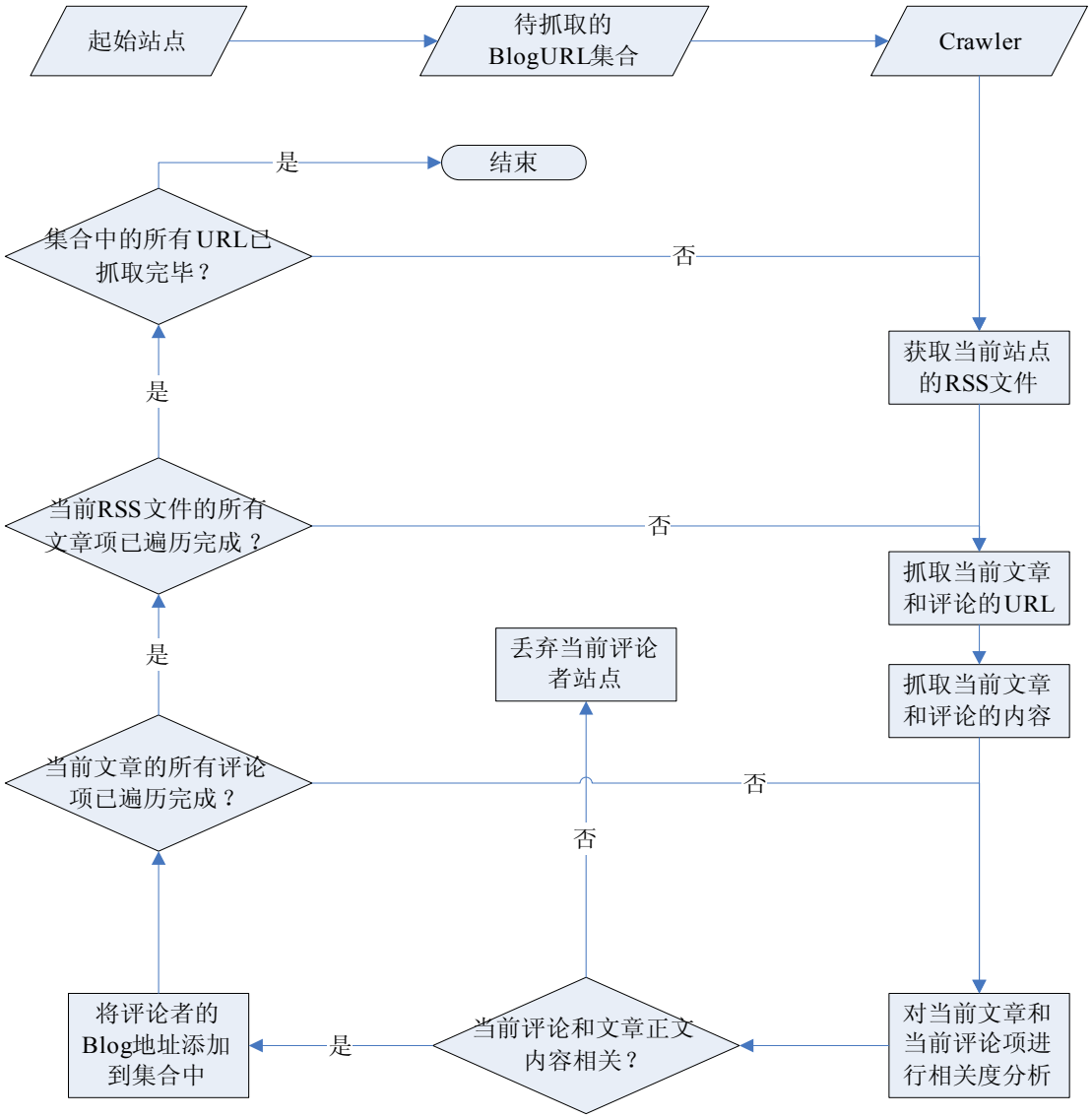

图 3-2 爬虫抓取算法

Fig. 3-2 Flow Diagram of Crawler

2．内容相关度分析

接下来，我们对 Blog 文章的结构和版面进行分析。通常，一篇 Blog 文章可以被分为 3 个部分：标题，正文和评论，如图 3-3 所示。通常的社区发现方法都是仅仅对链接的结构特点进行分析，结合图论的相关算法进行区分。而在本文中，我们要充分利用标题，正文以及评论的内容以及它们之间的相关度来动态筛选 Blog 节点。

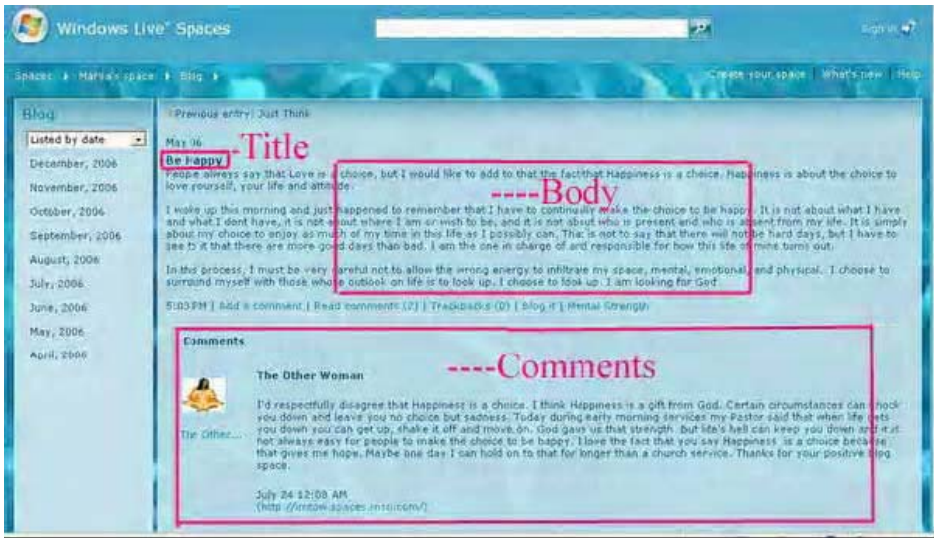

图 3-3 Blog 文章结构

Fig.3-3 Structure of Blog Document

为此，我们定义一些基本规则来过滤一些相关性不高的站点。

- 评论字数少于3个的站点，该评论者本次评论计数为零。我们认为，如果评论字数过少，则评论者极有可能是偶尔看到作者的文章，顺便留下了几个字的评论，或者说是一些常规性的无价值的留言。比如“沙发”，“顶”等等。评论者与作者之间的交流强度很弱，或者说内容相关性较低。
- 完全相同的评论内容和评论作者在同一篇文章的评论列表中重复出现3次以上的站点，该评论者本次评论计数为零。这种情况极有可能是评论者为吸引大众注意而进行的刷屏性质的恶意留言，或者是广告性留言。
- 同一个字或者词的出现字数占评论文本字数大于某一阈值的，该评论者本次评论计数为零。这种情况主要为了过滤用相同字词重复多次来增加评论量的评论者。
- 评论内容中如果出现人身攻击的恶意言词，或国家法律规定禁止在网络上发布的敏感性词语等，直接去除该评论者站点。这样可以保证挖掘出的社区是健康的、积极的社区。

利用以上这些规则，我们可以极大地缩小抓取的范围，从而更有针对性地抓取内容相关度更高，联系更紧密的Blog站点。

3．抓取结果

我们需要抓取 Blog 站点的链接结构和所有文本内容。

- 站点结构抓取

我们以微软的 Live Space 空间作为数据源来进行社区挖掘，仅考虑和分析其中的中文 Blog 站点。首先选取一个最近更新过，并且评论者相对较多的中文 Blog 站点作为爬虫的起点，逐步获取所有通过评论发生联系的 Blog 站点。根据六度分割理论，每两个人之间平均通过六个人就可以联系起来，为了避免爬虫无休止地抓取下去，我们规定爬虫的抓取深度为 2。即抓取起始站点，起始站点的评论者站点，对评论者站点发表评论的站点。当然，还包括这些站点之间的关系。初步抓取的结果是一个非连通图，用 xml 文件来存储。用来存储原始数据的 XML 文件如下图所示。

- 文本抓取

在抓取 Blog 网络结构的同时，同样以站点的 Rss 文件为线索，分别对所有站点的文章和评论进行抓取。根据 Blog 文章的结构特点，我们把抓取到的每篇文章分为 3 个部分：文章标题，文章内容和评论列表。

我们把抓取的站点结构和文本内容存储下来。为了较直观地表示Blog网络结构，我们定义一个XML文档格式来对抓取的Blog网络进行结构化存储。下图是一个抓取结果的示意图。该XML文件的根节点为<community>, 代表整个社区网络。它由多个<member>子节点构成, 每个<member>节点包括<author>, <url>, <sitename>, <level>, 和<commenter>子节点，各节点的含义如下：

<author>: 站点作者

<sitename>: 站点名称

<level>: 站点所处层次和深度，取值0，1，2

<url>: 站点的地址

<commenter>: 站点的评论站点地址。该节点还有一个count属性，用来存储该评论站点对当前站点发表评论的次数。

```
    <commenter count="3">http://majiao419.spaces.live.com</commenter>
    <commenter count="3">http://yuanhiq.spaces.live.com</commenter>
  </member>
- <member>
    <author />
    <sitename>异度空间</sitename>
    <level>1</level>
    <url>http://lovelyminlo.spaces.live.com/</url>
    <postnumber>23</postnumber>
    <commenter count="3">http://lovelyminlo.spaces.live.com</commenter>
    <commenter count="5">http://strawberrylr.spaces.live.com</commenter>
    <commenter count="6">http://youbearzhou.spaces.live.com</commenter>
    <commenter count="5">http://jamesxk1104.spaces.live.com</commenter>
    <commenter count="2">http://teikaika.spaces.live.com</commenter>
  </member>
- <member>
    <author />
    <sitename>燃烧的岁月</sitename>
    <level>1</level>
    <url>http://koengy.spaces.live.com/</url>
    <postnumber>39</postnumber>
    <commenter count="2">http://koengy.spaces.live.com</commenter>
    <commenter count="4">http://aopopa.spaces.live.com</commenter>
    <commenter count="3">http://zdz36.spaces.live.com</commenter>
    <commenter count="4">http://cymhaienng.spaces.live.com</commenter>
    <commenter count="1">http://binchengxi2006.spaces.live.com</commenter>
    <commenter count="1">http://groovebird.spaces.live.com</commenter>
  </member>
- <member>
    <author />
```

图 3-4 存储原始网络的 XML

Fig. 3-4 XML Sample of Raw Blog network

3.3.2 网络可视化及降噪

1．网络可视化

上节中我们把抓取的原始网络存储在了 xml 文件中，现在用网络分析分析工具对网络进行可视化。我们使用 Pajeck 来完成这项工作，Pajeck 是目前最流行的网络分析工具。

通过 Pajek 可完成以下工作：在一个网络中搜索类（组成，重要结点的邻居，核等）；析取属于同一类的结点，并分别地显示出来，或者反映出结点的连接关系（更具体的局域视角）；在类内收缩结点，并显示类之间的关系（全局视角）。除普通网络

（有向、无向、混合网络）外，Pajek 还支持多关系网络，2-mode 网络（二分（二值）图—网络由两类异质结点构成），以及暂时性网络（动态图—网络随时间演化）。Pajek 是专门用来分析大型网络（含有成百上千个结点）的专用程序，包含如下六种参数：

(1) Networks（网络）－主要对象（结点和边）。默认扩展名为.net。在输入文件中，

网络有多种表现方法：

- 利用弧线/边
- 利用弧线列表/边序列（如：1 2 3－从1 到2 的连线和从1 到3 的连线）
- 矩阵格式
- UCINET，GEDCOM，化学式

(2) Partitions（分类）－它指明了每个结点分别属于哪个类，默认扩展名为.clu。

(3) Permutations（排序）－将结点重新排列，默认扩展名.per。

(4) Clusters（类）－结点的子集（如：来自分类中的一个类）。默认扩展名.cls。

(5) Hierarchies（层次）－按层次关系排列的结点，例如：

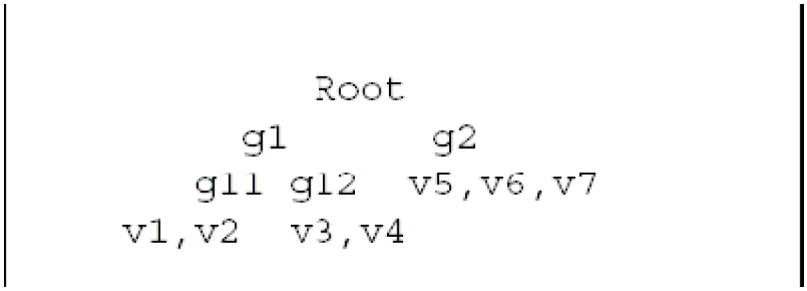

图 3-5 层次示意图

Fig. 3-5 Sample of Hierarchies

根结点 Root 下面有两个子群 g1 和 g2。其中 g2 是一个叶结点，包含 v5、v6、v7 三个结点。g1 又包含两个子群 g11 和 g12。默认扩展名.hie。

(6) Vectors（向量）－指明每个结点具有的数字属性（实数）。默认扩展名.vec。

2. 降噪

为了更好地进行主题分析，我们对抓取到的站点进行噪音过滤。去除活动频率较小以及对社区的主题贡献较小的节点。首先，去除只有出度而没有入度的点；其次，去除入度和出度均小于 2 的节点。这些 Blogger 在社区中的重要程度较低，或者在社区中的交互活动较少，可能会对主题的挖掘造成影响。

3.3.3 基于社会网络分析的 blog 社区发现

社会网络分析不仅是一种认识世界的方法，也是一种全新的研究世界的范式和技术。它将弥补甚至取代个体主义方法补充甚至超越主流的统计方法。下面分五个部分详细介绍社会网络分析中常用的网络变量及其在社区发现中的应用。

1. 入度(Indegree)

入度是社会网络中心度的一种形式，用来计算社会网络中其它行动者与特定行动者之间关系个数。在 Blog 网络中，入度一般用来提供其他成员阅读或引用某特定成员发布的讨论主题的人数信息。入度在 Blog 网络中的应用一般有以下几种情况：

- 入度值(Indegree Value) 某特定成员的入度值的高低表示该成员与其他成员的交互状况。通常用来描述特定成员被其他成员的认可和欢迎程度。成员的入度值高表明该成员的讨论主题的观点在 Blog 网络中具有很强的影响力。
- 平均入度(Average Indegree) 平均入度描述整个 Blog 网络的交互协作特征。Blog 网络具有高平均入度说明其成员发表的讨论主题之间相互引用的程度非常高，成员进行了较深的批判性思维，表明整个 Blog 网络进行了高质量的知识建构。
- 入度差异(Indegree Divergence) 入度差异用来描述成员对 Blog 网络的集体智慧的贡献的差异大小。表明知识架构网络中是否存在社会不公的现象。如果成员的入度差异大说明网络中部分成员发表的讨论主题特别受其他成员的欢迎，而另一部分成员发布的讨论主题被其他成员所忽略。这种情况表明在 Blog 网络中成员进行集体智慧发展的努力程度不均衡，成员对 Blog 网络的贡献存在较大差异，网络内部存在社会不公。Blog 网络的协调与促进人员如社区的知识管理者应该采取相应的措施消除这种情况。

2. 出度(Outdegree)

出度也是计算社会网络中心度的一种形式，研究角度与入度相反。它计算特定行动者与其他行动者发生交互关系个数。在 Blog 网络中出度一般用来表示特定成员阅读或引用其他成员发布的讨论主题的条数。在描述 Blog 网络中成员的交互模式方面出度与入度相类似，也是从三个方面进行表述的。

- 出度值(Outdegree) 出度值的高低也表示 Blog 网络中某个特定成员与其他成员的交互状况。与出度相反，它用来表明特定成员与其他成员交互的主动性和积极性。如高出度值表示特定成员通过阅读或引用其他成员的发布的讨论主题，主动积极地创建与其他成员的联系进行积极的集体智慧的发展。
- 平均出度(Average Outdegree) 平均出度与平均入度具有相同的功能，都是用

来描述整个 Blog 网络的交互协作特征的网络变量。

➢ 出度差异(Outdegree Divergence) 出度差异也是被用来描述 Blog 网络中成员对集体智慧发展贡献的差异大小，证明 Blog 网络中是否存在社会劳动不公的现象。与入度差异相比，出度差异更倾向于描述 Blog 网络中成员进行积极的自主学习的程度。Blog 网络的出度差异大说明部分成员进行知识建构的兴趣和动机存在问题。Blog 网络的协调与促进人员应该关注这些成员。

3. 中介度(Betweeness)

两个非邻接的行动者间的相互作用依赖于网络中的其他行动者。特别是位于两个行动者之间路径上的那些成员。它们对这两个非邻接行动者的相互作用具有某种控制和制约作用。这些行动者被称为中介者，所以中介性就是衡量中介者存在于其它任两个行动者路径上的重要程度。

在 Blog 网络中，中介度主要取决于信息流经过中介者的程度。如果信息流经常间接地经过某个特定成员再传播到其他的成员，那么该成员就具有了高中介性。如果 Blog 网络中具有较高的平均中介性，说明 Blog 网络中一定存在多个知识中介者。这些中介者在 Blog 网络中具有重要的社会位置，影响着 Blog 网络中的信息流动。这种情况是 Blog 网络所不希望的，违背了知识建构的民主性原则。因为 Blog 网络成员间需要具有直接的信息流动，而不是经过少量几个中介者然后由他们进行再传播。

4. 密度(Density)

密度是用来表示行动者的关系是否紧密。它用来测量社会网络中行动者之间的连结程度。密度越高，代表行为者之间的关系越紧密。密度是社会网络中实际联系的数目与所有可能的联系的数目的商。它的值在0和100%之间。因此在Blog网络中密度计算能反映了成员参与协商讨论的积极程度。网络密度的测量是根据有向图密度公式 $D(n_l)=\frac{l}{n(n-1)}$ 进行的。在Pajek中可以得到输入网络的密度，如下图所示：

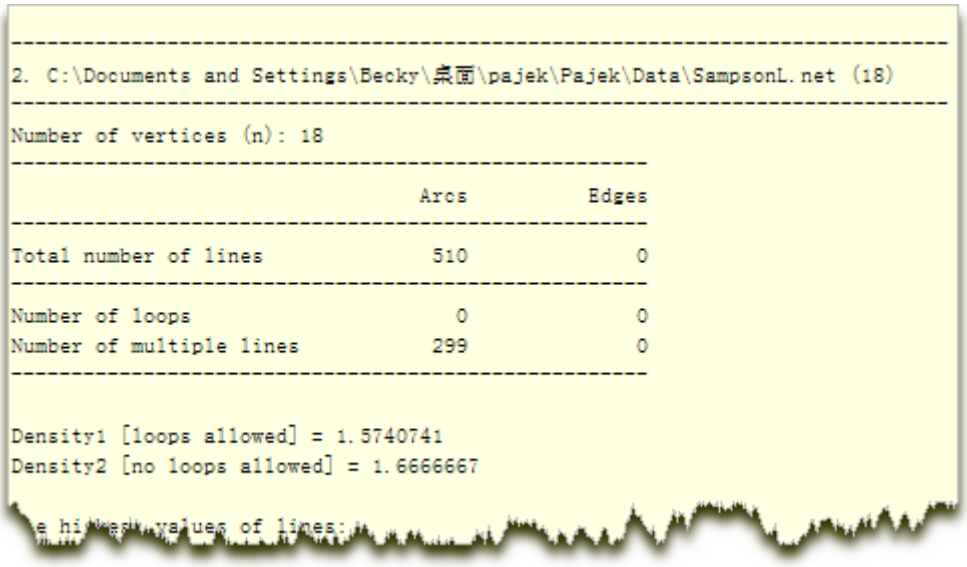

图 3-6 样本网络密度测量结果

Fig. 3-6 Destiny of Sample Network

这一结果显示，在所有个样本博客组成的网络中，实际存在的连接数为 510 条。该网络的密度为 1.5740741。以上的密度结果是基于整个网络进行的，考虑到网络规模对密度值的影响，又引入绝对密度公式 $D(n_l)=\frac{l}{4cr^3/3d}$ 对网络密度进行测量。Pajek 算出图直径为 4，测量过程见下节，据此算出网络的绝对密度为 18.4。这些数值显示，在整个网络范围之内点与点之间连接非常稀疏，样本博客之间没有普遍的、密切的交流关系.而网络绝对密度数值则显示，在网络连接相对紧密的最大关联图范围之内，样本博客的联系较密切。

5. 可达性(Reachability)

在Pajek中选择输入的网络模型，可以输出网络的直径。如下图所示：

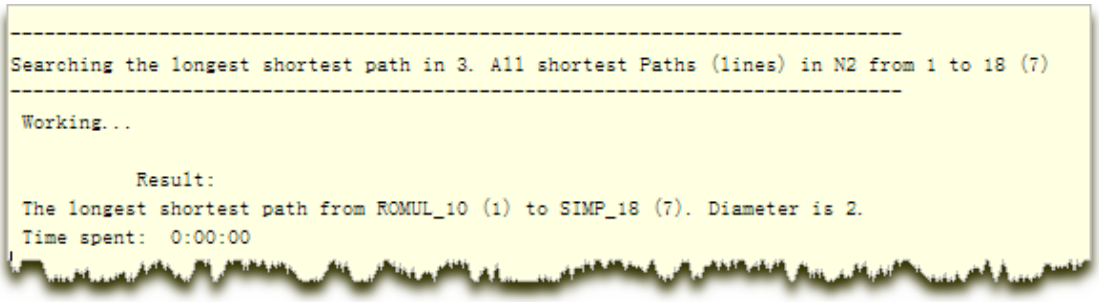

图 3-7 样本网络直径测量结果

Fig. 3-7 Diameter of Sample Network

对于一个有18个点、510条连接的有向网络来说，直径2为意味着最大关联图规模较小，只有很小一部分点实现了互联。下面的分析结果进一步证明了这一结论。

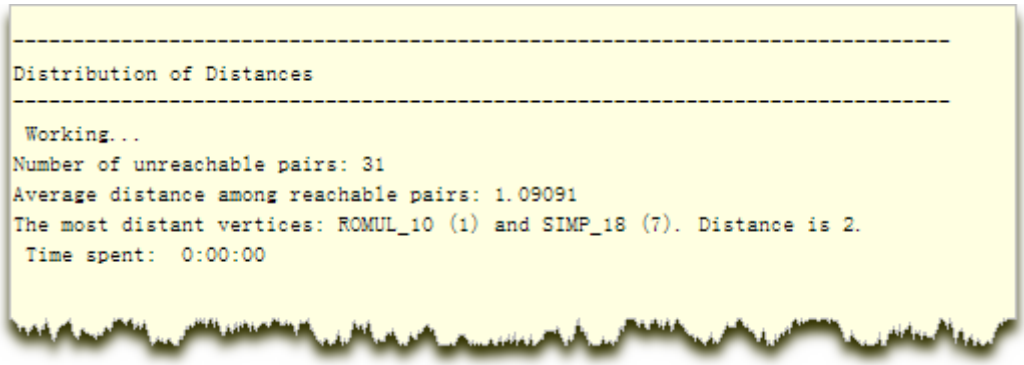

图 3-8 样本网络可达性测量结果

Fig. 3-8 Reachility of Sample Network

在考虑连接方向的前提下，全部样本博客中只有少数的点对能够到达对方，这说明网络的可达性非常微弱。

6. 凝聚力(Cohesion)

社会网络的凝聚力是通过派系 Cliques 的生成来证明的。在社会网络中，派系指的是至少包含三个行动者的最大完备的子群，是建立在行动者互惠关系基础上建立的凝聚子群。派系可以通过凝聚指数 Cohesion Index，C-Idx 来鉴定。凝聚指数是指派系内部的关系个数大于派系间的关系个数的程度。社会网络分析一般用凝聚指数 C-Idx，派系个数和派系中的成员个数来描述 Blog 网络的知识建构的品质。首先，凝

聚指数小说明 Blog 网络的不同派系的成员之间存在较多的互动关系，或者一个成员在不同派系承担多种角色。这样就有利于扩大整个 Blog 网络的信息流动。其次，派系的多少也能在一定程度上反映 Blog 网络中成员的互动关系。因为派系多说明成员之间的互动关系密集，将有利于知识建构。最后，单个派系的成员数多就意味着在以派系为单位的范围内成员交流的范围变大，也能在一定程度上促进员工之间的知识建构。

在 Blog 网络中，社会网络分析使用图论工具和代数模型描述网络的关系。使用网络变量描述成员的互动模式和网络特征，如使用入度和出度两个网络变量表达网络成员与其他成员的交互关系的紧密性。使用图论可视化工具形象地表达社会网络的关联性凝聚力交互的密度分布情况。一般而言，目前对于“社会网络”的使用有两种含义。一种是将网络作为一种分析工具指的是本文中社会网络分析；另一种则是将网络视为一种现实存在的实体，从而使之成为一种多学科关注的研究对象指的是本文中的社会网络。

7. 点度中心度(Centrality)

下图是计算点度中心度的例子：

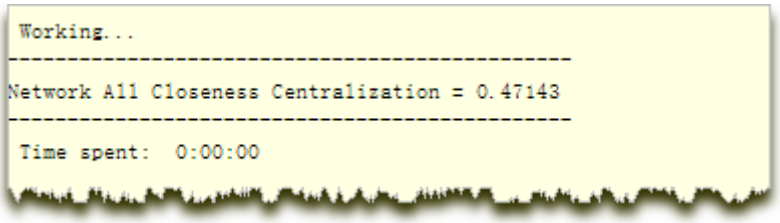

图 3-9 点度中心度测量结果

Fig. 3-9 Centrality of Sample Network

对这些中心度较强的点进行分析可以发现，它们大部分都处于同一聚类当中，即以一为起点的小圈子。它们的共同特点是外中心度远远大于内中心度，这说明它们之所以能够获得较大的点度中心度，并不是因为它们得到众多博客的追捧，而是因为它们自身向外连接了很多博客。这些点构成了一个个以他们本身为中心的星形网络，连接以外向、单向为主。可见，是过大的外中心度干扰了中心度的测量，造成了它们“位于网络中心”的表象。如果按照博客受追捧的程度即内中心度计算个体博客在局部的中心性，则点度中心度较高的博客都集中在以为起点的聚类当中。一个聚类中存在多个中心点博客，说明该聚类的互联性较强，各点的连接比较充分，而同时也说明，该聚类中不存在实际意义上的“中心点”，像这样内部充分连接的小群体在整个网络中

不具备代表性，是偏居博客网络一隅的特殊现象。

综上所述，根据社会网络分析的这些参数和理论，对已抓取的 Blog 网络进行分析，可以得到潜在的社区和核心站点等非常有价值的信息。

3.3.4 社区主题挖掘

1.主题词

Blog 社区是由站点和站点之间的联系构成的，而每个社区的文章和评论必然是围绕着一定的主题展开的。因此找到一种可行的方法挖掘出社区的主题是十分有必要，也是很有意义的。

主题一般是与某个事件，人物或者现象有关的话题，主要表现为词汇和短语等。比如：关于 2008 年奥运会的主题可以包括刘翔，“鸟巢”，北京，福娃等词汇。而与微软相关的主题则可以包括 Vista，比尔。盖茨，MSN 等等。所有这些主题词基本都是名词、代词和实体词，我们把这些可以表现文章主题的名词词汇和短语称为主题词，而把主题定义为一组相关主题词的集合。根据这个定义，我们对数据集中的所有 Blog 文章进行主题词抽取，每篇文章分为标题、正文和评论 3 个部分分别进行。

- 首先用天津海量科技公司的中文分词工具对所有目标文本进行分词，得到标记了词性的文本；
- 然后对分词结果进行筛选，仅保留普通名词、代词和专有名词。
- 最后，构造停用词表，去除停用词。我们把大多数文章都会出现的常用词收集起来，构成一个约包含200个中文词的停用词表。在筛选过程中去掉停用词表中的词可以提高主题分析的准确率和效率。

这样，每篇文章的标题，内容和评论都表示为一些主题词的集合。为了使挖掘的结果更有价值和意义，我们把描述提取的主题词进行分类。根据主题的规律和特点，我。把提取的主题词按词性划分为“人名”，“地名”，“时间词”，“其它专有名词”四类，再挖掘出各类别中的热门主题。

2. 主题挖掘

(a) 主题词频率

由于主题必然是社区成员讨论最频繁的话题，所以我们可以为每个集合中的主题词构造频率函数，表示该主题词在表示某一主题时出现的频率。以 TF-DF 方法为基础和启发，我们构造如下的频率函数：

- 主题词频率 $tf=\frac{t_x}{m}$，t_x是主题词t 出现的次数，m是抓取到的原始主题词总数
- 文本频率 $df=\frac{d_x}{n}$，d_x是包含主题词t 的文本数目，n是所有文本总数
- 主题词频率 $f=e^{tf+df}$

我们把主题词的频率用标题，正文和评论三个部分的频率加权组成的频率函数来描述：

$$F_t=w_tf_t+w_bf_b+w_cf_c \qquad (3\text{-}1)$$

f_t为标题频率，f_b为正文频率，f_c为评论频率，w_t，w_b，w_c3 个权重系数表示主题词频率在标题，正文和评论中的重要程度。

(b)文章时空分布

除了出现频率上的特征之外，热门主题还具有一些时间和空间上的特征，可以作为挖掘的依据。

- 时间特性。如果在一段的时间之内有较多与某个主题有关的文章连续发表，则可以

认为这个主题是热门主题。建立主题词的时间标度公式：

$$I_t=\frac{e^{n-\lambda}}{\frac{1}{n}(|p_{1t}-f_t|+|p_{2t}-f_t|+...+|p_{nt}-f_t|)+\varepsilon}\times\frac{n}{N} \qquad (3\text{-}2)$$

$$f_t=\frac{1}{n}(p_{1t}+p_{2t}+...+p_{nt}) \qquad (3\text{-}3)$$

其中，λ，ε为平滑参数，n表示包含主题词t的文章数目，N为所有文章数目，p_{it}表示第i篇文章的发表日期。该公式反映了文章发布的时间集中程度，同时还考虑了极端情况下的平滑。用它可以计算出关于某主题的文章发表频度。I_t的值越大，表示该主题的文章发表的时间越集中，而且发布的数量也较多。

- 空间特性。如果在某段时间内，较多的社区成员都发表了关于同一主题的文章，也可以认为该主题是热门主题。构造同主题文章的分布公式：

$$R_t = \frac{n_t}{s} \qquad (3\text{-}4)$$

s 为社区内站点总数，n_t 为发布了与主题 t 相关文章的站点数目。通过计算这个数值可以反映出主题 t 在社区中的被关注程度。

综合考虑以上两个特性并加上权重系数，我们得到主题词 t 的时空分布函数：

$$T_t = w_i I_t + w_r R_t \qquad (3\text{-}5)$$

(c)评论反馈

最后，我们根据文章评论的特征来挖掘主题。

- 评论回应间隔

我们认为，如果一篇文章发表后，大量评论是在较短时间间隔内做出回应的，则可以认为该文章相关的主题是大家所关注的话题。构造评论回应间隔公式：

$$I_t = \frac{1}{n}\sum_{i=1}^{n} e^{\frac{1}{(t_{ci}-t_p)+\varepsilon}} \times \frac{n}{N} \qquad (3\text{-}6)$$

其中，n 为与主题词 t 相关的评论数量，N 为所有评论数量，t_{ci} 为第 i 篇评论的发表时间，t_p 为主题文章的发表时间，ε 为平滑参数。I_t 值越大，表示评论的回应间隔越短。

- 评论长度

如果关于某个主题的评论平均长度超过一定值，说明成员对该主题比较关心，愿意花较多的时间来讨论。我们认为这样的主题也是话题。构造平均长度公式：

$$L_t = \lg(\frac{1}{n}\sum_{i=1}^{n} c_i) \times \frac{n}{N} \qquad (3\text{-}7)$$

n为与主题词t有关的所有评论数目，N 为所有评论数量，c_i 为第i 篇评论的字数。由此可以计算出每个主题词的平均评论长度。综合考虑评论反馈的2个特征并加上权重系数，同时考虑到评论的数量，我们得到评论特征函数：

$$C_t = (w_i I_t + w_l L_t) \times \frac{n}{N} \qquad (3\text{-}8)$$

至此，我们从 3 个方面来分析了社区主题的挖掘方法。把上面 3 个方面的因素综合起来，加上一定的权重，就可以对所有主题词的进行排序，得到最受关注的主题。

第四章 系统设计与实现

4.1 系统设计

4.1.1 系统设计目标

设计基于结构分析和内容分析相结合的 Blog 社区发现应用系统有重要的现实意义。因为 Blog 之间的连接是一种特殊的 Web 资源，起到连接 Blog 站点的作用，并隐含着被链接者之间的逻辑和语义等关系，同时也在一定程度上反映了链接者之间的社会联系。所以根据 Blog 页面之间的超链接关系，结合社会网络分析理论，按照设定的标准进行构建 Blog 社区，可以发现 Blog 世界里的重要资源和信息，促进 Blog 社区的更好发展。另外，加入对文章评论的内容相关度进行分析，能够对 Blog 信息集合进行有效的划分和直观易懂的描述，使最终得到的满足相关条件的站点数量将极大地减少，对不满足条件的站点进行有效过滤。在这个站点集合基础上可以发现和挖掘更准确，更稳定和更相关的 Blog 社区，从而有助于进一步挖掘社区的主题和其它重要信息，为我们提供更加丰富和准确的信息。

4.1.2 系统架构

通过学习和研究国内外 Blog 社区发现和挖掘技术，在参考链接结构分析技术，社会网络分析理论以及文本主题建模等技术的基础之上，本文设计并开发了基于社会网络分析和文本内容分析的 Blog 社区发现应用系统。该系统主要包括原始数据获取、基于社会网络分析的潜在社区发现和社区主题提取等部分，其体系结构如图 4-1 所示。

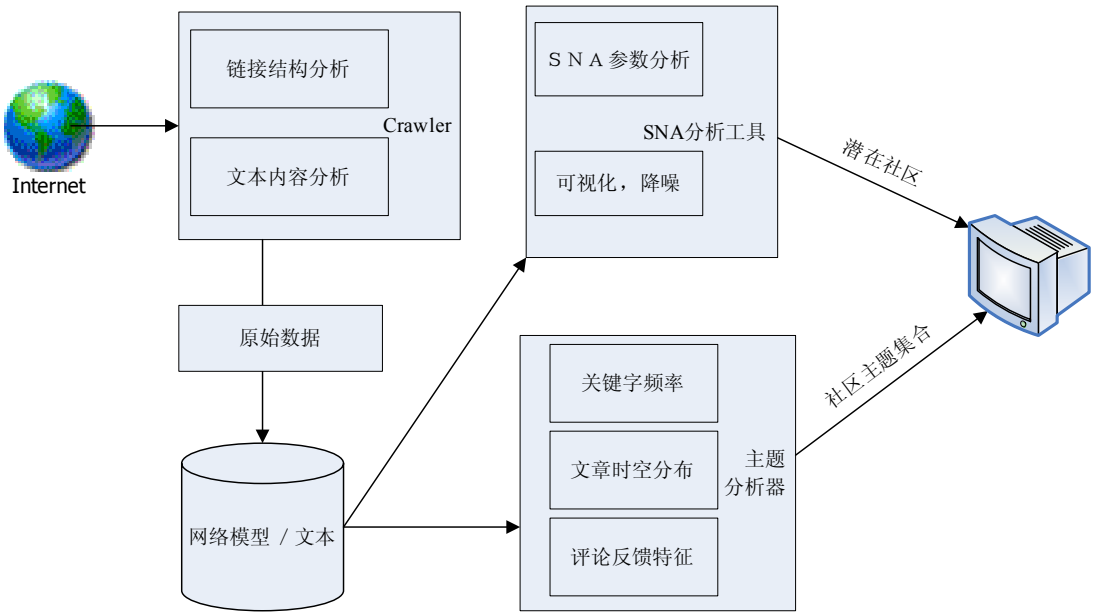

图 4-1 Blog 社区发现应用系统体系结构

Fig. 4-1 Architecture of Blog Community Discovering System

4.1.3 主要模块设计与分析

根据以上的系统结构图，Blog 社区发现系统主要包括以下 3 个模块

1. 原始数据采集与爬虫设计

由于因特网过于庞大，Blog 网站数目也数不胜数，再加上软硬件设备以及时间因素等客观条件的限制，不可能对所有 Blog 网站数据进行分析和研究，或者是所有特定类别的 Blog 网站数据进行研究。因此，我们必须首先获取要研究和分析的数据集。鉴于代表性和普及性，我们以微软的 Live Space 空间为数据来源，并精心选取一个比较热门，受关注的 Blog 站点为起点。然后设置一定条件，由爬虫在 Web 上抓取 Blog 页面的文本内容和链接关系。而对 Blog 原始数据的获取又是本系统中进行社会网络分析和文本内容分析算法的基础。图 4-2 是该模块的类图设计。其中，Crawler 是主要类，实现了站点抓取，内容相关度分析和所得网络的存储等核心功能。BlogSite 类描述了 Blog 站点，包括作者，地址，抓取文章数目，评论站点等详细信息。Commenter 类则描述了评论站点的信息，包括评论站点地址，评论次数等关键信息。而 RawNetwork 类则包含了抓取的整个网络的节点和关系。

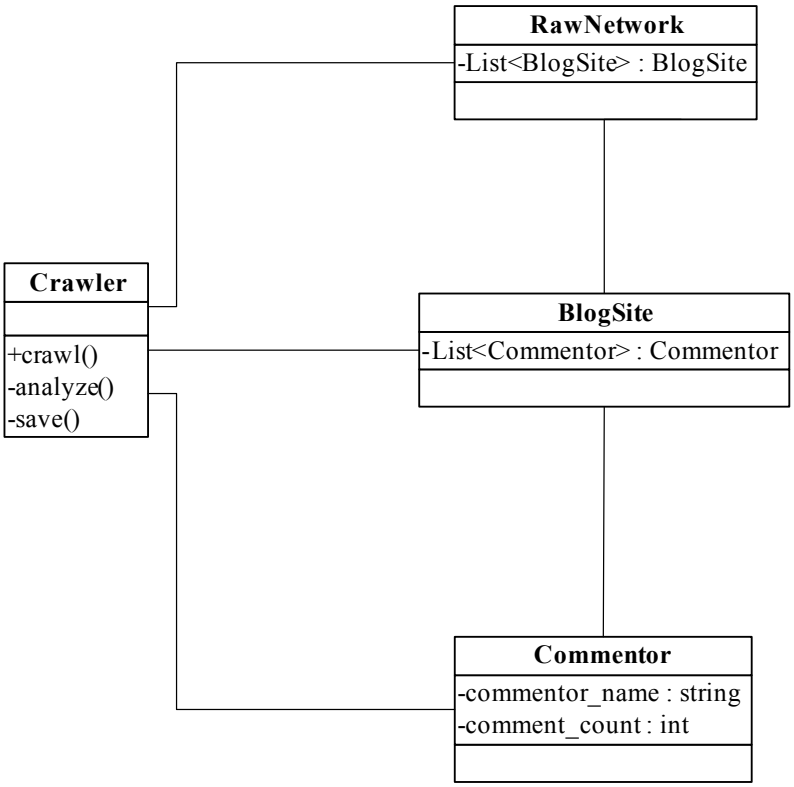

图 4-2 爬虫类图设计

Fig. 4-2 Class Diagram of Crawler

2. 网络建模与可视化

这部分主要完成把之前存储的 Xml 数据转换为 Pajek 能够识别的数据格式，然后用 Pajek 进行可视化，并选取相关参数对网络进行分析。Pajek 可以识别 6 种类型的文件，这里我们将原始数据转化为网络类文件的格式提供给 Pajek。

3. 主题分析器

根据第二步得到的结果，我们的主题分析器对潜在社区内 Blog 站点的文本进行内容分析。主题分析器的算法按照第三章的说明编写。图 4-3 是该模块的类图设计。其中，TopicMiner 是主要类，实现了分词接口的调用，主题词词频分析，文章时空分布特征分析，评论反馈特征分析等核心算法和操作。HSLCall 封装了对分词工具 dll 文件的调用，完成了从 C++到 C#的转换，并暴露出相关的接口。而 TopicWords 则描述了各类主题词按 3 种方法排序后的状态。

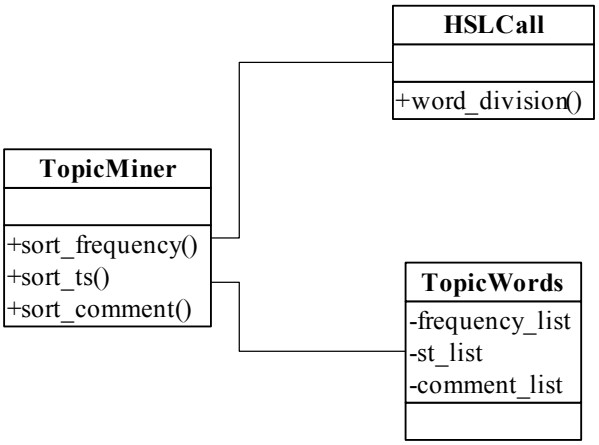

图 4-3 主题分析器类图

Fig. 4-3 Class Diagram of Topic Analyzer

4.2 系统实现

作为系统实现要考虑现有编程的相关技术，系统使用 C#作为开发语言，开发环境 IDE 选用了 Visual Studio 2005，Blog 网络可视化及社会网络分析工具 Pajek 和 Ucinet。在系统中，使用模块化设计、事件代理、面向接口等设计理念，同时使用了 Template Method、Observer 等设计模式，从而使系统具有良好的扩展性

4.2.1 原始网络采集

本文中抓取算法的核心就是首先找到 Blog 站点的 RSS 文件，然后以 RSS 文件为核心进行链接关系的搜寻和文本的抓取，由于我们的抓取目标限定在 Live Space 社区之内，因此我们可以用一个正则表达式来提取站点的 RSS 文件的地址：

(?<=application/rss\+xml.*?href="").*?.\(rss|rdf|feed|atom)

用这个正则表达式可以得到 rss，atom，feed 或 rdf 等各种规范和标准的 Feed 地址。下图是爬虫的顺序图实现。

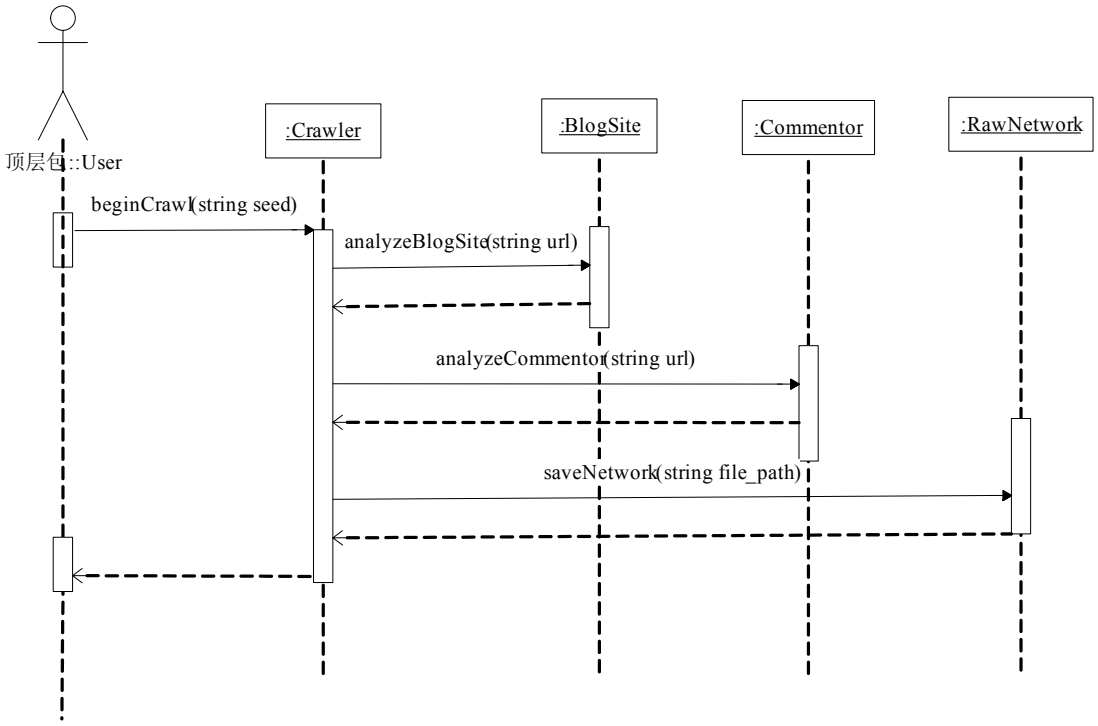

图 4-4 爬虫的顺序图实现

Fig. 4-4 Sequence Diagram of Blog Crawler

4.2.2 网络可视化与社区发现

在这个环节中，主要利用网络分析软件 Pajeck 来完成。我们首先将存储在 xml 文件中的网络节点和关系读取出来，转化为 Pajek 能够识别的数据格式。然后进行参数选择，利用 Pajek 进行分析。

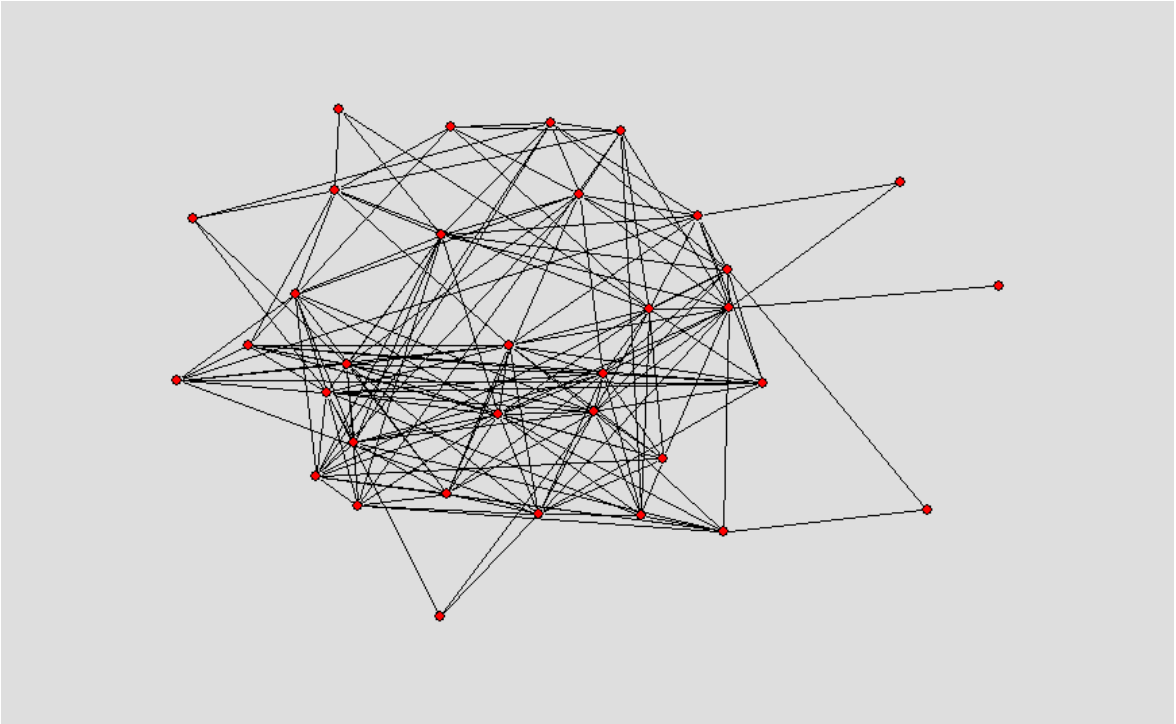

图 4-5　Pajek 样本网络可视化示意图

Fig. 4-5 Pajek Visualization

4.2.3 主题分析器实现

由于主题分析器是在分词工具分词的结果上进一步处理，挖掘出社区主题信息的。在本文中，我们使用天津海量科技的海量分词系统，可以比较准确地对中文文本分词。如下图是对一段中文的分词运算结果：

图 4-6　分词工具示例

Fig. 4-6 Demo of Word Dividing Tool

由于该分词工具是由 C++编写的，因此在用 C#实现系统的时候无法直接调用，而需要用到。Net 平台中的跨语言的调用机制 PInvok，进行 dll 导入和声明，并且封装在一个类中，然后暴露出相关的接口，方便系统的调用。在我们的方法中，我们关注的仅仅是名词，因而，我们把分词结果中的所有名词重新存储下来，然后进行分析。这样一方面在很大程度上减少了运算量，另一方面在结果的准确性上也可令人接受。下图是主题分析器的实现序列图。

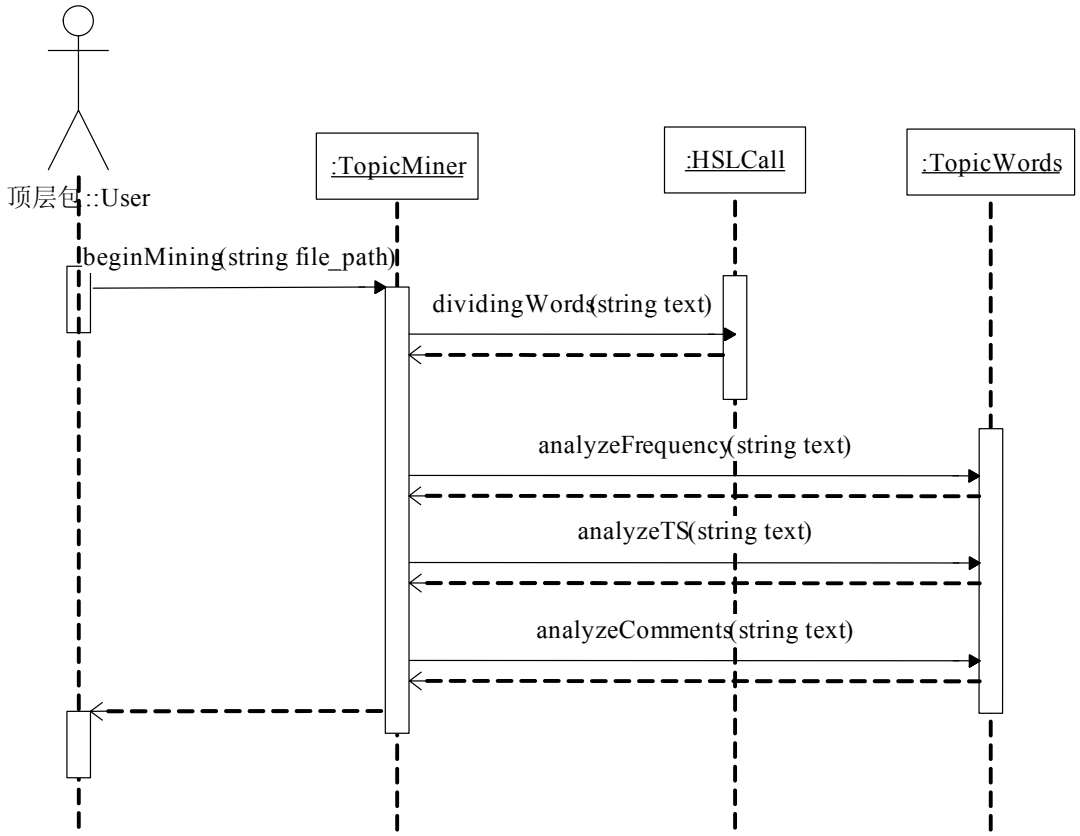

图 4-7 主题分析器顺序图实现

Fig. 4-7 Sequence Diagram of Topic Analyzer

第五章 实验与分析

5.1 实验设计

为了验证本文所研究的结构分析与内容分析相结合的分析方法的可行性，我们按照第三章介绍的总体框架和过程，设计了相关的实验。同样，对应本文社区发现方法的三个步骤，我们的试验也分为分三步进行：(1)抓取原始的 Blog 社会网络，包括结构和文本；(2)发现潜在社区及其结构分析；(3)分析和抽取社区主题。

可以知道，起始站点的选择将在很大程度上影响我们抓取的 Blog 网络，也在很大程度上决定了我们最后所发现的社区。比如，如果种子的入度太少，或发布的文章数量较少可能会造成抓取的网络规模过小等结果。因此，我们必须选取一个合适有效的起点。另外，为了更好，更直观地了解本文提出的基于内容相关度的 Blog 网络抓取方法的可行性，本文还作了一个对比实验。即从同一个起点出发，进行了两次网络抓取，一次使用了内容相关度算法，而另一次没有使用。下面将详细分析实验过程。

5.1.1 原始数据集的抓取

前面提到，为了得到一个较好的结果，我们必须精心挑选一个种子站点进行抓取，该站点应该至少满足以下两个方面的要求

- 文章发布量较大，更新频率高
- 文章的评论较多，好友链接较多

在本次实验中，选取了 http://atiger.spaces.live.com 作为起始站点。该站点已开创多年，作者发布文章频率较高，涉及内容也十分广泛。并且该站点访问率很高，大量文章都有较多的评论。从这几方面看，都符合条件，是一个比较好的种子站点。

由第四章介绍的爬虫算法，我们抓取到的原始数据集总共包括 485 个 Blog 站点，2007.9-2007.11 月期间的大约 5244 篇文章，以及 16514 条评论回复。下图是用 Pajek 对抓取结果的展现。

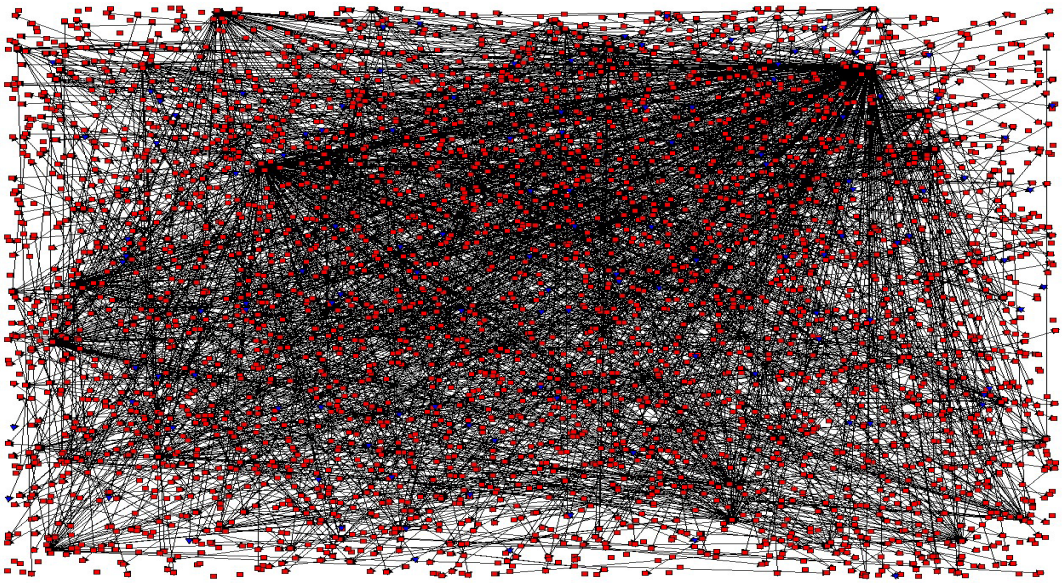

图 5-1　抓取的原始网络

Fig. 5-1 Raw Blog Network

另外，我们同样以该站点为起始站点进行了第二次抓取，不同的是在这次抓取中我们没有使用内容相关度的算法进行动态筛选和过滤。结果，采用内容相关度算法所抓取的站点要比全部抓取减少了近三分之一。而我们通过人工访问那些被去除的站点并查看其内容，发现这些被过滤掉的站点大多数都与我们抓取的社区联系较松散，入度一般都为 1 或 0。这说明我们用内容相关度算法进行原始网络抓取的效果较好，提高了性能，在很大程度上减少了数据处理量。下表是两次实验的结果对比。

表 5-1　对比实验结果

	结合内容分析的抓取	无内容分析的抓取
起始站点	http://atiger.spaces.live.com	http://atiger.spaces.live.com
抓取站点数目	485　个	1472　个
抓取文章数目	5244　篇	18451　篇
抓取评论数目	16514　条	57363　条
运行时间	23　小时	97　小时

5.1.2 潜在社区发现

接下来我们运用社会网络分析的理论对原始网络进行分析，找出其中潜在的Blog社区。我们主要利用度，可达性，密度和点度中心度等参数来进行分析。

5.1.3 社区主题挖掘

最后，我们对发现的社区进行文本分析，得到社区的主题集合。通过使用分词工具进行分析分词和初步处理，共抽取主题词 5856 个。在主题词频率方面，我们需要确定 3 个权重参数来得到主题词的频率得分。我们认为标题的概括性较强，故赋予较大的权值 0.5，而文章正文权重为 0.3，评论权重为 0.2。在时空分布上，我们认为时间分布因素和空间分布因素的比重相当，各赋予权重 0.5。

5.2 实验结果与分析

通过相关工具和社区发现系统的分析计算，得到社区结果如下：

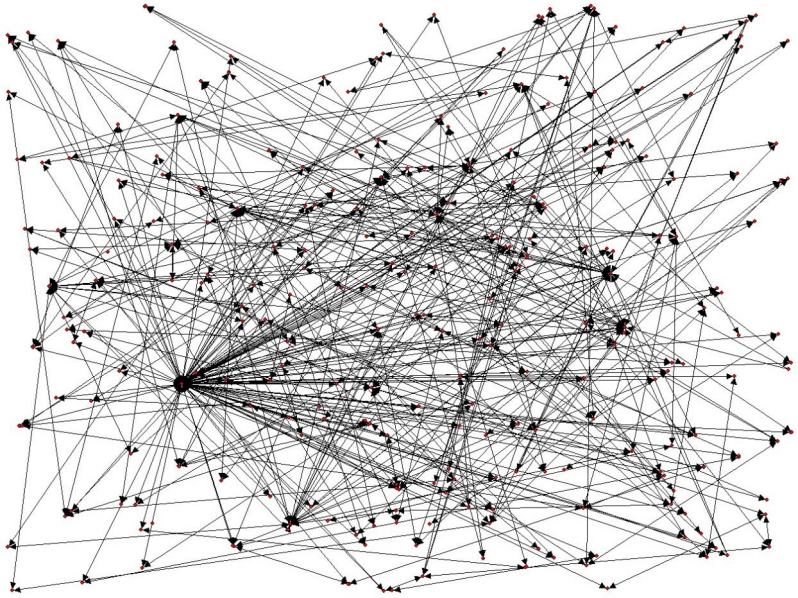

图 5-2 潜在社区

Fig. 5-2 Latent Blog Community

在以上的社区中，进行文本内容分析和主题挖掘，我们得到分值排在前几位的主题如下表所示：

表 5-2 社区主题

	人名	地名	时间	其它专名
主题词频率	孔子 0.03796 金庸 0.03774 苏轼 0.03725 白雪公主 0.03647 张小娴 0.03613	中国 0.07428 北京 0.07421 上海 0.07413 日本 0.05404 法国 0.05156	冬天 0.08312 中秋 0.05546 秋天 0.05235 新年 0.03529 夏天 0.02723	博客 0.02375 英语 0.01934 奥运 0.01924 佛教 0.01456 恐怖片 0.01103
时空分布	金庸 0.10694 王菲 0.09735 柏拉图 0.07546 周瑜 0.07438 三毛 0.06812	中国 0.39824 北京 0.38685 上海 0.38541 美国 0.25717 香港 0.25365	冬天 0.44581 周末 0.42072 秋天 0.41452 十一 0.40649 七夕 0.37713	博客 0.47455 英语 0.46727 QQ 0.42989 菩萨 0.40173 微软 0.39545
评论反馈	周杰伦 0.05731 赵敏 0.05494 董洁 0.05367 齐国力 0.05224 刘德华 0.05208	北京 0.16559 中国 0.16483 上海 0.16362 日本 0.15847 法国 0.12385	冬天 0.07232 中秋 0.06984 冬日 0.06804 除夕 0.06643 十一 0.06537	博客 0.43695 英语 0.41819 日语 0.40971 海归 0.38266 狮子座 0.37042

第六章 总结

6.1 结论

基于结构和内容分析的 Blog 社区发现主要研究如何结合对文章内容和评论的内容相关度分析，运用社会网络分析方法针对 Blog 页面间的链接关系对 Blog 网络进行有效的划分和归类。本文在发现社区的基础上进一步进行了信息挖掘，得到社区的主题。本论文的具体工作总结如下：

1. 当前 Blog 的发展及其数据特征的分析

随着 Blog 的不断发展及人们对 Blog 依赖程度的不断加深，Blog 数据资源也在多元化、复杂化发展。本文对当前 Blog 数据的复杂性及特征分布进行了较全面的分析，这是促进其它新兴学科与技术形成的必要前提。

2. 基于结构分析和内容分析的 Blog 社区发现算法研究

对 Blog 社区链接结构进行合理的描述和分析，是 Blog 社区发现研究的重要内容，一个很好的 Blog 社区发现算法有助于改善现有 Web 信息搜索结果。文中研究了基于链接分析的 Blog 社区发现典型算法，并提出了结合内容分析与结构分析的改进算法。

3. 基于统计分析的主题挖掘

在本文中，我们提出了一种挖掘 Blog 社区主题的方法。首先提取出主题词，然后对主题词进行聚类分析，充分利用文章的标题，正文和评论各方面的特征来计算主题词在各方面的特征，得到社区的最受关注的主题。本文提出的方法主要是基于统计学模型，结合一些 Blog 的行为特征和模式对 Blog 社区关注的问题进行分析和挖掘，下一步工作可以引入语义模型来提高准确性。

4. Blog 社区发现应用系统的构建

本文根据传统社区发现方法提出了改进算法，设计了具有实用价值的基于链接结构分析和文本内容分析的 Blog 社区发现应用系统，根据社区发现算法对无关 Blog 信息作了有效过滤，并对核心模块进行了开发和实现。

6.2 未来工作展望

Blog 社区发现技术仍是一个较新的发展领域，发现和分析 Blog 社区结构具有很重要的实际价值。本文的工作还存在一些问题和不足，需要对以下方面进行深入的研究与探讨：

1. Blog 社区结构统计特征度量方法的完善

因为 Web 拓扑结构除了一些稳定的统计特征以外，更细粒度拓扑结构特性目前还没有更好的度量方法，将链接信息和其他信息配合起来，进一步探讨具有实用价值的、更精确的社区结构及特性分析意义重大。

2. Blog 社区定义的明确和深化

现有的 Blog 社区定义仍然是比较初步和模糊的，没有一个统一的定义。由于没有统一明确的定义，也无法对发现的社区结果进行有效的验证和评判。因此下一步的工作需要为 Blog 社区给出一个形式化的定义，并探索验证社区发现结果的方法和技术。

3. 更加有效的内容分析方法

Blog 社区虽然是由超链接结构所决定，但它本身具备概念的特点。本文仅仅从统计学角度对文章和评论内容进行了粗略的分析，没有引入真正的语义模型，因此没有充分利用 Blog 的文章内容。如何从语义内容上组织 Blog 社区，如何将语义信息加入社区的发现方法，找出 Blog 网络中的语义社区，也是下一步将要研究的工作之一。

4. 社区发现算法的优化

Blog 社区是动态变化的，包括内容、链接结构等，所以增量计算算法研究变得尤为重要。因为大多算法采用的是离线算法，尽管在技术上可以考虑以前的计算结果，但对整个图形的迭代计算需一定的代价。因此，需要探讨在线的算法适应 Blog 社区的变化。

参考文献

[1] Qamra. A, Tseng. B and Chang. E. Y. 2006. Mining blog stories using community-based and temporal clustering. In Proceedings of the 15th ACM international Conference on information and Knowledge Management (Arlington, Virginia, USA, November 06 – 11, 2006).

[2] Glance. N, Hurst. M, Tomokiyo. T. BlogPulse: Automated Trend Discovery for Weblogs Presented at the Workshop on the Weblogging Ecosystem at the 13th International World Wide Web Conference(2004).

[3] Sekiguchi. Y, Kawashima. H, Okuda. H and Oku. M. 2006. Topic Detection from Blog Documents Using Users' Interests. In Proceedings of the 7th international Conference on Mobile Data Management (Mdm'06) - Volume 00 (May 10 – 12, 2006).

[4] R. Swan and D. Jensen. Timemines: Constructing timelines with statistical models of word usage. In KDD-2000 Workshop on Text Mining, August 2000.

[5] Li. B, Xu. S and Zhang. J. 2007. Enhancing clustering blog documents by utilizing author/reader comments. In Proceedings of the 45th Annual Southeast Regional Conference (Winston-Salem，North Carolina，March 23 – 24, 2007).

[6] Shen. D, Sun. J, Yang. Q and Chen. Z. 2006. Latent Friend Mining from Blog Data. In Proceedings of the Sixth international Conference on Data Mining (December 18 – 22, 2006)

[7] Teng. C and Chen. H. 2006. Detection of Bloggers' Interests: Using Textual, Temporal, and Interactive Features. In Proceedings of the 2006 IEEE/WIC/ACM international Conference on Web intelligence (December 18 – 22, 2006)

[8] Lin. Yu-Ru, Sundaram, Hari. Chi, Yun. Tatemura, Jun. Tseng, Belle. Splog Detection using Content，Time and Link Structures. 2007 IEEE International Conference on Volume, Issue, 2-5 Page(s):2030 – 2033. July 2007

[9] Chin. A and Chignell. M. 2006. A social hypertext model for finding community in blogs. In Proceedings of the Seventeenth Conference on Hypertext and Hypermedia (Odense, Denmark, August 22 – 25, 2006).

[10] Mei. Q, Liu. C, Su. H and Zhai. C. 2006. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In Proceedings of the 15th international Conference on World Wide Web (Edinburgh, Scotland, May 23 – 26, 2006).

[11] Chi. Y, Tseng. B. L. and Tatemura. J. 2006. Eigen-trend: trend analysis in the blogosphere based on singular value decompositions. In Proceedings of the 15th ACM international Conference on information and Knowledge Management (Arlington, Virginia, USA, November 06 – 11, 2006).

[12] 刁倩，张惠惠，王永成，何骥. 中文文献自动分类中的知识库构造及其仿人算法. 情报学报 Vol. 19，No. 3. June 2000.

[13] Efimova, L. et al. Finding ``the life between buildings'': An approach for defining a weblog community, AOIR Internet Research 6.0: Internet Generations, Chicago, 2005.
[14] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. J. Mach. Learn. Res. , 3:993–1022, 2003.
[15] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In WWW, pages 491–501, 2004.
[16] S. C. Herring, L. A. Scheidt, S. Bonus, and E. Wright. Bridging the gap: A genre analysis of weblogs. hicss, 04, 2004.
[17] T. Hofmann. Probabilistic latent semantic analysis. In UAI, 1999.
[18] D. A. Huffaker and S. L. Calvert. Gender, identity, and language use in teenage blogs. Journal of Computer-Mediated Communication, 10(2), 2005.
[19] K. Ishida. Extracting latent weblog communities: A partitioning algorithm for bipartite graphs. In WWW2005，2nd Annual Workshop on the Weblogging Ecosystem, 2005.
[20] A. McCallum, A. Corrada-Emmanuel, and X. Wang. Topic and role discovery in social networks. In IJCAI, 2005.
[21] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In WWW, 2006.
[22] Kumar, R. et al. Trawling the Web for emerging cyber communities. Computer Networks. Amsterdam, Netherlands, 1999.
[23] D. Shen, J. -T. Sun, Q. Yang，and Z. Chen. Building bridges for web query classification. In SIGIR, 2006.
[24] Aschenbrenner. A., and Miksch. S. Blog mining in a corporate environment, Technical Report ASGAARD-TR-2005-11, Smart Agent Technologies, 2005.
[25] Hoyt C. Mining the blogosphere，the HUB Magazine，January 10，2006，
[26] Malkin. M. All about the Minnesota school shooter, March23, 2005, http://michellemalkin. com/archives/001837. htm，last accessed on November 1, 2006.
[27] Nicolov. N., Salvetti. F., Liberman. M., and Martin. J. H. Computational approaches to analyzing weblogs. In Papers from 2006 AAAI Spring Symposium, 2006.
[28] Nardi, B.A. et al. Why we blog. In Communications of the ACM, Vol. 47, No. 12, 2004, 41–46
[29] Torio.J. Blogs, A Global Conversation，Master's Thesis，Syracuse University，2005
[30] Lada A. Adamic and Natalie Glance. The political blogosphere and the 2004 u.s. election: Divided they blog. Proceedings of KDD Workshop on Link Analysis and Group Detection LinkKDD, 2005.
[31] Tyler, J. R. et. al. E-Mail as Spectroscopy: Automated Discovery of Community Structure within Organizations. The Information Society, Vol. 21, No. 2, 2005, 143 – 153.
[32] Flake, G. et al. Self-Organization and Identification of Web Communities. IEEE Computer, Vol. 35, No. 3 (2002), 66-71.
[33] 周涛，张际平 WiKi 社群的社会网络分析. 华东师范大学硕士学位论文 2005.

致　　谢

历时 10 余月终于完成了学位论文，本人的学生生涯也将随之结束，踏上工作岗位。短短几十页，洋洋数万字，凝聚了自己不懈的努力和执着的追求，也包含了几位老师的谆谆教诲和良苦用心。回首两年半的硕士生活，有学习的艰辛、项目的苦与乐、实习的付出与收获，还有生活中的欢笑和泪水……所有这些点点滴滴都将随着这本论文缓缓驶入记忆的长河，成为一段美好的珍藏。在这个收获的季节里，在硕士生涯的最后阶段，我要对我研究生期间的老师、同学和朋友深情地说一声：谢谢你们！

感谢我的导师吴刚老师。他不但是我学习上的严师，也是生活中的益友。他身先士卒，言传身教，以自己的行动感染和教导着我们。他以严谨踏实，一丝不苟的态度指导我的学习和研究，在他的严格要求和耐心指导下，我的理论知识和实践能力都得到了极大的提高。同时他还给了我一个自由的空间，让我能够充分发挥我的能力，教给了我很多做人的道理，使我受益终生。

感谢实验室的姜丽红老师和蔡鸿明老师。他们在我做实验室的项目过程中，给予我细心指导，帮助我顺利的完成实验室的科研项目；同时在我论文方面也给予我无微不至的关怀。在此，我要对这两位老师表示由衷的感谢。

最后，感谢读研期间的同学们：肖伟、宗裕朋、张辰、朱艳、邵彬彬、莫振华等。期间所有同学们在课业和研究上的切磋琢磨、相互鼓励、浓情厚谊永志不忘，本文的完成，也离不开他们无私的帮助和支持。愿他们踏出校门之后，前程似锦，大展宏图。

感谢每一位关心、帮助、支持我的人！感谢软件学院！感谢上海交通大学！

无论何时，无论何地，我都将铭记校训：引水思源，爱国荣校！

攻读硕士学位期间已发表或录用的论文

[1] 张浩，吴刚. Blog 社区热点主题挖掘. 《计算机工程》增刊（已录用）