哈 尔 滨 工 业 大 学

硕士学位论文开题报告

题 目：高校教师社会网络的自动构建

院　　（系）　计算机

学　　科　计算机科学与技术

导　　师　王晓龙

研　究　生　王长伟

学　　号　**09S003037**

开题报告日期　**2010.09.06**

研究生院培养处制

1、课题背景

随着计算机的普及以及互联网的迅猛发展，大量的信息以电子文档的形式出现在人们面前。为了应对信息爆炸带来的严重挑战，研究人员提出了多种技术帮助人们在海量信息源中迅速找到真正需要的信息。正是基于这种背景，本课题就信息的快速获取、动态更新、社会网络挖掘和社区发现、个性化服务等技术进行研究，以企自动化构建高校计算机教师的社会网络，通过整合各种信息为相关用户更好的服务。下面是课题背景的详细介绍

1.1 人物搜索技术

随着互联网对人们生活的渗透，个人信息也越来越多地出现在网上。此外，随着信息服务业的发展，个体和机构对信息的需求进入更加精细化的阶段[1]。在互联网应用走向多元化的今天，搜索引擎的搜索功能日益完善，新闻，图片、音乐、游戏等搜索逐渐成为必备功能，而互动问答、购物、影视、行业、学术搜索逐渐成为新的焦点，然而人物搜索作为新生事物，正处于兴起的阶段。两年前、市场上还没有成熟完善的人物信息方面的搜索引擎，人物要查找人物的信息时，只能借助于通用的搜索引擎。直到 2008 年，才有成型的人物搜索引擎的问世[2]。

英文人物搜索引擎比较成熟的产品有 Spock[1]、雅虎人物搜索[2]、USA search[3] 等。Spock 可以搜索出人物在互联网上公开的个人信息。利用雅虎人物搜索引擎可以搜索出人物的电话、住址、email。USA search 可以搜索出人物的年龄、生日、城市、电话等。

中文人物搜索引擎比价成熟的产品有优库[4]、中文雅虎人物搜索引擎[5]和中国人同学录[6]等。优库是收费的中文人物搜索引擎、从网络资源中收集人物信息，包括姓名、性别、生日、身高、体重、电话、Email 等个人基本信息。中文雅虎人物搜索的特色是对人与人之间的关系的抽取，侧重于知名人士的娱乐性内容。

1.2 社交网络服务

社交网络服务（SNS，Social Network Service）主要作用是为一群拥有相同兴趣与活动的人创建在线社区。早期社交网络的服务网站呈现为在线社区的形式，用户通过聊天室进行交流。互联网的发展尤其是 Web2.0 技术为社交网络带来了重大改变，Blog 等新的网上交际工具的出现使用户可以通过网站上创建的个人主页来分享信息。更为成功的代表则是验证了“六度分隔理论”的社交服务网站，如 Facebook[7]、Myspace[8]，以及国内的人人网 [9]、开心网 [10] 等。

社交网络为信息的交流与分享提供了新的途径。多数社交网络会提供让用户交互起来的聊天、寄信、文件分享、博客和讨论群组等方式途径。社交网络网站以用户的真实社会关系为基础，加强了用户之间彼此的交流沟通，并提供了各种新颖的应用，吸引了大量用户的参与。以开心网为例，截止 2010 年 6 月注册用户达到 8500 万，日活跃用户

---

[1] http://www.spock.com/
[2] http://people.yahoo.com/
[3] http://www.usa-people-search.com/
[4] http://www.ucloo.com/
[5] http://.people.cn.yahoo.com/
[6] http://www.chinaren.com/
[7] http://www.facebook.com/
[8] http://www.myspace.com/
[9] http://www.renren.com/
[10] http://www.kaixin001.com

约为 2000 万至 2500 万。

1.3 互联网个性化服务

所谓个性化信息服务是指能够满足用户的个体信息需求的一种服务，即用户可以按照自己的目的和需求，在某一特定的网上功能和服务方式中，自己设定网上信息的来源方式、表现形式、特定网上功能及其他的网上服务方式等，或通过对用户个性、使用习惯的分析而主动地向用户提供其可能需要的信息服务。

互联网上的信息总量很大，而每一个用户只是对其中的一小部分感兴趣，不同用户之间有着不同的兴趣爱好，不同的信息需求。尽管搜索引擎能够帮助用户对互联网上的信息进行查找，但是其不能针对用户的不同兴趣爱好、不同的信息需求提供不同的服务，检索结果中仍然包含对用户来说没有用处的内容。网络用户的信息需求与现有的信息查询技术之间的矛盾日益尖锐。而互联网上针对不同用户的个性化服务必是信息爆炸时代的必然结果。

早期的个性化 web 信息服务主要由新闻裁剪、股票报价和目录推荐等内容组成，这些内容是通过浏览器的 cookie 文件和一些信息信息服务商提供的智能后端工作实现的。当前流行的个性化信息服务则在社交网络服务网站中得到了很好的体现。通过提供兴趣小组和信息推送等服务，使用户能够定制自己需要的信息。

2、课题研究的目的和意义

本课题旨在通过运用自然语言处理相关技术，针对高校计算机领域教师，自动化的构建出一个社会网络为用户服务。整个过程的实现包括教师个人信息的抽取和自动更新、教师研究信息的获取和更新、领域专家的跟踪以及个性化信息服务等。

该社会网络系统自动采集教师信息对其进行一个比较全面的展示，包括个人信息和研究信息，通过信息整合等技术以更加直观的方式展现。这将极大地方便那些希望了解某教师信息的用户，能够提供全方位、多角度、垂直化的服务。同时，对于教师而言，作为该社会网络的直接参与用户，不需要自己填写更多的信息即可以得到个性化服务，包括兴趣小组、研究信息推送等。因此，该课题是信息展示和服务的一种新的突破，必将为领域用户带来更好的体验。

3、国内外研究现状和前景

3.1 信息抽取

信息抽取研究旨在为人们提供更有力的信息获取工具，以应对信息爆炸带来的严重挑战。与信息检索不同，信息抽取直接从自然语言文本中抽取事实信息[3]。信息抽取系统的主要功能是从文本中抽取出特定的事实信息。比如，从新闻报道中抽取出恐怖事件的详细情况：时间、地点、袭击目标等；从经济新闻中抽取出公司发布新产品的情况：公司名、产品名、发布时间、产品性能等；从病人的医疗记录中抽取出症状、诊断记录、检验结果、处方等。通常，被抽取出的信息以结构化的形式描述，可以直接存入数据库中，供用户查询以及进一步分析利用。

从自然语言文本中获取结构化信息的研究最早开始于 20 世纪 60 年代中期，这被看做是信息抽取技术的初始研究，它以两个长期的、研究性的自然语言处理项目为代表[4]。20 世纪 80 年代末，得益于消息理解系列会议(MUC, Message Understanding Conference)

的召开，信息抽取研究蓬勃开展起来。正是在 MUC 系列会议的的推动下，信息抽取发展成为自然语言处理领域的一个重要分支。MUC 定义的信息抽取任务的各种规范、确立的评价体系已经成为信息抽取研究事实上的标准。

此后，一系列信息抽取领域的国际测评会议的召开，如多语种实体评价任务(Multilingual Entity Task Evaluation , MET)、自动内容抽取(Automatic Content Extraction, ACE)、文本理解会议(Document understanding Conferences , DUC)等，也对信息抽取技术的发展起到了很大的推动作用。

早期的信息抽取多是基于规则来做[5]。尽管正则表达式和基于规则的信息抽取系统不需要任何训练，但是此方法只能限定于特定领域，并且需要专家来制定抽取规则，实现起来费时费力且不具备通用性。得益于自然语言处理技术的发展，很多机器学习方法在信息抽取领域中得到应用。例如支持向量机(SVM)[6-8],隐马尔科夫模型(HMM)[9]，条件随机域(CRF)[10,11]，和马尔科夫逻辑网络(Markov Logic Network)[12]。

研究人员发现，如果将多种不同的学习模型组合成一个集成系统，则系统的性能经常会得到明显改善[13],当前对各种集成技术（如 boosting, bagging, 和 stacking 等）的研究非常热门[14,15]。基本思想是所有的学习模型都在某些方面有所偏置，而通过对多个不同的模型的平均，可以有效地消除这些偏置。尽管统计机器学习方法在信息抽取中比基于规则的学习方法具有优越性，但是却需要大量标注好的语料的支持，而创建新的标注语料库资源是十分高代价的。因此一部分研究人员开始关注如何从现存的小规模已标注语料通过自扩展(bootstrapping)方法生成大规模的标注语料库[16]。

于琨等提出了双层级联信息抽取模型用于从简历中提取个人信息[17]。该方法将建立文本分解为文本块，并将简历中包含的信息分解为概要信息与详细信息。首先对简历中的文本块进行切分与分类，抽取出概要信息，然后选择可能包含详细信息的文本块，将其切分为文本串，再通过文本串的分类抽取出详细信息。Kristjansson 等开发了一个交互式信息抽取系统帮助用户从邮件中建立联系方式数据库[18]。唐杰等提出了一种提取研究人员信息的统一方法[19]，首先将页面分割成文本串并赋予所有可能的标签，然后应用 CRF 模型来确定最终的信息类型。

3.2 重名区分

重名区分(name disambiguation)可以分为以下两类情况：

1) 同一个实体或人被标以不同的名字标记。例如，作者名“David S.Johnson”在不同的发表刊物中以不同的名字出现，包括“David Johnson”、“D.Johnson”或者“D.S.Johnson”。
2) 多个实体或人被标以同一个名字标记。例如现实生活中的很多人的名字一样。

对中文数据来说，不存在同一个人的名字因为缩写原因而造成的重名区分问题，而一人具备多个姓名或者误拼写造成的第一类情况是非常少的，因此中文处理更多的集中于重名区分的第二类情况。对于本课题而言，则是在收集某一个教师的论文信息时，可能因为重名原因把不属于该教师的论文包括进来。去除因为重名带来的数据杂质，提高系统的精确度，是一个非常迫切的问题。

名字不确定对于数据整合带来了很大影响，降低了信息检索的准确度。很多科研人员在这方面进行了研究。重名区分是身份模糊(identity uncertainty)的一个特例[20]。针对身份模糊问题使用不同的方法在多个领域做了大量研究，例如记录连接(record linkage)[21]，重复记录检测和消除(duplicate record detection and elimination)[22,23,24]，引用匹配（citation matching）[25]等。

重名消岐是真实世界信息整合和数据管理的一个研究热点。一些工作只是基于基本

的引用数据进行作者重名区分[26,27]。Han[26]提出了两种基于有监督学习的方法，分别是使用朴素贝叶斯推理模型和支持向量机模型(SVM)。Han[27]使用了 k-way 谱聚类方法作者名消岐，对论文引用的每个属性（合作者名，标题，刊物名）的作用进行了评价，同时考虑了如数据集大小、特征权重等特征的影响。

除了使用论文引用的基本元数据外，其他研究也尝试从互联网上获取额外信息协助重名消岐。Kang[28]利用合作者网进行重名区分，要进行消岐的作者的合作者通过基于网络的技术获得。姓名对被作为关键词提交到网络搜索引擎，返回的结果中同时包含两个作者名，而返回文档中新出现的作者名则作为原始姓名对的合作者。Tan[29]也通过搜索引擎来进行重名消岐，输入引用信息进行查询，然后利用返回结果的链接信息作为特征进行重名区分，因为不同的论文引用查询得到的结果如果包含相同的链接信息表明它们可能来自同一作者。

本课题要从网络中获取某一教师的论文信息，而由于重名等原因，得到的结果可能包含一些非该教师的论文。如何从结果中把这些杂质去除，与重名消岐问题具有一定得相似性，可以借鉴相关方法进行处理。

3.3 社会网络构建及社区发现

社会网络是指社会个体成员之间因为互动而形成的相对稳定的关系体系，社会网络关注的是人们之间的互动和联系,社会互动会影响人们的社会行为。现实世界中，网络关系从宏观到微观无处不在，人类社会中的任何交往活动以及各种有形无形的社会关系本身也是各类不同层次的、交叉重叠的社会网络。

一个社会网络是由多个点(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。社会网络中所说的“点”是各个社会行动者，而社会网络中的“边”指的是行动者之间的各种社会关系。关系可以是有向的，也可以是无向的。同时，社会关系可以表现为多种形式，如人与人之间的朋友关系、上下级关系、科研合作关系等，组织成员之间的沟通关系，国家之间的贸易关系等。社会网络分析(Social Network Analysis，SNA)就是要对社会网络中行为者之间的关系进行量化研究，是社会网络理论中的一个具体工具。以关系作为基本分析单位的社会网络分析方法今天已经在社会学、教育学、经济学等诸多领域得到了广泛的应用，被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。

社区发现目前是复杂网络领域研究的一个热点问题，虽然有很多方法，但是还存在一些未解决的基本问题，如社区概念虽然大量使用，但却缺少严格的数学定义；大多数社区发现算法虽然性能优越，但所需计算量却很大。对于一般网络发现算法来说，其基本要求是可发现网络的最自然的分割，无需事先指定要分解成多少个社区，同时不存在如层次聚类算法那样的缺陷。Girvan 和 Newman 所提出的 GN 算法部分地达到上述要求，该方法的思路如下：社区之间存在的少数几个连接应该是是社区间通信时通信流量必经之路。如果考虑网络中某种形式的通信并且找到具有最高通信流量的边，该边就应该是连接不同社区的通道。 将这样的边去除就可以获得网络最自然的分解。

社会网络在合著网络方面的研究成果比较多。在这些研究成果中，作者往往根据各自的研究目的选取一定范围的学者，或者是某一数据库中有关一个特定主题的所有文献的作者，或者是一个专业数据库中所有文献的作者。在构建社会网络时，将这些作者作为网络中的节点,而作者之间的合著关系则构成作者之间的连线，然后应用社会网络分析法对合著网络进行分析。Liu Xiaoming 等人选择对 ACM 和 IEEE 各自的数字图书馆会议以及 ACM / IEEE 联合召开的数字图书馆会议 JCDL (Joint Conference on Digital Library) 的文献进行研究。他们首先根据社会网络理论建立起作者之间的合著网络，然后应用社会网络分析法中的中心性分析研究该合著网络。同时还在 PageRank 方法的基础

上提出了 AuthorRank 方法，并应用该方法对合著网络进行分析。最后对两种方法的分析结果进行了一定的比较[30]。

4、课题研究内容

高校教师网目前立足于海天园知识服务网站 IT 部落，旨在利用机器学习、数据挖掘等自然语言处理相关研究知识，自动化的构建一个社会化网络。该系统的基本用户群体定位为高校教师（目前限于计算机领域教师），但是向全体互联网用户提供有关教师个人信息、教师研究信息、教师活动信息等内容，实现一个有关高校教师的更直接、高集成、全方位、多角度的信息展示平台，并在此基础上打造一个大量科研人员参与的学术活动平台。

系统前期主要任务为自动化信息获取和全方位信息展示。自动收集教师个人信息、教师研究信息（主要论文）以及互联网能够采集到的有关教师的新闻、博客等信息。在这些信息的基础上进行相关研究，包括信息的整合和自动更新、研究论文收集整理、社会网络挖掘、社区发现、研究兴趣发现及论文推荐等。此外，论文信息的整合还要考虑中英文论文信息收集和整理。在实现大规模数据的自动收集、整理后，引入用户参与。这里的用户主体为高校教师，但提供所有研究人员使用，实现一个完整的社会服务网络。

下面对课题涉及的研究内容进行详细阐述：

1) 教师个人信息的收集。教师的个人信息是指姓名、性别、职称、研究方向等内容，是教师信息的基本组成部分。如何从互联网上获得教师的这部分信息，是系统所要解决的首要问题。传统的基于模板的信息抽取方式，需要针对不同格式的网页设计不同的抽取模式，这是需要耗费大量的时间和人力的。运用机器学习的方法来简化整个过程，自动化的对教师信息进行抽取是该研究主要关注的问题。
2) 教师研究信息的收集。对于高校教师来说，发表的论文情况最能反映其研究能力。因此，课题要从互联网上收集教师的论文信息。首先，要设计爬虫爬取论文信息，包括获得论文题目、作者信息、发表的期刊会议、发表时间等。本文通过适当的关键词到互联网学术搜索引擎中检索教师的论文，再把检索得到的论文信息作为查询串到学术网站进行再次搜索，最终得到每一篇论文的详细信息。但是，由于教师重名或者搜索引擎返回的结果不够精确等原因，结果中一部分论文可能并不是该教师的论文，这就需要通过适当的算法去除这部分杂质，提高结果的准确度。因此，重名消岐是这部分所要解决的第二个问题。
3) 社会网络构建。整个课题研究的主要问题就是通过自动构建出高校教师社会网络，为其提供个性化服务。而社会网络的基本信息来源就是论文数据中合作者关系。如果两个作者之间合作了一篇论文，那么很明显二者之间在现实的社会中必然有着比较紧密的联系。通过从论文的合作者关系中抽取教师之间的联系，组建出与现实吻合的社会网络。然后，如何找到该社会网络中联系比较紧密的小群体，即社区，是我们要研究的另一个问题。在通过论文合作者关系组成的社会网络的同一个社区内的用户，应该具有同一个方向的研究内容和研究兴趣。如果能从高校教师社会网络中挖掘出一个个社区来，也就能找到该社区中的专家人物并在后续工作中对其进行跟踪。

5、课题研究方案及进度

5.1 已完成工作

5.1.1 教师个人信息收集及实验

本系统目前已收集全国重点院校 120 所共计 4134 名计算机相关专业的教师的信息，且数据仍在不断更新。每个教师包含的已有信息项如表 1

表 1 系统教师信息项

序号	信息项	说明
1	姓名	
2	性别	
3	职称	教授、讲师等
4	是否院士	是否为中科院或工程院院士
5	导师类型	硕士生导师、博士生导师
6	研究方向	教师的研究领域
7	电话	
8	电子邮件	
9	大学	教师所在的大学
10	个人页面链接	信息来源页面

在系统收集到的数据基础上，进行了教师个人信息的抽取实验。我们提出了一种基于块划分的网页信息抽取模型从中文教师页面中抽取感兴趣的教师信息项。在该模型中，首先对包含教师信息的 HTML 页面进行预处理，包括规整 HTML 标签等，并进行分词和词性标注。然后利用具有语义化分作用的标点符号和 HTML 标签共同将整个页面分割为一个个句子，我们认为句子之间相互独立，即上一句的内容对当前句子内的信息抽取没有作用。条件随机域模型(Contional Random Fields, CRF)被作为标注模型对句子中的信息进行标注。

表 2.教师个人信息抽取实验中要抽取的信息项

信息类型	抽取任务	标记
基本信息	性别	G
	出生日期	BI
	导师类型	T
	职称	P
教育信息	本科毕业时间	BD
	本科学校	BU
	本科专业	BM
	硕士毕业时间	MD
	硕士学校	MU
	硕士专业	MM
	博士毕业时间	PD
	博士专业	PU
	博士学校	PM
联系方式信息	传真	F
	电话	PH
	邮箱	E

在实验中共使用了三类特征信息，词性特征，启发词特征和信息项值特征，详细介

绍如下：

- 词性特征。使用词性标注得到结果。
- 启发词特征。对信息抽取具有很强启示作用的词语。例如，在教师职称信息的前面一般会有“职称”，电话信息的前面会有“电话”这样的词。通过对这类情况的分析，我们得到了 40 个有效词汇。
- 值特征。原始的中文单词本身被用来作为特征。此外，一些信息项，如“性别”，其值只能是“男”“女”之一，即来自于固定的一个值集合中。我们根据常识信息对这些信息项值进行收集，最终得到了 34 个值信息词汇。

为了对该方法进行评价，我们从已有系统中收集了 636 个教师网页，并进行人工标注作为实验数据集。每个教师网页所标记的信息项如表 2 所示。首先针对每一个词生成特征，发现对于基本信息和联系方式信息该方法已经能得到较好的结果。而教育背景信息，因为前后文之间的联系比较复杂，存在远距离控制等，信息抽取的结果不够好。为此，在接下来的实验中将有关教育背景的词特征扩展到句子级，实验表明有效提高了信息抽取的结果。

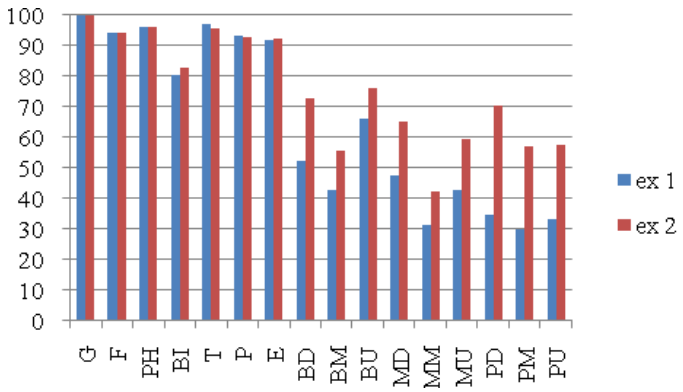

图 1. 实验 F-值

表 3. 实验结果

信息项	Ex1(词语级特征)			Ex2(句子级特征)		
	Prec.	*Rec.*	*F1*	*Prec.*	*Rec.*	*F1*
G	99.76	100	99.88	99.76	100	99.88
F	94.33	95.00	94.05	94.33	95.00	94.05
PH	96.98	95.11	96.03	96.98	95.11	96.03
BI	79.52	81.27	80.24	81.66	83.92	82.69
T	98.78	95.17	96.92	97.30	93.69	95.44
P	92.69	93.54	93.10	91.99	93.57	92.76
E	95.19	89.48	92.07	94.24	90.98	92.41
BD	59.70	47.76	52.24	72.43	73.45	72.76
BM	40.43	46.29	42.92	50.80	61.70	55.47
BU	63.27	68.94	65.97	74.07	79.24	76.29
MD	55.19	44.22	47.49	72.43	61.46	65.19
MM	32.19	33.23	31.33	44.59	46.36	42.49
MU	56.19	37.33	42.64	63.00	60.61	59.58
PD	40.48	31.14	34.72	71.29	69.99	70.26
PM	28.51	32.21	29.91	55.39	60.16	57.01
PU	33.18	32.94	33.08	55.46	60.30	57.72

5.1.2 论文信息收集

论文信息最能反映一个高校教师的研究情况。针对系统中已有的 4000 多个高校教师，系统从互联网上获取论文信息。具体流程如图 2。

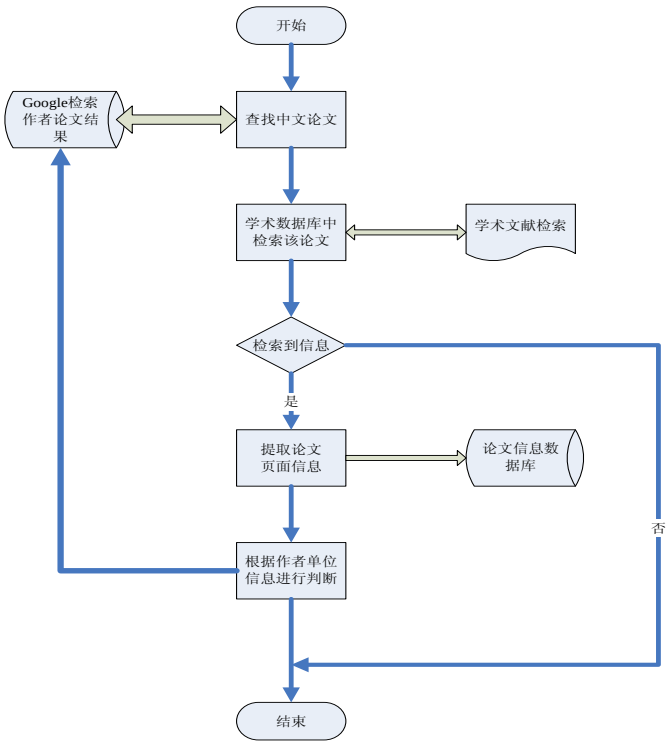

图 2. 论文获取流程

对于系统中已经收集到的每一个高校教师，把姓名和学校组合后作为关键词放到谷歌学术搜索引擎中进行检索，对于返回的检索结果，再到国内著名的学术网站万方[7]、知网[8]和维普资讯[9]进行查询，得到更详细的论文信息并存储到数据库中。最终，得到了 4 万多篇相关论文，其中近 1 万 5 千篇能通过作者单位信息进行确认。

5.2 剩余工作进度安排

研究的目的是为广大互联网用户，解决他们遇到的问题，这是每一个科研人员的理想和最终动力。本课题的提出就是为了满足相关用户的需求，为他们提供高质量的服务，因此在整个过程中，如何把现有的成果转化到已有系统中，是我们一直要做的一件事情。每一个阶段结束后，都会对现有系统进行改进和完善。

为了完成整个课题，还需要做以下两部分内容：

1) 论文信息整理。在之前的工作中通过设计爬虫获得了教师论文信息，但是得到的结果中有一部分并不是正确的结果。例如因为重名等原因，得到的论文信息可能是另外一个与其重名的教师的论文。因此需要从论文信息中去掉该部分杂质，提高准确度以便下一步更好的抽取社会网络。在之前获取论文详细信息的过程中，一部分论文能够得到作者的单位信息。我们假设每一个高校计算机学院的教师没有重名的，因此，可以通过作者单位信息对该论文进行一个初步的验证，判断是否该论文是当前查询教师的论文。再以得到验证的论文信息作为准确信息，对未得到验证的论文信息进行判断，从而得到较为准确的结果。

---

[7] http://www.wanfangdata.com.cn/
[8] http://www.cnki.net/
[9] http://www.cqvip.com/

2) 社会网络抽取和社区发现。在该阶段主要从已经获得的论文合作者信息中抽取作者之间的关系，利用这种联系组建出一个社会网络。如果两个作者合著一篇论文，那二者在现实中也必然存在紧密关系，研究兴趣和方向相同或者来自同一个实验室，可以用点来代表每一个论文作者，如果两个作者合著了一篇论文，则在二者之间连一条边，对所有的论文进行这样的处理，最终就会得到一个网络图出来。在建立网络的过程中，还需要考虑边的权重等信息。之后，要从社会网络中挖掘出社区。所谓社区，就是在社会网络中相互之间联系比较紧密的小群体，比如同一个实验室的研究小组，成员经常合作一些论文，他们就构成了社会网络中的一个社区。社会网络中的社区发现已经有了比较多的研究，我们需要对一些方法进行尝试，以期得到更符合高校教师社会网络的社区发现方法。得到社区后，很容易就能够发现该社区中的专家人物，这里的专家人物是指在该社区中起到核心作用的人，比如一个实验室的负责人，在该社区中活动较为频繁，与社区成员及社区外部之间存在较多的联系。对于专家人物，是人们比较关心的对象，而课题就可以对这部分人物的行动进行跟踪。专家跟踪包括从互联网上获取该人物相关的新闻、博客以及其他所有相关信息内容，通过整合处理等，展示出该专家的社会活动情况和其他多方面的信息。

剩余工作安排如下表

时间	工作安排
2010.09 — 2010.10	重名区分和社会网络的调研。
2010.11 — 2011.01	基于已有数据进行重名区分实验，并将实验结果应用到实际系统中
2011.02 — 2011.05	社会网络实现及相关实验，并应用到实际系统中
2011.06 — 2011.06	工作总结及毕业设计论文撰写

6、课题的研究条件和困难

该课题已经具备基本的研究条件，但是在实际系统的实现过程中，如何把研究结果与应用结合起来，让系统能够反映出最新的研究信息，这是需要仔细思考和解决的。同时，系统的实用化，能够为真实用户服务，保证稳定性、易用性、可靠性，是所面临的一个难题。相信在周围老师同学的帮助下，能够得到相应解决。

7、参考文献

[1] Xiangfen Wei,Ning Jia,Quan Zhang. Research on a Model of Extracting Persons' Information Based on Statistic Method and Conceptual Knowledge. The 7th International Conference on Chinese Computing, WuHan, China, 2007:2-6

[2] 任宁. 大规模真实文本中的人物职衔信息抽取研究.北京语言大学硕士论文.2008:4-7

[3] 李保利,陈玉忠,俞士汶. 信息抽取研究综述[J].计算机工程与应用,2003 ,39 (10) :1 - 5.

[4] Gaizauskas R, Wilks Y, Information Extraction: Beyond Document Retrieval. Journal of Documentation, 1997.

[5] S. Soderland. Learning information extraction rules for semi-structured and free text. Machine learning, 1999,34(1): 233-272

[6] Rohit J. Kate. A dependency-based word subsequence kernel. Annual Meeting of the ACL, Honolulu, Hawaii, 2008,ACL:400-409

[7] A. Culotta, J. Sorensen. Dependency tree kernels for relation extraction. Proceedings of the

Association for Computational Linguistics, Barcelona, Spain ,2004, Morristown, NJ, USA,ACL.

[8] D. Zelenko, C.Aone, A. Richardella. Kernel methods for relation extraction, The Journal of Machine Learning Research, March. 2003,3: 1083-1106

[9] D. Freitag, A. McCallum. Information extraction with HMM structures learned by stochastic optimization. Proceedings of AAAI, 2000, AAAI Press:584-589

[10] A. Culotta, A. McCallum, J. Betz. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. HLT-NAACL, New York ,2006,ACL:296-303

[11] F. Peng, A. McCallum. Information extraction from research papers using conditional random fields. Information Processing and Management, 2006,42(4): 963-979

[12] H. Poon, P. Domingos. Joint inference in information extraction. Aaai Conference On Artificial Intelligence, Vancouver, British Columbia, Canada, 2007, AAAI Press : 913

[13] Dietterich, Thomas G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine Learning , 2000,40(2):139-157

[14] Radu Florian, Abe Ittycheriah, Named Entity Recognition through Classifier Combination. Proceeings of the Seventh CoNLL conference, 2003:168-171

[15] E.M. Kleinberg. A Mathematically Rigorous Foundation for Supervised Learning. In J. Kittler and F. Roli, editors, Multiple Classifier Systems. First International Workshop, MCS 2000, Cagliari, Italy, volume 1857 of Lecture Notes in Computer Science, Springer-Verlag, 2000,pp.67–76

[16] Blum A, Mitchell T. Combining labeled and unlabeled data with Co-Training. In Proc. the 11th Annual Conf. Computational Learning Theory (COLT,1998), Madison, USA, July 24-26, 1998, pp.92-100

[17] K. Yu, G. Guan, and M. Zhou, "Resume Information Extraction with Cascaded Hybrid Model", In Proc. of ACL'2005

[18] T. Kristjansson, A. Culotta, P. Viola, and A. McCallum. Interactive information extraction with constrained conditional random fields. In Proc. of AAAI'04, 2004

[19] J. Tang, D. Zhang, and L. Yao. Social network extraction of academic researchers. In Proc. of ICDM'07, 2007, pages 292–301

[20] H.Pasula, B.Marthi, B.Milch, S.Russell, and I.Shpitser. Identity uncertainly and citation matching. In Proceedings of Neural Information Processing Systems: Natural and Synthetic, number 15,2002

[21] I.P.Fellegi and A.B.Sunter. A theory for record linkage. Journal of the American Statistical Association,1969,64:1183-1210

[22] D.Bitton and D.J.DeWitt. Duplicate record elimination in large data files. ACM Transactions on Database Systems,1983,8(2):255-265

[23] M.-L.Lee, T.W.Ling, and W.L.Low. Intelliclean: a knowledge-based intelligent data cleaner. In In 6th International Conference on Knowledge Discovery and Data Mining, pages 290-294,2000

[24] A.E.Monge and C.Elkan. An efficient domain-independent algorithm for detecting approximately duplicate database records. In Research Issues on Data Mining and Knowledge Discovery, pages 23-29,1997

[25] A. McCallum, K. Nigam, and L. H. Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. In Knowledge Discovery and Data Mining, pages 169-178,2000

[26] H. Han, C. L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis. Two supervised learning approaches for name disambiguation in author citations. In Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, pages 296–305, Tuscon, USA, June 2004.

[27] H. Han, H. Zha, and C. L. Giles. Name disambiguation in author citations using a k-way spectral clustering method. In Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, pages 334–343, Denver, USA, June 2005.

[28] I.-S. Kang, S.-H. Na, S. Lee, H. Jung, P. Kim, W.-K.Sung, and J.-H. Lee. On co-authorship for author disambiguation. Information Processing and Management, 45(1):84–97, January 2009.

[29] Y. F. Tan, M.-Y. Kan, and D. Lee. Search engine driven author disambiguation. In Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries, pages 314–315, Chapel Hill, USA, June 2006

[30] Leydesdorff L, Vaughan L. Co-occurrence matrices and their applications in information science: extending ACA to the Web environment [J]. Journal of the American Society for Information Science and Technology, 2006, 57 (12) : 1616-1628