北 京 大 学

硕士研究生学位论文

题目：基于概率模型的名人网页相关度评价

姓　　名： 刘晓莉

学　　号： 10208074

院　　系： 信息科学技术学院

专　　业： 计算机系统结构

研究方向： 网络与分布式系统

导　　师： 李晓明　教授

二零零五年五月

版权声明

摘　　要

本文的工作是在北京大学网络实验室、北京大学计算语言所与北京大学—IBM 创新研究院联合研发的天网知名度系统(Fame)中开展的。针对原有系统名人网页相关度评价中存在的问题，本文中提出了一种基于概率模型的名人网页相关度评价模型。

首先，针对 Fame 系统中名人网页相关度评价的特点，构建基本相关度评价模型。构建基础是 Okapi BM25 检索模型，在其基础上引入 HTML 标记权重系数，改进 Okapi BM25 公式，弥补其没有考虑 HTML 标记的不足。利用 Fame 系统数据集进行评测，实验结果表明 HTML 标记系数的引入提高了系统相关度评价质量，同时显示该基本模型优于原有系统中的相关度评价模型，提高了系统性能。

其次，由于不同领域名人的属性信息对其相关度评价有不同的贡献，本文中构建了区分领域的多层次实体模型，来更好地描述用户的信息需求。同时在基本模型基础上引入属性信息权重系数，使基本模型从不支持结构化查询需求改进为支持多层次实体模型。各领域的权重系数通过训练集训练的方式获得，避免了人工赋予方法的不确定因素。选取对系统相关度性能提高最大的一组权重系数作为模型中的领域参数，该套参数通过测试集的测试，证明有较好的适用性。

再次，采用了伪反馈和用户反馈两种相关反馈方法，为实体属性信息进行权重的自动调整，以达到系统相关度评价的进一步优化。通过实验得出的结论是：1）初始检索的质量很大程度地影响伪反馈的效果。应该先对初始检索模型进行优化，再使用伪反馈，这个顺序很重要；同时初始检索的质量需要达到一定高度后，使用伪反馈才能提高系统检索质量，目前系统的初始检索质量仍不适宜直接进行伪反馈。2）用户反馈在总体上自动优化了属性信息权重，提高了系统相关度评价质量。3）用户反馈的效果受名人实体属性信息词数的影响，属性信息越丰富，采用用户反馈后评价质量提高的概率越大。

关键词：信息检索，相关度评价，概率模型，相关反馈

Probabilistic Model-Based Relevance Evaluation of Famous People's Web Pages

LIU Xiaoli (Computer Architecture)
Directed by LI Xiaoming

Abstract

Tianwang Fame is an individualized information retrieval system. According to the main problems of the original relevance evaluation models, a probabilistic model-based relevance evaluation model is proposed in this dissertation to improve the relevance ranking of famous people's web pages.

First, a basic evaluation model is built on the basis of Okapi BM25. The author introduced an HTML weight to Okapi BM25. The test on Fame data collection shows that this basic model brings improvements to the system.

Second, a multi-level area-differed entity model is built to fully describe the entities. The author brings an area-differed attribute weight to the basic model. The selection of the weights' values is based on the training process on Fame's training set. The variables are tested be effective to improve famous people's relevance evaluation.

Last but not least, both pseudo-feedback and user feedback methods are used to carry out relevance feedback to provide an automatic justification for the relevance weights and refine the system. Several conclusions are drawn from the experiments:

1) The quality of the initial evaluation affects the pseudo-feedback a lot. The refinement to the initial ranking model should be carried out first and then the pseudo-feedback. The order is very important.

2) User feedback improves the relevance ranking quality in total.

3) The result of user feedback is highly connected to the number of the words in entity's attributes. The richer the attributes are, the bigger the probability of improving relevance evaluation quality after user feedback.

Key words: information retrieval, relevance evaluation, probabilistic model, relevance feedback

目　录

图 目 录

表 目 录

??? 引言

WWW（World Wide Web）作为一种全新的信息资源，获得了极大的发展，为人类信息获取提供了一个丰富的宝库。信息的有效检索随之变得举足轻重。

1.1 项目背景

大量实验和研究表明，WWW 上整体网页的数量以指数形式增长[1, 2, 3]。根据中国互联网络信息中心 CNNIC“2004 年中国互联网络信息资源数量调查报告”[4]，截至 2004 年 12 月 31 日，全国域名数为 1852300 个，与 2003 年同期相比增长 56%；网站数为 668900 个，同期相比增长 12.3%；网页总字节数增长最快，同期相比增幅为 238%。网页总数为 6.5 亿个，同期相比增长 108.6%；平均每个网站的网页数为 1297 个，同期相比增长 147.5%。在线数据库数为 306000 个，同期相比增长 80.1%。截至 2005 年 5 月，Google 公布的数据是搜索 8,058,044,651 张网页[5]。

目前人们在网上寻找信息时，大部分还是基于传统的信息浏览方式，主要工具是浏览器。通过浏览器获取信息主要有三种方式[6]：

1）直接向浏览器输入一个该信息源的网址（URL），例如，http://net.pku.edu.cn，浏览器将返回所请求的网页，用户可以根据该网页内容及其包含的链接文本或图像的引导，获得自己需要的内容；

2） 登录到某个知名门户网站，例如 http://www.yahoo.com.cn，根据该网站提供的分类信息和相关链接，进行网上“冲浪”，寻找自己感兴趣的内容；

3） 访问某个搜索引擎网站，例如，http://e.pku.edu.cn，输入自己关心信息的关键词，根据返回的相关网页列表、摘要和链接，试探寻找自己需要的信息。

这三种方式各有特点，各有相对最适合的应用场合。第一种方式的应用是最有针对性的，例如要了解北京大学网络实验室在做些什么工作，得知该实验室的网址是 http://net.pku.edu.cn，于是直接把这个 URL 输入浏览器就是最有效的

方式。第二种方式的应用类似于读报，用户不一定有明确的目的，只是想看看网上有什么有意思的消息；当然这其中也可能是关心某种主题，例如体育比赛、家庭生活等。第三种方式适用于用户大概知道自己要关心的内容，例如，“和谐社会”,但是不清楚哪里能够找到相关信息(即不知道哪些 URL 能给出这样的信息)；在这种场合,搜索引擎能够为用户提供可能是相关网页的一个网址及其摘要的列表，由用户一个个试探，看是否是自己需要的。现在的搜索引擎技术已经能够做到在多数情况下满足用户的这种需要。

然而，上述这些没有覆盖人们的所有信息需求。例如，一个人可能会关心最近半年来网上出现了哪些关于他（她）的信息，一个企业可能要关心它做了一次大规模促销活动后一个月内网上有什么反响,一个政府机构可能会关心在一项政策法规颁布后网上的舆论。对于这样的信息需求，目前的网上信息系统都不能很好地满足。在上面三种信息获取方式中，只有第三种方式可以勉强地、间接地提供这类信息服务，但是需要通过不断提供各种查询词、反复试探，繁复、效率不高而且很不方便。

以一个例子来说明“繁复、效率不高、很不方便”。比如某著名“信息技术”公司的“总经理”“王晓东”希望了解最近一段时间来网上有些什么关于他的信息（即他最近在网上的“知名度”如何)。为此，他登上某个搜索引擎，例如 http://e.pku.edu.cn，输入名字“王晓东”。极大的可能是，搜索引擎返回给他上万个条目，大致一看，许多虽然含有“王晓东”三个字，但和他一点关系都没有。于是他下一步输入“总经理”，利用搜索引擎提供的“结果中查询”功能，将上万个条目限制到上千个，但是其中大部分仍然还是和他无关。他当然可以一个一个查看，记录下确实和他相关的，然后仔细研究其内容；但这显然“繁复、效率不高、很不方便”。这里的问题在于，现在的搜索引擎一般都是通用的，要准备响应用户提出的任何查询词,同时没有预先保存关于查询用户的任何指定信息的特征，因此给出的返回信息只能尽量“包罗万象”，谈不上针对性。利用多重关键字进行限制能起到一些作用，但效果还是不够好。而且由于搜索引擎是把查询作为一个无结构的词串来处理,会把用户输入的多重关键词查询到的文档集合取严格交集返回，因此当输入查询词比较多的时候，又常会导致找不到任何信息。

因此迫切需要一种能够为用户自动收集、分析和整理具有预定特性信息的信息服务系统,其特点是如下两个方面的结合：

1）大规模网上信息的收集(主要是 Webpage 收集和整理)；

2）用户预先提供尽量确定的目标信息特性。

目前的 Web 搜索引擎能够较好地完成 1)，而信息过滤、智能检索等技术能够为实现 2）提供一定的基础。将二者结合起来，有可能实现一种网上预定特性信息的收集、评价与分发系统，其特点是：系统持续不断地从 WWW(或企业网内部)上收集和保存网页（或任何指定格式的文件），并把满足要求的网页（或文件）以指定的方式加工、存储和分发（例如按照评分、更新时间、文件大小等指标、进行排序、分类、自动摘要和 Email 通告等)。

上述技术目标就是天网知名度项目的立论动机。为进一步明确信息的预定特性，本项目限定用户提供的预定信息为一个或多个实体（包括人和公司/机构）的描述信息（例如人名、工作单位、行业与社会职务分类、或者公司/机构名称、主要业务、产品等)，这样系统将自动为用户指定的实体对每个网页进行相关度评价，并把相关的网页进行汇集、排序和加工。用户可以由此定期和定量地获得网上对其自身做了报道或描述的相关网页，由此可以产生一种优质的个性化网上知名度信息服务。

2002 年 7 月始,北京大学网络实验室、北京大学计算语言所与北京大学—IBM 创新研究院联合研发天网知名度系统。天网知名度项目在天网搜索引擎的基础上，结合中文信息处理的资源和先进技术，以名人实体为起点，针对名人 WWW 网页的特点，创建用于表示其特征的用户属性信息表示，建立相关度评价模型，进行名人网页的过滤和评价工作，并提供个性化检索和定制信息的主动推送服务。

1.2 相关工作

1.2.1 天网知名度原有系统

查询输入、文档表示和相关度评价是信息检索模型的三个基本方面。在天网

知名度系统中，查询是用户在注册时填写的名人属性信息，系统中为了对其进行更好的描述建立了名人实体模型，包括 8 类属性信息：领域、姓名、工作单位、职业/职务、兼职、社会形象、特征词、代表作；文档是系统收集的网页；本文基于概率模型，通过改进相关度评价算法和采用相关反馈来提高名人网页相关度评价的质量。

相关网页排序结果的优劣是系统服务质量的最根本体现，因此名人网页的相关度评价算法是系统的关键所在。相关度，在本系统中认为是用户注册信息(代表用户信息需求)与网页的匹配程度。

原有系统中有两个相关度评价模型[6]：

1）基于信息提取的布尔加权模型，简称模型 1。其评价方法是对网页表示库中的每一个网页，检查其人名列表，检索用户信息库，对其中已注册的人名（实体名）建立一个该网页对该人名的相关度评分初值；对检索出的注册名人实体列表，检查该网页中的二元关系和实体信息库，对符合匹配的关系为该网页的相关度评分增加一定分值，同时利用排除词表过滤掉重名的无关网页；对网页分词中的有效词（对语义理解有效的大部分实词）分别检索实体信息库的八类信息，分不同情况为该网页对名人的相关度评分增加不同分值；对网页分词中的有效词检查其 HTML 标记，分不同情况为该网页对名人的相关度评分增加不同分值；根据网页长度、网页中的人名个数等因素调整其相关度评分值。

2）组合向量空间模型（Combined Vector Space Model，CVSM），简称模型 2。其评价方法是，对于实体属性的八类信息分别创建八个向量，每个向量的维数是该类信息包含的词的个数。相应的，根据实体属性对应的八个向量对实体相对应的网页分别提取相应的八类信息的向量表示，分别计算这八对向量中两两之间的相似度，然后根据每类信息各自对相关度的贡献大小对这些相似度加权求和，形成最后的网页信息与实体属性的相关度评价结果。网页文档向量各个维的权重根据该词的绝对词频及其 HTML 标记等信息来组合计算。

在对相关度评价结果的评测中，模型 2 优于模型 1，原有系统的相关度评价质量以模型 2 为准。本文中对原有系统相关度质量进行改进的参照为模型 2。

Fame 中模型 2 相关度评价结果与 Google、Baidu、Tianwang 以 Fame 中[人名+单位+职业]和[人名+单位]等信息作为查询关键词返回的结果，采用 P@20 进

行的评测比较,结果显示 Fame 在该项指标下，与 Google 的检索结果基本相当，并优于 Baidu 和 Tianwang 的检索结果[6]。

但是，从多层次的相关度判别角度出发（本系统中网页的相关度分为高、中、低三个层次），对网页不再仅进行二元的相关或不相关的判断时，高相关度的网页排在前面才是一个实用的检索系统质量的更可靠指标。原有系统中，高相关度网页被排到后面的情况仍较常见。

通过分析，系统中原有相关度评价模型 2 主要存在下列三个不足：

1） 向量空间模型的思想是测量 query 向量和文档向量之间的相似度。但是向量空间基本模型中并没有给出相似度的计算公式。通常使用向量之间夹角的余弦或向量内积作为相似度衡量标准，无论采用哪种方法向量中每个纬度的权重要如何计算都是对检索质量致关重要的[7]。模型 2 采用 CVSM 计算相似度时主要依据是属性信息词的词频（tf）和 HTML 标记。没有考虑逆文档频（idf）和文档长度（dl）的作用。很多研究和实验表明二者在相关度评价中是非常重要的因素。tf,idf[7],dl 是现有大部分搜索引擎在权重计算方法中都包含的因素[8]。idf 代表了一个词的区分度，在大量文档中都出现的词与仅在少量文档中出现的词对网页相关度评价的贡献是不同的。文档长度对相关度评价也是有影响的，一个查询词在一篇内容重复冗长的文档中出现 2 次，与在一篇简要精炼的文档中出现 2 次，对相关度的贡献是不同的，如果不考虑文档长度，那么就会使评价标准偏向于内容重复冗长的文档。因此需要在网页的相关度评价中加入对 idf 和 dl 因素的考虑（同时还要考虑到文档集合中所有文档的平均长度，avdl，作为比较依据）。

2） 不同领域名人的实体模型没有进行区别。通过观察和实验发现，不同领域的名人，其高相关度网页的内容有比较明显的差别，且有一定规律可循，体现了各类属性对相关度评价质量的贡献不同。这与名人的领域有关，不同的领域性质，决定了舆论的不同特性。其中政府类名人的高相关网页往往比较正式，关于其出席某次会议、发表某个讲话等，关于其个人的专门报道比较少，这与其工作性质的要求有关，其相关网页中职业、职务的出现率较高；科教类名人的高相关网页内容也比较正规，其职业、职务类属性在其相关网页中出现率较高，而且其相关网页往往围绕其科研领域这个主题，即特征词属性；而媒体、演艺类名人的高相关网页则常常围绕他们的代表作、特征词类属性展开，比如对演艺类名人相

关网页的主题往往离不开其拍的电影、推出的专辑等。因此在实体模型中需要根据领域对名人的属性信息进行区别对待，根据领域提高其相应“精华”属性对相关度评价的作用力。

3） 参数问题。首先，模型 2 即组合向量空间模型中引入八个向量，因此也引入了 8 个属性信息词权重系数，其取值由人工赋予，存在一定不确定因素；其次属性信息词权重系数的调节需要人工参与，该模型中的属性信息词权重由文档集决定，是固定值，实际系统需要根据使用效果自动调节内部参数，实现进一步的优化。要使系统具备参数自动调节以实现自动优化的功能，首先需要改变赋值方式，用训练的方式替代人工赋予，减少不确定因素；其次需要提供属性信息词权重的自动调节机制。

本文中将针对模型 2 的以上三个不足，提出基于概率模型的相关度评价模型，以提高 Fame 系统的相关度评价质量。

1.2.2 Ask Jeeves “Famous People Smart Search”

2004 年 8 月，Ask Jeeves 在其英国站点 www.ask.co.uk 增加了名人智能搜索“Famous People Smart Search”，该功能是为了满足大量搜索用户关于名人信息的搜索请求而推出。“名人搜索”包括搜索音乐家、电影明星，政治家，运动明星，历史人物及其他被关注人群。根据用户输入的人名，将人物图片，背景传记或者新闻等不同类型的信息资源综合成一个结果，形成对该名人的一个答案返回给用户。例如，输入 Queen Elizabeth，会返回英国现任女王的信息，包括图片、传记及女王的官方网站，但是结果也会同时给出 16 世纪英国女王 Queen Elizabeth I 的链接。“Famous People Smart Search”收录了 2，000 多位世界著名人物的信息[9]，它的目标是对已收录的名人给出一个包括其传记等信息的答案，所有查询同一个名人的用户会得到相同的答案。天网知名度系统中则是由用户来预定目标名人特性，不同用户查询的同一个名人在天网知名度系统中是被当做不同实体的，系统根据用户对目标名人的信息描述，对网页进行相关度评价，返回其相关网页，按相关度从高到低的顺序或按时间顺序等，是根据用户的需求提供个性化名人检索服务，目前只针对 WWW 网页资源进行评价。

1.2.3 专家查找

在相关研究和系统中，专家查找系统是信息检索、知识管理、计算机支持协作工作（CSCW）等几个领域的交叉研究点[10]。目前已经存在一些专家查找系统[4]。专家查找系统主要是通过专家模型对可得资源进行匹配来识别某领域的专家，其目标是根据用户的需求寻找该领域的人物（专家）。天网知名度系统是在用户基本明确目标人物的前提下，为用户进行相关信息获取。

文献[10]中对专家查找领域做了整体的介绍和分析。进行专家查找的两个最主要的动机是：专家是一种信息源和专家具有实施某项组织或社会功能的能力。一种实现方法是依靠专家数据库的支持，其中 SPUD（Microsoft）、CONNEX（Hewlett-Packard）和 SAGE People Finder 是使用这种方法的典型系统。专家数据库的缺点是：人工建设数据库工作量非常大、数据容易过时而且对于专家描述的完整性和针对性往往不够。个人网页也是专家查找时的一个重要信息来源。如果专家的个人主页能够及时更新并且包含了相应的关键词，那么通过按主题搜索，就可以找到这些网页。这些网页显然是比专家数据库更好的信息源，但是通用的搜索引擎并不容易满足这种需求，主题搜索返回的网页太多，而且它基于关键词匹配，用户需要不断调整关键词才能找到最合适的专家，比较麻烦。

由于上述方法的缺点，一种能够从信息资源中自动发现专家信息的系统成为需要。这方面比较早期的工作，可以在 HelpNet 中看到，用户输入对某个领域的专家需求，系统返回给用户一个专家人名列表，以可能解决这个问题的概率排序。该系统事先要求人们从已有的一些主题中选择可以表明他们特长的主题，并且填写自己的能力可以为该方面提供一个满意答案的概率，系统基于这些数据使用概率模型计算满足用户需求的专家的排序。Expert/ Expert- Locator(EEL)[11]可以使用自然语言查询，采用潜在予以索引（Latent Semantic Indexing，LSI）返回一个最相近的研究组。ContactFinder[12]是一个智能代理，从讨论组中提取某些特定领域的联系方法。还有的系统[13]根据参考链接来提取专家信息等等。

文献[10]中，作者对专家查找系统进行了一个系统的、结构化的描述：

1） 专家识别的基础：包含各种信息源，不同类型的文档、数据库、个人等。组织内部的所有文档都是识别专家的潜在资源，但是往往由于所有权、隐私等原

因不是所有的资源都可以使用，专家系统需要对可利用资源进行识别和采集。系统可以在最初定义它所处理的资源类型。

2） 专家信息提取：从系统获取到的资源中提取专家信息，采用的技术可以分为领域知识独立的和领域知识驱动的。

3） 专家模型：对个人的技术/技能的元描述，或者这种描述的一个独立保存的、有组织的结构。这种描述和数据库里的一条记录是不同的，它包含了一定的不确定程度，其建立是通过对专家信息的分析和有可能的推理。一个专家属性信息包含所有信息包括个人详细情况、他（她）的专长的模型以及与其他专家的关系。专家模型主要有两种产生方式：查询时产生和预定义。查询时自动产生的方式，可以使用信息检索系统检索资源数据库，从匹配的数据中提取专家模型，缺点是响应慢；预定义专家模型的方法，使用搜集代理定期的收集专家信息并且构建更新专家属性。专家模型，以及专家属性和其他相关信息。就形成了一个专家空间。

4） 查询机制：用户可以直接向系统提出需求或者由系统从用户的行为中进行分析推理。

5） 匹配：对于信息需求或者专家需求，主要是利用专家模型或专家属性信息来匹配，使用包括准确的关键词匹配和基于各种检索模型的相似度匹配的检索技术，甚至一些推理的机制，比如“如果 x 是主题 y 方面的专家，那么她可能（不）懂得主题 y”,或者“如果 x 知道主题 y,那么她可能知道谁是主题 z 方面的专家”。

6） 输出表示：像在信息检索技术中一样，可以对识别的专家应用一些不同的排序机制。除了专家本人的详细情况，其机构的或社会的关系以及专家间的网络也可以被利用。

7） 适应和学习：通过收集用户反馈信息使系统个性化，系统不光要能提供满足需求的专家，同时还要能够识别最合适的人选来满足用户的需求。使用户可以参与排序，通过用户的反馈来完善专家模型。由于专家是社会结构的一部分，系统也要收集对用户对专家的评价。 文献[11]提到对他们的系统通过对系统中接收的每个查询进行学习的计划。

专家作为一种信息源的不争事实，使专家查找系统成为一个重要而且非常有意义的领域。

1.3 本文工作

查询输入、文档表示和相关度评价是信息检索模型的三个基本方面。在天网知名度系统中，查询是用户在注册时填写的名人属性信息，系统中为了对其进行更好的描述建立了名人实体模型，包括 8 类属性信息：领域、姓名、工作单位、职业/职务、兼职、社会形象、特征词、代表作；文档是系统收集的网页。

本文的主要工作集中在相关度评价方面，提出了一种基于概率模型的名人网页相关度评价模型。通过改进相关度评价算法和采用相关反馈来提高名人网页的相关度评价质量。

首先通过分析，发现系统中原有的相关度评价模型主要存在下列三个不足：

1） 模型 2 采用 CVSM 计算相似度时主要依据是属性信息词的词频（tf）和 HTML 标记。没有考虑逆文档频（idf）和文档长度（dl）的作用。很多研究和实验表明二者在相关度评价中是非常重要的因素。

2） 不同领域名人的实体模型没有进行区别。通过观察和实验发现，不同领域的名人，其高相关度网页的内容有比较明显的差别，且有一定规律可循，体现了各类属性对相关度评价质量的贡献不同。因此在实体模型中需要对名人的属性信息进行区别对待，根据领域提高其相应“精华”属性对相关度评价的作用力。

3） 参数问题。首先，模型 2 即组合向量空间模型中引入八个向量，因此也引入了 8 个属性信息词权重系数，其取值由人工赋予，存在一定不确定因素；其次属性信息词权重系数的调节需要人工参与，该模型中的属性信息词权重由文档集决定，是固定值，实际系统需要根据使用效果自动调节内部参数，实现进一步的优化。要使系统具备参数自动调节以实现自动优化的功能，首先需要改变赋值方式，用训练的方式替代人工赋予，减少不确定因素；其次需要提供属性信息词权重的自动调节机制。

针对系统中原有名人网页相关度评价模型的不足，为提高系统相关度评价质量，本文提出了一种基于概率模型的名人网页相关度评价模型：

1） 首先，针对 Fame 系统中名人网页相关度评价的特点，构建基本相关度评价模型。构建基础是 Okapi BM25 检索模型，在其基础上引入 HTML 标记权重系数，改进 Okapi BM25 公式，弥补其没有考虑 HTML 标记的不足。利用 Fame 系统

数据集进行评测，实验结果表明 HTML 标记系数的引入提高了系统相关度评价质量，同时显示该基本模型优于原有系统中的相关度评价模型，提高了系统性能。

2） 其次，由于不同领域名人的属性信息对其相关度评价有不同的贡献，本文中构建了区分领域的多层次实体模型，来更好的描述用户的信息需求。同时在基本模型基础上引入属性信息权重系数，使基本模型从不支持结构化查询需求改进为支持多层次实体模型。各领域的权重系数通过训练集训的方式获得，避免了人工赋予的不确定因素。选取对系统相关度性能提高最大的一组权重系数作为模型中的领域参数，该套参数通过测试集的测试，证明有较好的适用性。

3） 再次，采用了伪反馈和用户反馈两种相关反馈方法，为实体属性信息进行权重的自动调整，以达到系统相关度评价的进一步优化。通过实验得出的结论是：第一，初始检索的质量很大程度地影响伪反馈的效果。应该先对初始检索模型进行优化，再使用伪反馈，这个顺序很重要；同时初始检索的质量需要达到一定高度后，使用伪反馈才能提高系统检索质量，目前系统的初始检索质量仍不适宜直接进行伪反馈。第二，用户反馈在总体上自动优化了属性信息权重，提高了系统相关度评价质量。第三，用户反馈的效果受名人实体属性信息词数的影响，属性信息越丰富，采用用户反馈后评价质量提高的概率越大。

另外，在系统方面，05 年 5 月，为 Fame 系统构建了新一批名人实体数据集。加入新名人实体 150 人，人工制作了相应的名人实体属性信息库和标注了相关度高、中、低三个等级的名人实体相关网页数据集（含网页 3887 篇，平均每位名人实体 26 篇相关网页）。在天网知名度系统中，用户预定的目标信息特性即名人实体属性信息对应于普通信息检索中的查询输入，系统依据它们对网页进行相关度评价。名人实体相关网页数据集用来评测系统评价算法的质量。本次数据集采取的是招募志愿者参与的方式来完成的，26 位来自校内外的志愿者参加了 2005 年 5 月“天网志愿者”活动，为该批名人实体数据集的制作付出了辛勤的劳动。

1.4 论文组织

本文共分为六章，论文是按照如下方式来组织的。

第一章，引言，介绍了天网知名度系统产生的背景以及一些相关工作，并分析系统原有相关度评价模型中存在的不足。

第二章，天网知名度系统，介绍系统流程，基本模块和实体数据集，包含了05 年 5 月份对系统原有流程中网页搜集模块搜集模式的改进和同期 150 位新增名人实体的实体数据集的构建。

第三章，基于概率模型的相关度评价，首先，针对 Fame 系统中名人网页相关度评价的特点构建基本相关度评价模型，构建基础是 Okapi BM25 检索模型，在其基础上引入 HTML 标记权重系数，改进 Okapi BM25 公式，弥补其没有考虑 HTML 标记的不足；其次，模型的实现中，针对名人网页特点，采用信息提取与基本相关度评价模型相结合的方式；再次，介绍系统中采用的两种评测方法，包括系统原有的评测方法和新采用的 DCG 评测方法 ;最后,通过 Fame 数据集对 HTML 标记权重系数引入前后的相关度评价质量进行评测比较,并对基本评价模型在系统相关度评价质量方面带来的提高进行评测。

第四章，支持多层次实体模型的相关度评价，是在第三章的基本评价模型基础上的进一步改进。首先,针对不同领域名人属性信息对相关度评价贡献的区别，构建多层次的名人实体模型；其次，进一步改进相关度评价模型，以支持这种多层次的名人实体模型；再次，使用训练集训练的方法给出参数设置，避免人工赋予的不确定性实现不同领域名人的个性化对待；最后，通过测试集的测试验证选定参数的适用性和对系统相关度评价质量的提高。

第五章，相关性反馈，在系统中分别采用伪反馈和用户反馈两种相关性反馈方法，实现参数自动调整、使系统相关度评价质量实现自动优化。通过实验比较两种反馈方法性能的差异并通过进一步实验分析了影响两种反馈效果的因素,得出如下结论：1）初始检索的质量很大程度地影响伪反馈的效果。应该先对初始检索模型进行优化，再使用伪反馈，这个顺序很重要；同时初始检索的质量需要达到一定高度后，使用伪反馈才能提高系统检索质量，目前系统的初始检索质量仍不适宜直接进行伪反馈；2）用户反馈在总体上自动优化了属性信息权重，提高了系统相关度评价质量；3）用户反馈的效果受名人实体属性信息词数的影响，属性信息越丰富，采用用户反馈后评价质量提高的概率越大。

第六章，总结和展望，首先对本文的工作进行总结，然后对将来进一步的工

作进行展望。

最后是参考文献和致谢。

第二章 天网知名度系统

天网知名度系统在天网搜索引擎的基础上，结合中文信息处理的资源和技术，以名人实体为起点，针对名人 WWW 网页的特点，创建用于表示其特征的名人属性信息表示，建立相关度评价模型，进行名人网页的过滤和评价工作，并提供个性化检索和定制信息的主动推送服务。本系统目前的访问地址为 http://net.pku.edu.cn/~fame 或 http://162.105.80.47。

2.1 系统流程

天网知名度系统在天网搜索引擎的基础上，增加了名人实体属性信息收集、信息提取等部分，并根据名人网页的特点，建立了相关度评价模型，对每个网页进行内容的相关度评价和褒贬性评价，并对相关网页提供多种排序机制，为用户提供检索服务，或以邮件等形式主动地向用户推送信息。

天网知名度系统流程图如下图 2-1 所示：

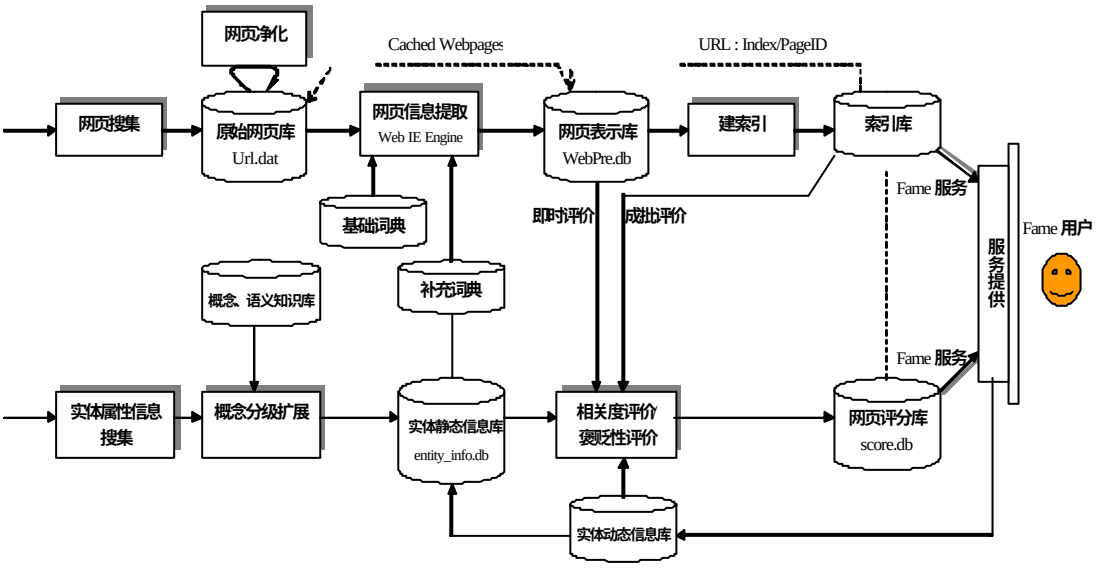

图 2-1 天网知名度系统流程图

图 2-1 的上半部分主要是天网搜索引擎的系统模块，网页信息提取除外。其

中网页搜集模块是以初始时人工制作的训练集中的网页为种子，使用天网的网页抓取模块，为系统抓取中文网页，形成原始网页库（Url.dat）。原始网页库是按系统需求结构存储的文本文件，其中网页以抓取顺序存放，存放网页内容之前都附加了固定格式的头信息（URL，时间，编码等）。先对原始网页进行网页净化的预处理，去掉一些网页中的噪音，下一步进入网页信息提取模块，在该模块中进行中文分词、词性标注、命名实体识别和实体二元关系提取，生成网页表示库（WebPre.db）。网页表示库是 BerkeleyDB 数据库格式，在这里，每个原始网页对应一个<key,value>对，其中 key 对应的是该网页的 URL，而 value 中存放网页长度（以词为单位）、网页文本词串、词串对应的词性标记串和 HTML 标记串、网页中提取出的人名列表、单位机构名列表、人名与单位关系列表、人名与职务关系列表。这一步是后面对网页根据实体属性信息进行相关度评价的基础。之后对网页表示库建立索引，主要是依据词频和位置等信息建立的，形成索引库。

图 2-1 的下半部分，最左边是收集用户注册信息模块，即在用户到系统注册时填写实体属性信息的模块，下一个模块对这些属性进行实体信息的概念扩展，提取用户实体信息补充词典，并形成实体信息库 entity_info.db，存放每个实体的属性信息。中间部分，是根据表示用户查询需求的实体属性信息对网页进行相关度评价的模块。这个模块基于网页表示库、实体信息库和实体属性信息描述，首先过滤出包含注册实体名的网页，然后根据网页内容与实体属性信息间的匹配程度对该网页中出现的每个实体进行相关度评价和褒贬性评价，得到网页评分文件，进一步处理得到网页评分库 score.db。

图 2-1 最右边的部分是用户服务界面，用户登录系统后，选择要查询的注册实体，系统便通过检索网页评分库 score.db 和网页索引库，返回给用户一个按照相关度由高到低排序的网址及自动生成的该网页的摘要的列表。同时提供多种排序方式，如按照相关度、时间、褒贬性等排序。

天网知名度系统服务器包括两台 IBM Netfinity 7600 服务器。主要配置为，cpu 类型：Intel Pentium III；处理器主频：700MHz；处理器个数：2 个；内存容量：512MB；硬盘类型：SCSI；随机硬盘容量：60GB。目前其中一台用来作为系统的主要服务器，在随机硬盘基础上挂载了一块 70G SCSI 硬盘，另一台主要用来进行一些数据的备份。

下面几个小节中分别介绍系统的主要模块—网页搜集模块、网页分析与索引

模块、网页评价模块和用户界面模块。其中网页搜集模块小节包含了 05 年 5 月份对系统原有流程中网页搜集模块搜集模式的改进，实体数据集小节包含了同期 150 位新增名人实体的实体数据集构建。

2.2 网页搜集模块及改进

原有系统中，网页搜集模块是以人工制作的名人实体网页数据集中的相关网页为种子，使用天网的网页抓取程序，为系统抓取中文网页，保存在原始网页库中。针对名人网页的特点，相关网页的一个基本条件就是网页中出现实体名。要为名人实体提供尽量丰富的相关网页信息，需要系统有足够丰富的网页资源。但是由于目前 Fame 服务器的现有配置下资源有限，无法存储超大规模的网页集合。目前系统原有的搜集模块抓取到的原始网页数量虽然达到 80 万，但是相当一部分网页中并没有出现系统中已经注册的实体名。

针对这个问题，我们对原有抓取模块进行改造，改变了搜集方式，不再自主抓取网页，而是采用从天网按实体名过滤网页的方式。这种搜集模式依托于天网搜集技术和天网更为丰富的数据源。既增加了 Fame 系统中名人网页的数量又节省了系统资源。 过滤的方法是，使用天网知名度系统中已经注册的实体名作为查询输入，访问天网查询端，将查询到的包含该人名的网页过滤到 Fame 系统中，网页数量的阈值定为 10，000 篇。如果天网查询端返回的网页数量大于 10，000 篇，那么选择返回序列中的前 10,000 条网页加入 Fame 系统，存入原始网页库，阈值的制定是基于对数据量与系统效率两个因素的考虑而做出的平衡。

过滤采取的是定期进行的方式。2004 年 5 月初从天网过滤到天网知名度系统中包含系统中注册实体姓名的网页 29 万余篇。从天网过滤网页部分的代码由由 Fame 项目组姚从磊同学完成。

改造后，系统提高了存储效率，更好地依托了天网搜索技术。

2.3 网页分析与索引模块

该模块在天网搜索引擎的基础上进行了扩展，主要功能是对搜集来的原始网

页进行内容解析、转变格式、分词与词性标注、信息提取、建立网页库索引、评分库索引、自动生成摘要等功能。

网页分析部分，是对抓取来的原始网页进行分析和预处理，主要完成如下工作：分析原始网页，提取 URL，日期，长度等网页属性信息；分析网页内容并处理 HTML 标记；将繁体字网页转化为简体字网页，避免繁体字在进行分词处理的困难；网页净化，去除网页中的噪音链接信息，调用天网的网页净化库程序是一个可选模块；提取出网页中的文字信息，并根据原始网页的内容，组织为文字段落；对文字串进行切词、命名实体识别、基于命名实体的二元关系提取等分析处理；匹配网页的原始内容和文字串分析后的结果，为每一个词语加上它的 HTML 属性；将处理结果保存为原始网页库。

该部分的结果是形成网页表示库 WebPre.db，存为 Berkeley DB 数据库。在这里，每个原始网页对应一个<key,value>对，其中 key 对应的是该网页的 URL，而 value 中存放网页长度（以词为单位）、网页文本词串、词串对应的词性标记串和 HTML 标记串、网页中提取出的人名列表、单位机构名列表、人名与单位关系列表、人名与职务关系列表。

网页索引在这里主要用来对原始网页库、评分结果库和网页属性库建立索引，以提供查询接口给用户检索模块使用（包括按实体编号查询对应的网页信息、按网页编号查询网页属性、摘要等）。网页索引主要完成以下功能：把原始网页追加到原始网页数据库，并给原始网页数据库建立索引（增量的方式）；同步更新网页属性信息库（增量的方式），为增加的新网页分配唯一的文档编号；根据网页褒贬性评价的结果文件更新网页属性数据库中褒贬性评价属性；根据网页相关度评价和褒贬性评价的结果文件，更新网页—实体数据库（增量的方式），提供给用户界面模块的用户检索功能使用。各数据库的结构如表 2-1 至表 2-3[6]所示：

表 2-1 原始网页库（WebDataTxt.db 和 WebData.db）结构

	字段名	内容
Key	Docid	网页编号
Data	Docinfo	原始网页内容，包括 url、lastmodifiytime、content

表 2-2 网页—实体评分库（eid_id.db）结构

	字段名	内容
Pkey	Eid	实体编号
Data	Readflag	已读/未读标志
	Score	网页对该实体的相关度的分
	Docid	网页编号
	Time	网页最后更新时间
	PositiveScore	网页褒义得分
	NegativeScore	网页贬义得分
	ScorePolarity	网页褒贬义得分

注：docid 以下的字段为网页属性，为提高检索效率，也作为本库的成员。

表 2-3 网页属性库（urlinfo.dat）结构

	字段名	内容
Pkey	docid	网页编号
Skey	url	网页地址
Data	Time	网页最后更新时间
	PositiveScore	网页褒义得分
	NegativeScore	网页贬义得分
	ScorePolarity	网页褒贬义得分

2.4 网页评价模块及改进

网页评价模块的工作是在前面网页分析模块的基础上，根据用户注册的实体属性信息，对网页进行有关实体的相关度评价以及网页内容的褒贬性评价。该模块是决定天网知名度系统质量的关键。

由于用户预定了其信息需求，即注册了名人实体属性信息，天网知名度系统根据这些属性信息建立适应网上名人特点的相关度评价模型。针对名人相关网页的特殊性，名人的相关网页必然要有其姓名出现，所以网页相关度评价模块首先根据用户注册实体的姓名对经过了网页分析模块处理的网页内容进行过滤，仅对

出现注册实体的网页进行相关度评价，在很大程度上提高了系统的效率。系统原有的相关度评价模型有两个：基于信息提取的布尔相关度评价模型（模型 1）和基于组合向量空间的网页相关度评价模型（模型 2）。

1）基于信息提取的布尔加权模型，简称模型 1。其评价方法是对网页表示库中的每一个网页，检查其人名列表，检索用户信息库，对其中已注册的人名（实体名）建立一个该网页对该人名的相关度评分初值；对检索出的注册名人实体列表，检查该网页中的二元关系和实体信息库，对符合匹配的关系为该网页的相关度评分增加一定分值，同时利用排除词表过滤掉重名的无关网页；对网页分词中的有效词（对语义理解有效的大部分实词）分别检索实体信息库的八类信息，分不同情况为该网页对名人的相关度评分增加不同分值；对网页分词中的有效词检查其 HTML 标记，分不同情况为该网页对名人的相关度评分增加不同分值；根据网页长度、网页中的人名个数等因素调整其相关度评分值。

2）组合向量空间模型（Combined Vector Space Model，CVSM），简称模型 2。其评价方法是，对于实体属性的八类信息分别创建八个向量，每个向量的维数是该类信息包含的词的个数。相应的，根据实体属性对应的八个向量对实体相对应的网页分别提取相应的八类信息的向量表示，分别计算这八对向量中两两之间的相似度，然后根据每类信息各自对相关度的贡献大小对这些相似度加权求和，形成最后的网页信息与实体属性的相关度评价结果。网页文档向量各个维的权重根据该词的绝对词频及其 HTML 标记等信息来组合计算。

在对相关度评价结果的评测中，模型 2 优于模型 1，原有系统的相关度评价质量以模型 2 为准。Fame 中模型 2 相关度评价结果与 Google、Baidu、Tianwang 以 Fame 中[人名+单位+职业]和[人名+单位]等信息作为查询关键词返回的结果，采用 P@20 进行的评测比较，结果显示 Fame 在该项指标下，与 Google 的检索结果基本相当，并优于 Baidu 和 Tianwang 的检索结果[6]。本文中对原有系统相关度质量进行改进的参照为模型 2。

本文的工作主要集中在这个模块，首先通过分析，发现原有相关度评价方法存在的不足，然后针对这三个问题提出一种基于概率模型的名人相关度评价模型，以提高系统相关度评价质量，将在第三章至第五章中详细论述。

另外，网页的褒贬性评价有着现实的检索需求，是个性化检索应该具备的功能之一。与网页内容的相关度评价相比，网页内容的褒贬性评价更需要做篇章的

分析理解，考察其中词语、句子、段落等的修辞特点、情感好恶取向等，区分主体与客体，贯穿全文得到整篇网页对其中实体的褒贬性评价结果。在天网知名度系统中，对名人网页的褒贬性评价作了初步的尝试。各种修辞形式是有力而常用的表示情感好恶的手法，但是目前天网知名度系统中关于语言学专家知识的考察和利用只从基本词语入手，没有分析修辞形式。一方面是因为现有种文字信息处理关于修辞形式的识别技术还不成熟，另一方面，对基本词语的褒贬特征分析确实可以覆盖一部分修辞形式。天网知名度系统目前使用的褒贬词词库共计一千余条。利用这些词语的褒贬特征，对 75 万网页中的用词进行统计分析，作为整个网页的褒贬特征取向，该方法简单、高效，评价的结果经人工考察，达到了一定的用户满意度[6]。

2.5 用户界面模块

用户界面是天网知名度系统为用户提供服务的窗口，主要包括用户注册、用户检索和信息推送三个功能。

用户注册功能主要包括新用户注册自己的信息（用户名、密码、email 地址）和注册所关注名人实体的属性信息。名人实体属性信息主要包括：领域、姓名、工作单位、职业、兼职、社会形象、特征词以及代表作，其中特征词一项填写用户对该实体的关注点。该模块将这些属性分别导入用户信息库和实体属性信息库。其提交界面如图 2-2 所示。

用户检索，即天网知名度系统为用户提供检索服务的模块。首先，用户登录以后，可以看到系统已有的名人实体列表，用户可以选择访问。系统提供了查看全部相关网页、已读相关网页和未读相关网页三个进入点，同时还可以查看该名人实体的注册信息。界面如图 2-3 所示。其次，用户选择查看某类网页后，点击相应链接，就可以进入相关网页检索排序界面了。系统通过访问评分库和索引库，返回给用户一个该实体相关网页的网址和自动摘要列表。系统对这个列表提供三类排序方式：按相关度评价结果（按系统中的多种评价模型），按时间（时间新的网页排在前面）和按褒贬性（正面性、负面性）排序。如图 2-4 所示。网页列表中的摘要是动态生成的，系统根据所查询的实体名，动态的从网页中提取相关内容，形成摘要。再次，由于很多网页是会在较短的时间内从网上消逝的，因此

检索界面中，对每个网页提供了网页快照功能，点击该链接，可以从系统的原始网页库中取出该网页内容显示给用户，所以即使该网页已经从网上消逝，由于系统中保存了原始网页，用户仍然可以获得该网页。目前系统中仅存储了原始网页的文本内容，由于存储空间有限，没有保存网页中的图片数据。

图 2-2 名人实体属性信息注册界面

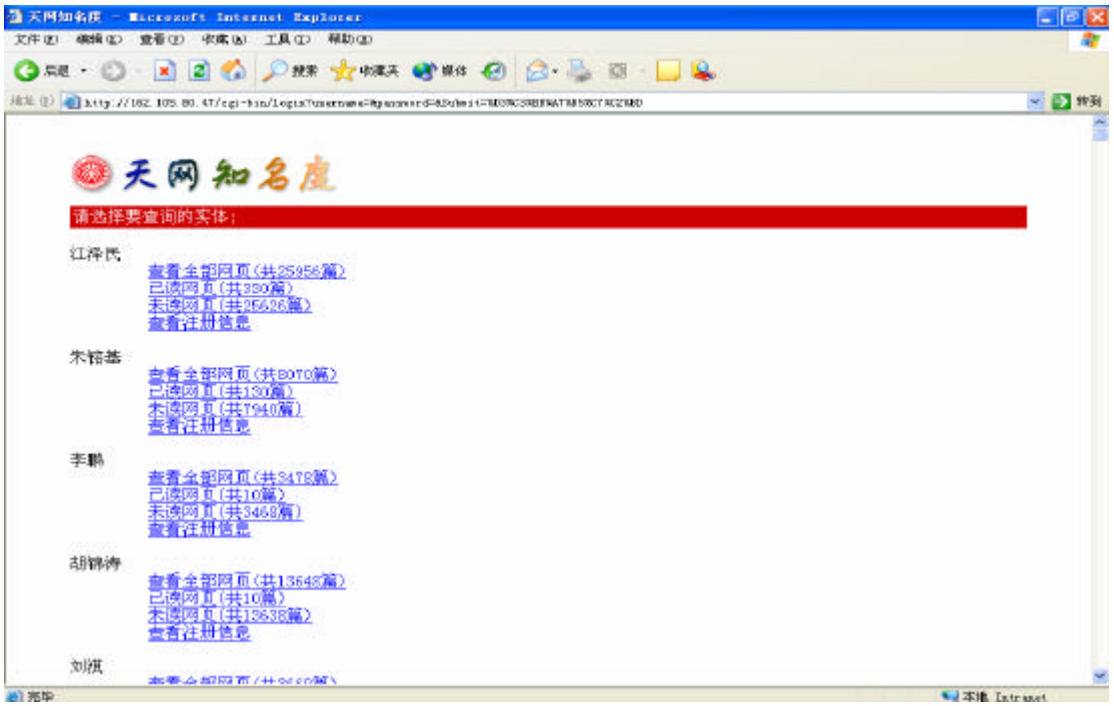

图 2-3 用户检索界面

图 2-4 实体网页的检索排序

信息推送功能是天网名人系统为方便用户使用而提供的，系统定期把新的检索结果通过 email 发送给用户。与传统的信息浏览方式不同，用户不需要每次向 Web 发出请求，等待 Web 检索后将满足用户需求的结果返回用户，只需要在首次使用时注册自己的信息需求，此后系统就会定期或定量的将用户订制的信息发送给用户，用户从主动地拉取信息变为只要等待信息的主动到达。

2.6 实体数据集及扩容

天网知名度系统的实体数据集包括名人实体属性信息库和名人实体相关网页数据集两部分。

天网知名度系统根据用户预定的名人实体属性信息对网页进行相关度评价。为保存用户的预定信息，我们构建了名人实体属性信息库，采用人工选取填写的方式。为了判定评价结果的质量，我们构建了名人实体相关网页数据集，采用人工制作的方式。为了保证相关网页数据集中名人实体间的均衡性，规定对每位名人收录 25 篇网页。系统第一批属性信息库和网页数据集中包含 8 个领域的 275

位名人，网页数据集中共 8776 篇网页。我们在 05 年 5 月份制作了第二批数据集，新增 8 个领域的 150 位名人，网页数据集共 3887 篇网页。这批属性信息库和数据集的制作采取的是招募志愿者参与的方式来完成的。26 位来自校内外的志愿者参加了 2005 年 5 月“天网志愿者”活动，为该批数据的制作付出了辛苦的劳动，非常感谢他们的工作。目前系统中共有 420 余位名人。

在确定新的 150 位名人名单时，作者首先进行了一次过滤，以名人实体的姓名为查询，使用天网搜索引擎，返回网页数目大于 1，000 条的实体才予以收录，确保收录的实体有一定的“知名度”，以避免第一批名人中部分实体相关网页数量较少的情况。这是在系统资源有限的条件下做出的一种权衡。

在制作名人实体属性信息库时，我们制定了“名人实体属性填写原则”，尽量丰富地选择能够描述实体特点的词语。在制作相关网页数据集时，制订了“相关网页选取及相关度判定原则”，数据集的质量对正确衡量相关度评价质量非常重要。主要包括：相关——有该人名出现，而且确实有关于该实体的论述。这是选择网页的前提，只在网页上的某个链接处出现这个人名或者网页内容为重名的另一个名人，则该网页为不相关网页；高度相关——集中的报道、介绍该实体，有代表性，内容较充实；中度相关——网页中有一部分内容是关于这个人物的。其它内容可能关于其它实体或其它事情；低度相关——网页中仅稍微有所提及这个人物。

相关网页数据集制作工具，沿用第一批数据集构建工具 MyFame，采用 Client/Server 结构，服务器端采用 SQL Server 管理和维护数据库，客户端支持多人并行工作。构建时，制作者首先以人名为查询到搜索引擎中获取相关网页，输入 MyFame，并对相关网页做出相关度高、中、低的相关度判定。MyFame 会实时从网上抓取相应网页，按指定结构存顺序放到数据库中，每个网页的存放结构为[6]：

网页编号——5 个字节，记录网页收集的自然顺序号，唯一标记网页语料库中的每一个网页；

相关度类别——1 个字节，取值范围从 1 到 4，表示该网页与该名人实体的相关程度，1 表示高相关度，2 表示中相关度，3 表示地相关度，4 表示未定义，这是兑相关度评价方法进行衡量的重要依据；

类别编号——共 9 个字节，前两个字节，用于标记网页实体的种类，00 表

示名人网页，其它预留，将来可以随着 Fame 的发展，加入如企业、产品等时使用；中间三个字节是名人所在的领域的编号，000 至 006 分别表示政府、科教、媒体、演艺和体育，其它预留；后 5 个字节为名人姓名的编号，随着名人加入的顺序分配，可以区别重名的不同名人实体；

网页 URL——255 个字节；

标志——1 个字节，0 表示训练网页，1 表示测试网页，为便于在使用时分离数据，用于对评价算法进行训练和评测；

网页内容的长度——7 个字节；

网页内容——不定长，每个网页内容完整地记录了网页的 HTML 源代码，不包含原始网页中的图片、视频等内容，面向基于文本内容的挖掘应用。

第二批实体属性信息库和相关网页数据集的制作采用的是同一个实体两个集合共同完成的方式，即同一个名人实体的属性信息和相关网页选择及判定由同一个制作者通过对该实体网页的浏览一同制作完成。制作时间约为 1 位名人/小时。

2.7 本章小结

本章主要介绍了天网知名度系统的工作流程，系统的主要模块—网页搜集模块、网页分析与索引模块、网页评价模块和用户界面模块，以及系统的实体数据集。其中网页搜集模块小节包含了 05 年 5 月份对系统原有流程中网页搜集模块搜集模式的改进，由自主抓取网页，改变为从天网过滤的方式。系统更好地依托了天网搜索技术，提高了存储效率。采用新的网页搜集方法，为系统新加入网页 29 万余篇。实体数据集小节包含了同期系统第二批实体数据集的制作，共新增 8 个领域 150 位名人实体的实体属性信息库和相关网页数据集（3887 篇网页）。目前系统中共有 425 位名人。

第三章 基于概率模型的相关度评价

概率模型是一种经典的信息检索模型，Okapi BM25 则是在此基础之上发展起来的一个比较成熟的实用模型，对词频、文档频和文章长度等有综合的考虑。针对系统中原有相关度评价模型中没有考虑文档频、文章长度等重要因素的不足，本章针对天网知名度系统名人网页相关度评价的特点，构建基本相关度评价模型。构建基础是 Okapi BM25 检索模型，在其基础上引入 HTML 标记权重系数，改进 Okapi BM25 公式，弥补其没有考虑 HTML 标记的不足。在模型的实现中，针对名人网页特点，采用信息提取与基本相关度评价模型相结合的方式，形成具有天网知名度系统特色的基于概率模型的相关度评价基本模型。同时，由于概率模型本身自然支持用户反馈，是后面第四章讨论用户反馈的基础，对提供自动的参数调节功能是一个有益的平台。

3.1 概率模型

3.1.1 经典的概率模型

信息检索中经典的检索模型有三个，分别是布尔模型，向量空间模型和概率模型。系统中已有的两个检索模型便是基于前两个模型进行设计的。

Maron 和 Kuhns 首先提出了概率检索模型[14]的思想。文档 d 可以用一个向量表示 $x=(x_1, x_2, ..., x_n)$，其中 x_i =0 或 1 表示第 i 个索引词不出现或出现。对于一个信息需求，每个文档对应概率 $P(rel \mid d)$（文档 d 满足该信息需求的概率）。概率模型的基本思想是对于文档集合中的所有文档按照对于当前信息需求的相关性概率，即 $P(rel \mid d)$ 排序。根据贝叶斯理论可以得到：$P(rel \mid d)=\frac{P(d \mid rel)P(rel)}{P(d)}$。其中 $P(d)$ 是一个先验概率， $P(rel)$ 是由信息需求决定的，因此对于同一个信息需求而言是常数。

Robertson 和 Sparck Johns 基于查询词在相关文档和不相关文档中的分布提出[15]：以$P(rel|d)>P(\overline{rel}|d)$条件作为文档 d 相关的基本条件，$P(\overline{rel}|d)$为文档 d 不满足该信息需求的概率。所有文档按照 $P(rel|d)-P(\overline{rel}|d)$ 差值的大小排序。该差值记作 $g(d)$。 $P(rel|d)$ 和 $P(\overline{rel}|d)$ 由文档 d 中的所有索引词在相关文档和不相关文档中的分布概率综合而得，即：$P(d|rel)=\prod_{i=1}^{n}p_i^{x_i}(1-p_i)^{1-x_i}$，$P(d|\overline{rel})=\prod_{i=1}^{n}q_i^{x_i}(1-q_i)^{1-x_i}$，$p_i$（$q_i$）是第 i 个索引词在相关（不相关）文档中出现的概率。于是 $g(d)=\sum_{i=1}^{n}c_i x_i+C$，其中 $c_i=\log\frac{p_i(1-q_i)}{q_i(1-p_i)}$，$C$ 对于同一个查询需求是常数。 p_i 和 q_i 的得到基于统计，对索引词 i，整个文档集合中的文档数量为 N，其中与查询 Q 相关的文档集合的文档数量为 R，相关文档集中包含该索引词 i 的文档数量为 r。对于文档集合有如下数据：

	Relevant	Non-relevant	Total
$x_i=1$	r	$n-r$	n
$x_i=0$	$R-r$	$N-n-R+r$	$N-n$
	R	$N-R$	N

于是，$p_i=r/R$，$q_i=(n-r)/(N-R)$，$g(d)=\sum_{i=1}^{n}x_i\log\frac{r/(R-r)}{(n-r)/(N-n-R+r)}+C$。

但是在实际的应用中，对于查询 Q，完整的相关文档集合和不相关文档集合都是不可能获得的，因此 p_i 和 q_i 的准确值也是无法获得的。概率模型的通常使用方法，就是在初始检索时使用对 p_i 和 q_i 经验的估计值，在首轮检索之后，通过某种相关反馈的方法选定一个相关文档集合，对完整的相关文档集合进行近似的模拟，通过查询词在该集合中的分布来计算 p_i 和 q_i。

这是就是经典概率模型 BIR 模型[16]（Binary Independence Retrieval Model）。它的优点是，基于贝叶斯概率论原理，利用相关反馈的归纳学习方法，

获取概率函数，有较好的理论基础，通过严格的形式化模型来计算文档与查询词的相关概率和不相关概率。它的缺点是，每个词的权重 x_i 只能为 0 或 1，没有考虑查询词在文档中的频率(tf)、词在文档集合中的频率（df)、文档长度、文档集合平均长度、HTML 标记等后来被证实在检索中具有重要影响的因素。这只是概率模型的一个最基本的原形，在后面讲到的 Okapi BM25 则已经比较成熟，考虑进了很多因素。它的另一个缺点是，它基于了各检索词之间相互独立的假设，但是在实际应用中没有实验明确表明它是个不良的假设[17]。

3.1.2 Okapi BM25

Robertson 等在 Okapi 系统中提出了日臻完善的概率模型计算公式--Okapi BM25[18]公式，在实际中有优良的表现，词频、文档频、文档长度、文档集合平均长度都被考虑其中：

$$\sum_{T\in Q\cap D} w^{(1)} \frac{(k_1+1)tf}{K+tf}\frac{(k_3+1)qtf}{k_3+qtf}+k_2|Q|\frac{avdl-dl}{avdl+dl} \qquad \text{(Okapi BM25)}$$

其中，Q 为查询，包含索引词 T ，D 为文档，

$$w^{(1)}=\log\frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)},$$

N 为文档集合中文档数量，n 为包含索引词 T 的文档数量，R 为与该信息需求相关的文档数量，r 为与该信息需求相关文档中包含词 T 的文档数量，tf 是查询词 T 在文档 D 中的出现频率，qtf 是查询词在查询 Q 中的出现频率，dl 为当前文档的长度，$avdl$ 为文档集合中文档的平均长度，$K=k_1((1-b)+b*dl/avdl)$，k_1，k_2，k_3，b 是可调参数。

只出现在少量文档中的词显然要比出现在大量文档中的词价值更高，Okapi BM25 中公式中的 $w^{(1)}$ 值可以体现出来。一个词在一篇短的文档和一篇长的文档中出现同样多的次数,那么显然该词在前者中的价值更高，公式中文档长度 dl 体现了对这种情况的考虑。对考虑文档长度因素时，Okapi 模型中使用的是文档长度

与文档集合平均长度的比值（$dl/avdl$），这样做的一个优点就是文档长度的单位，用词还是字来计算都不会有太大的影响，本系统中以索引词为单位计算长度。

文档的这些因素综合考虑到一起，可以有很多种的组合，Okapi BM25 是 Okapi 在 TREC 中验证的最有效的一个算法。调整参数k_1可以改变词频对词权重的影响程度，Okapi 通过实验，发现k_1=2 时是有效的，对初轮检索是一个比较安全的值。更高的k_1值会提高tf带来的影响，k_1=0 时则会消除掉tf的作用。如果b=1，那么假设是文档长度长是因为其中的内容重复，0 的假设是它们长是由于网页内容是多主题的。于是把b设定为靠近 1，比如 0.75，就会在文章冗长重复的情况下降低词频的作用力，当b=0 时，就没有长度调整的作用了，长的文章更占有优势。

上面这个公式保证了，首先，词频的作用不会太强（tf增加两倍，该词的权重不会增加两倍），其次，对在一篇长度等于平均长度的文档中出现一次的词，它的权重就是它的文档频。文档D的总的相关度分数就等于在该文档中出现的所有查询词的分值之和。文档按照这个分数排序，并按从高到低的顺序显式给用户[19]。

以上讨论的是 Okapi BM25 检索模型在普通信息检索中的一些优点。下面在它的基础上构建 Fame 系统中的名人网页相关度评价基本模型。

3.2 Fame 系统名人网页相关度评价基本模型

查询输入、文档表示和相关度评价是信息检索模型的三个基本方面。在天网知名度系统中，查询是用户在注册时填写的名人属性信息，系统中为了对其进行更好的描述建立了名人实体模型，包括 8 类属性信息：领域、姓名、工作单位、职业/职务、兼职、社会形象、特征词、代表作；文档是系统收集的网页；本文基于概率模型，通过改进相关度评价算法和相关反馈来提高名人网页评价的性能。

3.2.1 基本模型

Fame 系统中在 Okapi BM 25 基础上构建 Fame 系统中的名人网页相关度评价基本模型。首先对 Okapi BM 25 公式进行改进。

HTML（Hypertext Marked Language），即超文本标记语言,用 HTML 编写的超文本文档称为 HTML 文档，从 1990 年起 HTML 就一直被作为 WWW 的信息表示语言。HTML 定义了一系列标签来提供丰富的标记信息，标签是一类指令符号，可以控制文档在浏览器中的输出效果，用“ <标签名　属性> ”来表示。

WWW 网页检索源于纯文本检索，WWW 网页与纯文本检索的一个区别就是上面提到的，网页中含有丰富的标记。所以在对网页进行相关度评价时，应该充分利用丰富的 HTML 标记。

天网知名度系统中主要对网页中的文本信息进行检索，以此把 HTML 规范中的标记分为两类：第一类，与网页文字显示有关，控制其显示位置或特点等，如<TITLE>、<H1>…<H6>、<CITE>、<ANCHOR>、<STRONG>、<FONT>等；第二类，与网页文字内容的显示无关，如<IMAGE>、<TABLE>、<FRAME>等。其中，第一类标签中隐含了一些对于相关度评价有益的信息，比如文字内容在标题中出现，意味着对该段内容有较重要的意义；粗体类、大字号也都表现出了网页作者对其的强调和突出，这些信息都是对网页评价的一些良好的提示信息，应充分利用。第二类标记则对本系统检索的影响较小，不予考虑。

但是 Okapi BM25 公式中没有对 HTML 标记的考虑，作者在该公式中为词频 tf 引入一个 HTML 权重系数 w_{html}，改进这个不足。该参数的引入目的在于充分挖掘 HTML 标记有利于名人检索的潜在信息。同时由于 k_1 和 K 的存在，w_{html} 的成倍增长并不会使整个词的权重成倍增长，所以不会过于强调 HTML 标记的作用。于是得到名人网页相关度评价的基本模型（简称模型 3）：

$$\sum_{T\in Q\cap D} w^{(1)} \frac{(k_1+1)w_{html}tf}{K+w_{html}tf}\frac{(k_3+1)qtf}{k_3+qtf}+k_2|Q|\frac{avdl-dl}{avdl+dl} \qquad \text{（模型 3）}$$

天网搜索引擎的相关工作中通过大量的分析和实验，对于 HTML 标记的权重形成了一套参数取值方案[20]，模型 3 中 HTML 标记的权重分配以天网中的这套方案为参考。主要如表 3-1 所示。文档中的每个词，首先被赋予一个初始权重 w_0，如果一个词还被上述多种标记所标识，那么该词的 HTML 部分的权重就由这多种标记的权重（$w_t(tag_i)$ 表示 HTML 标记 tag_i 的权重）综合影响，例如，<big>

网络 </big>，在这种情况下，“网络”一词的 HTML 权重为：$w_{html} = w_0 + w_t(big) + w_t(b)$。

表 3-1 天网搜索引擎中部分 HTML 标签权重分配

标签	权重	标签	权重
<TITLE>	40	<DT>	4
<CITE>	8	<LI>	4
<EM>	2	<UL>	4
<STRONG>	4	<A>	4
<B>	4	<FONT SIZE=7>	16
<I>	2	<FONT SIZE=6>	12
<BIG>	4	<FONT SIZE=5>	8
<H1>	12	<FONT SIZE=4>	4
<H2>	8	<FONT SIZE=3>	1
<H3>	4	<FONT SIZE=2>	1
<H4>	1	<FONT SIZE=1>	1
<H5>	1	<FONT SIZE=+1>	4

3.2.2 模型的实现

如果按照普通信息检索的方式直接将 Okapi BM25 公式应用到天网知名度系统中，其中存在的问题是：

首先，Okapi 模型中，查询是被作为一个没有结构的词串来处理的，而天网知名度系统中的代表用户信息需求的“查询”是实体属性信息，这并不是一个无结构的词串，是划分了 8 个类别的属性信息。（该问题在第四章中解决。）

其次，天网知名度的特殊服务需求，要求实体名的识别达到 100%，如果有的人名没有识别出来，那么检索其相关网页就无从谈起了。因此在进行相关度评价之前，必须确保这些词能够 100%准确的识别。因此相关度评价流程不能像普通检索一样，需要对网页进行特殊的预处理。

再次，网页中出现该名人的姓名，是该网页与该名人相关的基本条件，这是 Fame 系统中的一个独特要求。而直接对网页使用 Okapi BM25 模型，很可能返回的某些“相关”网页中根本连实体名都没有出现过，更谈不上跟该名人相关，这个问题是由 Okapi BM25 的特点决定的，它采用的排序依据是所有词的一个综合

权重，不能依据其中的某个词做过滤，所以要在使用该公式之前，先用人名对网页进行过滤，才能保证 Okapi 模型评价符合 Fame 系统的基本要求。

基于以上问题，首先，在评价之前的网页分析阶段，对网页进行预处理，包括补充词典的使用、命名实体识别和二元关系识别为主的信息提取；然后，用实体名对网页先进行一次过滤以确保后面 BM25 公式评价的准确性；下一步，对于出现了该实体名的网页使用 BM25 公式进行相关度评价。下面主要介绍一下信息提取和补充词典。

3.2.2.1 信息提取

信息提取的任务包括从简到难的五个任务：命名实体--提取文本中的命名实体，实体名包括人名、地名、机构名等，另外还包括日期、时间、货币、百分比等，一般地，命名实体识别就是判断一个文本串是否代表一个命名实体，并确定其类别；实体关系--提取命名实体之间的重要关系(事实)，如人名与职务、机构与处所等；模板脚本--提取指定的事件，包括参与这些中的个体实体、属性或关系，如时间、地点、施事、受事、时间类别等信息组成的有关事件的描述；共指消解—对代词、名词等的共指分析，解决名词性短语的指代关系；模板合并—把相同的事件的模板合并为一个模板。在中文信息提取方面，目前，共指消解和模板合并还是信息理解的难点，仍处于研究阶段。但前三项任务效果比较好，特别是命名实体识别关系识别已经达到 90%以上准确率，可以达到实用水平[6]。

根据天网知名度系统进行名人检索的独特特点，进行网页的相关度评价之前，在网页分析模块，采用了命名实体识别、人物-职位和人物-单位二元关系识别等一些信息提取中较为成熟的技术对原始网页进行了分析处理，对网页中出现的实体名串、实体二元关系串都作了分类组织。该模块由北京大学计算语言学研究所研究开发。命名实体的识别和二元关系的提取都为进行相关度评价时提高准确度奠定了基础。

同时要说明的一点是，新的 Fame 系统中，从天网过滤网页时，是以人名为查询输入，从天网查询端获得返回列表的，天网目前没有使用命名实体识别功能，因此有一些人名在天网的切词模块处理后，是被切散的，没有被识别为一个词，但是这并不影响过滤，因为在天网中虽然没有被切为一个词，但是只要包含了姓名中所有字的网页都是会被返回的，同时由于天网中考虑了查询词在网页中出现

时相互间的距离，所以基本是可以满足过滤需求的，不会过滤不到网页，也不会过滤到太多无关网页。

3.2.2.2 补充词典

尽管网页分析时用以切词的基本词典已经包含九万多词条，并且在切词和信息提取中采用了规则和统计的方法对词典中未收录的词进行了识别，但是仍然不能够覆盖多变的民实体属性信息。比如，一些包含高频词的人名，科教领域名人的一些研究领域等专业化比较高的词语、演艺类领域名人如歌手多样化的代表作（歌名等），还有一些属性信息是比较长的词，超过基本词典的 4 个汉字的词长限制。

代表用户信息需求的名人实体属性信息，是对网页进行相关度评价的基本依据。而由于不能对这些词实现准确的识别，不少的实体属性信息在分词之后被切散了，变得支离破碎。对相关度评价准确性有很大影响。要保证网页相关度评价的准确性，首先要保证对实体属性的识别具有较高的准确率，其中，对人名的识别要求要达到 100%的准确，因为网页中出现名人的人名是该网页与该名人相关的基本条件，

为满足这种需要，Fame 系统以名人实体属性信息为依据，构建用户信息的补充词典。该词典优先于基本词典，首先对网页文本进行长词绑定，保证实体属性的完整性，提高网页信息与实体属性相关度评价的准确性[6]。

3.3 系统评测方法

对 Fame 系统的名人相关网页数据集中收录的网页，划分了三个相关度等级——高度相关、中度相关和低度相关。在对相关度评价算法进行评测时，依据的就是这种多等级的相关度判定。

3.3.1 系统中原有的评测方法

对于相关度评价模型质量的评定，系统中原有的评测方法[6]（后文简称评测方法 1）是：

对于有关实体网页的相关度，根据算法评价的结果与人工评判的结果（人工对实体相关网页进行了高、中、低三种不同相关度的判定）进行比较评测。基本原则是网页与实体的相关度分值应该满足下列关系：

[high]>=[mid]　　[high]>=[low]　　[mid]>=[low]

如果定义已知网页集合中相关度为高、中、低的个数分别为：

|high|=m；　|mid|=n；　|low|=k

那么总的关系个数为：Total= m*n + n*k + k*m

评价指标是有关实体网页相关度排序正确的关系数目与总的关系数目的比值。

原系统中采用了微平均和宏平均来表示总体效果：微平均是关于名人实体网页的所有相关度关系的准确率，是对所有关系的算术平均，其中相关网页多的名人的相关度评价队最终结果的贡献较大；而宏平均是先计算每个名人实体相关网页的相关度评价关系的准确率，再求这些准确率的算术平均，其中相关度评价准确率高的名人实体的相关度评价对最终结果的贡献较大。

3.3.2　DCG 评测方法

评测方法 1 对所有的相关网页都是同等重视的。但是一个实用检索系统的质量更应该由它检索高相关度网页的能力决定。本文中引入更符合用户的检索感受的 DCG 评测方法。

3.3.2.1 方法介绍

Järvelin 和 Kekä lä inen的论文[21]中提出了 DCG(Discounted Cumulative Gain)评测方法，该方法最主要的优点是可以把多等级的相关度判定结果统一处理，而不是只针对二元的相关度评价。大多数评测方法仍然是基于对网页二元的相关度判断的，即认为一个网页或者相关或者不相关。对进行多等级相关度评价的或连续相关度评价的方法不易进行评测。而 DCG 评测方法可以把多等级的相关度等级判定融合到一起，来对相关度评价算法进行评测，非常适合用来对天网知名度系统进行评测。

DCG 评测方法的主要思想是，对一个查询需求，用一个向量V来表示相关度评价方法返回的排序结果，$V =< G[1], G[2], G[3]... >$，其中$G[1]$表示返回队列中排

序为第一位的文档，$G[2]$表示第二位的文档,依次类推。在该方法中，通过对不同等级的相关度文档赋以不同的权重,来表示对它们的不同重视程度。假设对高、中、低相关度文档分别赋予权重 H、M、L。分别用各文档对应的权重来取代它在向量V中的位置，就可以得到新的向量V。用下面这个公式来对这个向量来计算这次排序结果的 DCG 值：

$$DCG[i]=\begin{cases} G[1], i=1 \\ DCG[i-1]+G[i]/\log_b i, otherwise \end{cases} \quad \text{（DCG 公式）}$$

其中的参数b可以用来控制排在后面的网页权重的衰减速度。在本次实验中b的值取 2，排在最前面的两个网页是不受惩罚的。在本系统中，出现最理想的 DCG 值的情况是，所有高相关的网页都排在中相关的网页之前，所有中相关的网页都排在低相关的网页之前，可以以此序列构造理想V向量，并计算出最理想的 DCG 值。取实际 DCG 值与理想 DCG 值的比值作为系统评测指标。

DCG 评测方法与评测方法 1 的主要区别是，评测方法 1 中高、中、低相关度网页是同等重视的，而 DCG 评测方法中，首先，由于公式中$\log_b i$项的作用，排在前面的网页的对该指标贡献更大；其次，由于对高、中、低相关度文档赋予了不同的权重，更重视的是高相关度网页的排序质量。DCG 评测方法的这两个特点使它更符合用户浏览相关网页时的习惯。

高、中、低网页赋予的权重会影响理想 DCG 值和当前排序结果的 DCG 值．权重的设置体现了对高中低网页的态度。当高、中、低网页的权重变化时，理想 DCG 值也是变化的。从实验中看出，随着高中低网页权重的变化，系统中已有的三个评价模型（模型 2、Okapi BM25、模型 3）之间的相互位置会发生一些变化。下面通过实验分别给高、中、低相关文档附以不同权重比例，考察 DCG 评测方法的表现。首先固定中、低相关文档的相关度权重比值,调整高相关文档的权重。然后固定高中相关文档的权重比值，调整低相关文档的权重。最终确定本系统中采用的权重比例。

3.3.2.2 参数选取

本节通过对不同等级相关文档赋以不同权重进行实验，通过区分度和稳定性两个方面来选择合适的权重参数。

当不同等级相关文档的权重比例发生变化时，理想 DCG 值也会发生变化。当赋予高度相关、中度相关和低度相关的网页同等权重时，每个相关网页都会影响到 DCG 值，当赋予高相关度网页的权重比中、低相关度网页高很多时，一旦高相关度的网页被检索到以后，DCG 值就很快达到最高点。

在文献[22]中，比较了高相关文档和一般相关文档（该文献中相关度等级分为两级：高相关和一般相关）分别取不同权值时 TREC 排序结果的变化，本文中给出了分别赋予高、中、低相关度文档不同权重比例时对三次检索结果（模型 2、Okapi BM25，模型 3）的评价情况，如图 3-1 至图 3-7 所示，图中 A:B:C 表示高、中、低相关度网页的权重值为别为 A、B、C。

从图中可以看出，三条数据线的相对位置是有变化的，说明相关度评价质量的排序会随着对不同等级相关网页的重视而发生变化。从图中可以看出，当所有相关的文档基本同等对待（图 3-1）和仅稍微提高高相关度网页的权重（图 3-2、图 3-3）时，评价结果基本相当，但对于三种评价方法的区分度比较小；当给高相关的文档高一些的权重（图 3-4 至图 3-6）时，曲线形状比较相似，排序结果比较稳定而且有一定的区分度；当赋予高相关的文档比特别大的权重（图 3-7）时，排序结果与其它曲线差别较大，这是因为当不同相关度等级文档赋予的权重比例很大时，DCG 评测方法基本被高相关度文档主宰了，基于非常少量的一批文档来对评价算法做评测的，使评测结果的可靠性下降，评测方法变得不稳定。

根据文献[22]中的建议，使用 DCG 时，两相邻的相关度之间选用小一些的比例（如 3：1、5：1），可以把所有相关文档都考虑进来提高它的稳定性，并且同时仍然对检索了高相关文档的系统以奖励。本系统中使用 DCG 评测方法时，高、中、低相关度文档的权重比例选取 100:20:4（图 3-7），排序结果比较稳定而且有适当的区分度。

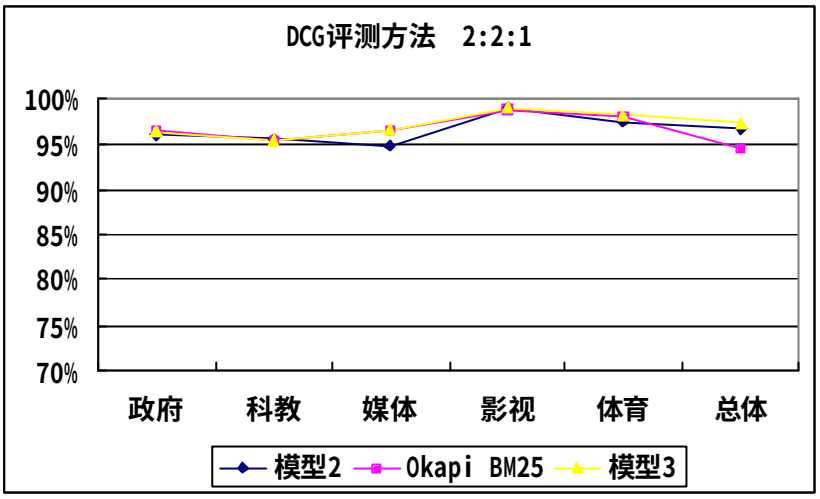

图 3-1　DCG 2:2:1

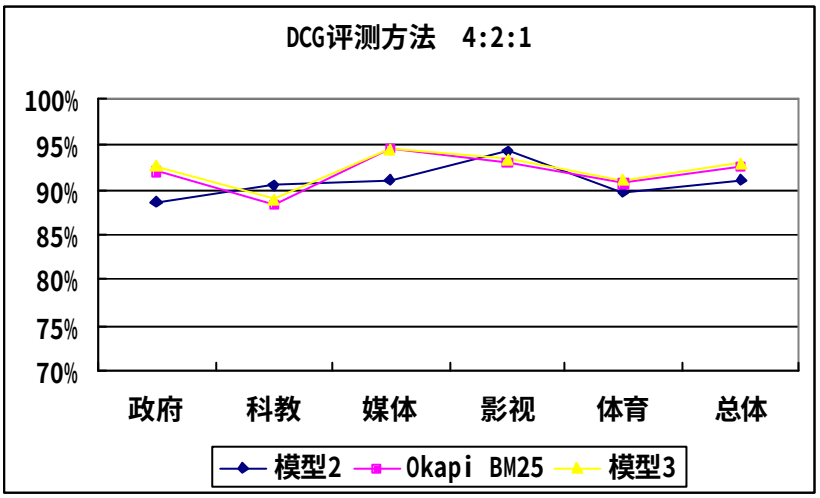

图 3-2　DCG 4:2:1

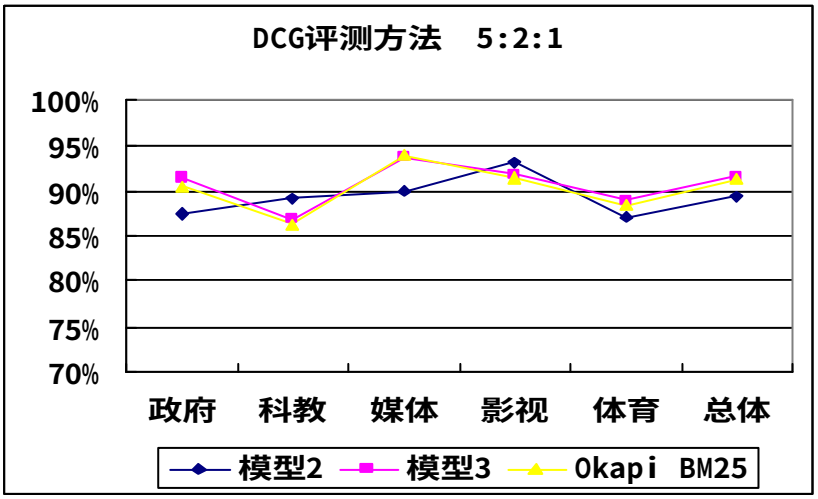

图 3-3　DCG 5:2:1

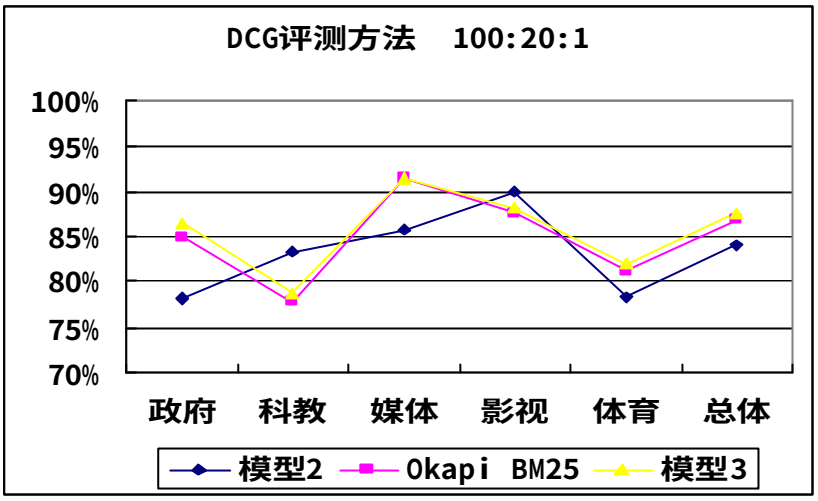

图 3-4　DCG 100:20:1

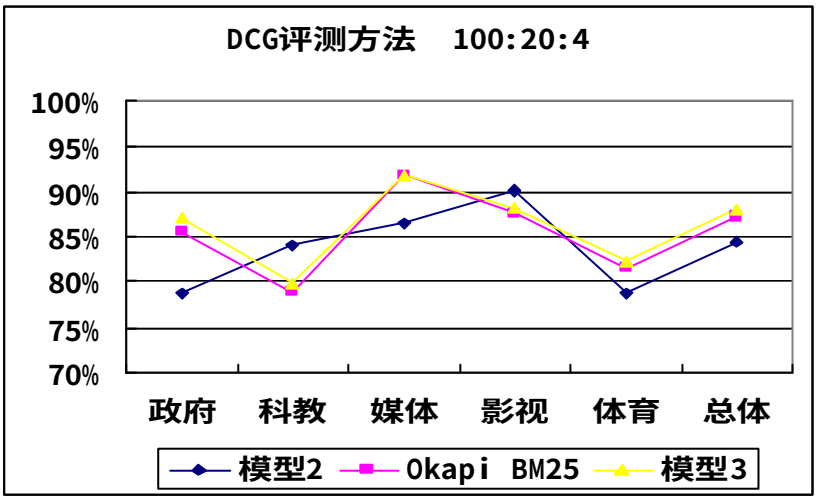

图 3-5　DCG 100:20:4

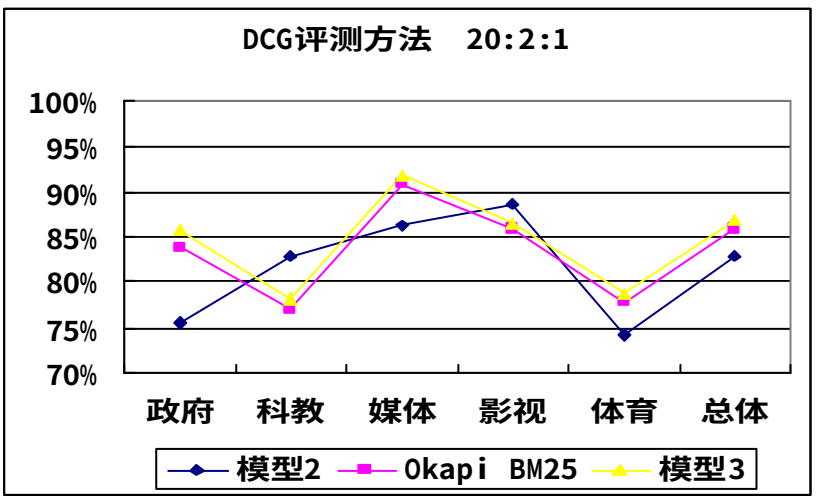

图 3-6　DCG 20:2:1

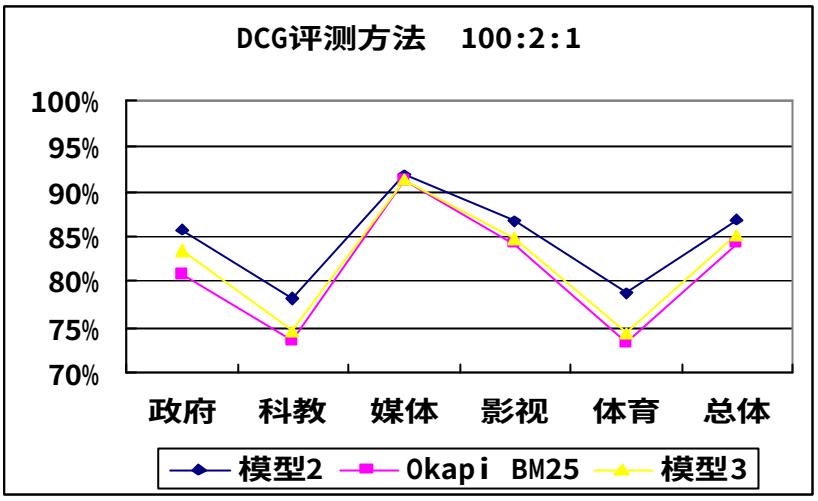

图 3-7　DCG 100:2:1

3.4 基本模型性能评测

3.4.1 实验设计

分别对系统网页集合进行三组相关度评价，

第一组：不采用补充词典，对原始网页库进行命名实体识别、二元关系识别、分词等网页分析处理，先以人名进行过滤，不考虑了 html 标记，使用原始的 Okapi BM25 公式进行评价；

第二组：在系统中采用补充词典，然后对原始网页库进行命名实体识别、二元关系识别、分词等网页分析处理，现以人名进行过滤，不考虑 html 标记，使用原始的 Okapi BM25 公式进行评价；

第三组：在系统中采用补充词典，然后对原始网页库进行命名实体识别、二元关系识别、分词等网页分析处理，现以人名进行过滤，考虑了 html 标记，使用模型 3 进行评价；

第四组：在系统中采用补充词典，然后对原始网页库进行命名实体识别、二元关系识别、分词等网页分析处理，使用原有系统中的 CVSM 模型进行评价；

用第一组评测结果与第二组评测结果对比，比较加入补充词典前后系统相关度评价质量的变化；用第二组评测结果与第三组评测结果对比，加入 HTML 标记权重系数前后系统相关度评价质量的变化；用第三组评测结果与第四组评测结果对比，比较在评价中对词频、文档频、文档长度和 HTML 标记进行综合考虑后的

模型 3 比系统原有模型 2（CVSM）对评价质量的提高。

使用 Okapi BM25 公式时，参数选取根据 Okapi 研究工作的建议，$k_1 = 1.2$，$b = 0.75$，$k_2 = 0$，$k_3 = 0 \sim 1000$。本系统中，选取 $k_3 = 10$，初始时没有相关文档集合，因此 r、R 都未知，根据使用概率模型的常见方法，取 $r = R = 0$，于是索引词的初始权重为：$w^{(1)} = \log \frac{N - n + 0.5}{n + 0.5}$。

3.4.2 实验结果及分析

第一组评测结果与第二组评测结果对比如图 3-8 所示；第二组评测结果与第三组评测结果对比如图 3-9，3-10 所示；第三组评测结果与第四组评测结果对比如图 3-11，3-12 所示。

从图 3-8 的评测结果曲线对比中可以看出，补充词典的加入，使用户属性信息在后面的网页分析中保持了完整性，使用户的信息需求与网页内容的匹配更加准确，提高了系统相关度排序质量。

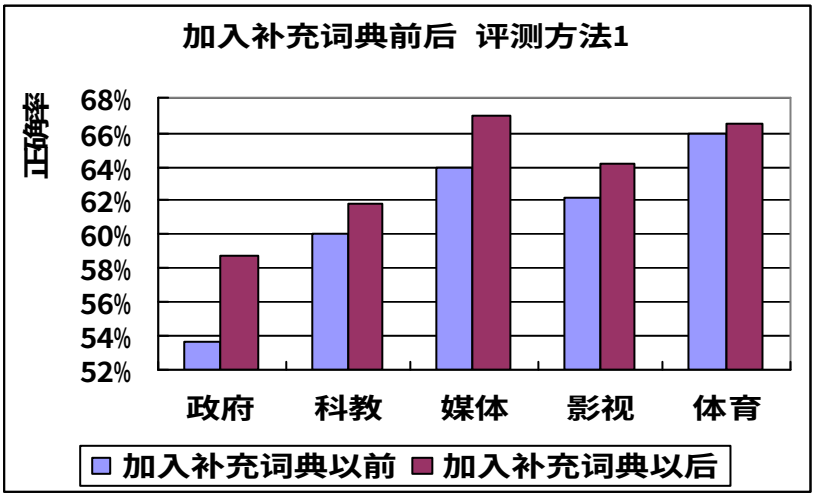

图 3-8　Okapi 模型在使用补充词典前后

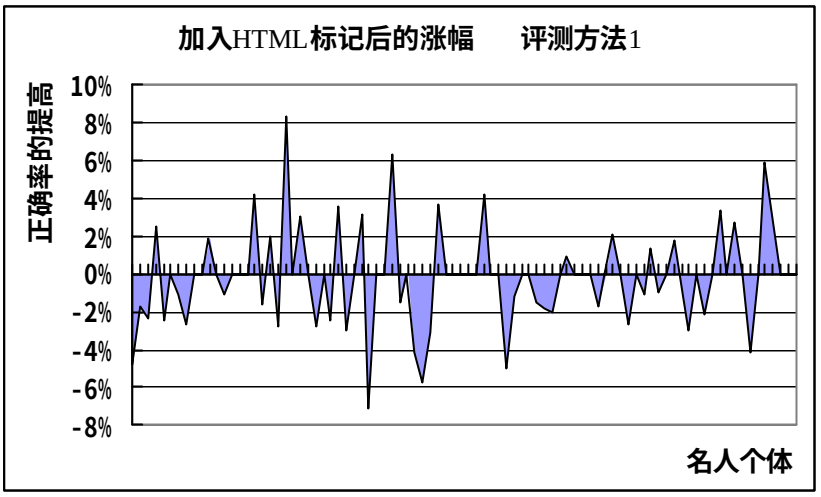

图 3-9　加入 HTML 标记后的涨幅　（评测方法 1）

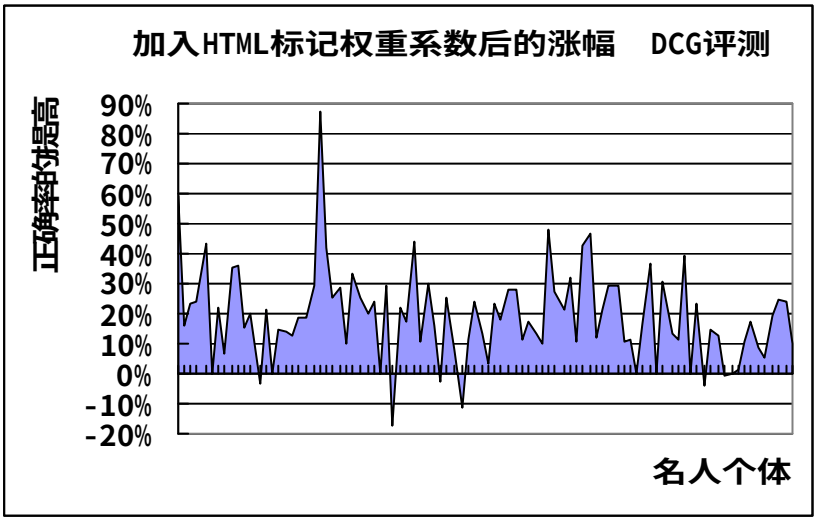

图 3-10　加入 HTML 标记权重系数后的涨幅　（DCG 评测）

图 3-9 和图 3-10 中均以名人个体为单位。纵轴均为在评测指标下提高的百分点，是实际值，而不是相对值，下文皆同。使用评测方法 1 得到的图 3-9 显示，加入 HTML 后，系统性能没有明显提高，名人实体的相关度评价质量有涨有跌。

图 3-10 使用 DCG 评测方法的结果曲线中则清楚地表现出加入 HTML 标记后，相关度评价性能提升显著。

两种指标评测结果的显著差异，反映出了，网上有相当数量的网页仍有很大的不规范性，相关度高的网页往往比较规范，因此加入 HTML 标记的考虑后，确实挖掘出了一些潜在信息，但是对于相关度低的网页，其网页的不规范性因素，也使得利用 HTML 标记进行潜在信息挖掘的工作变得无所适从，因此在图 3-9 中增减各半。

由于 Fame 是一个实用系统，所以更重视高相关网页的排序情况，根据加入 HTML 标记后用 DCG 评测后的优异表现，认为 HTML 标记所隐含的信息对提高本系统的相关度检索质量有明显贡献，模型 3 这种加入对 HTML 标记考虑的相关度评价模型优于基本的 Okapi BM25 公式，提高了系统相关度评价质量。

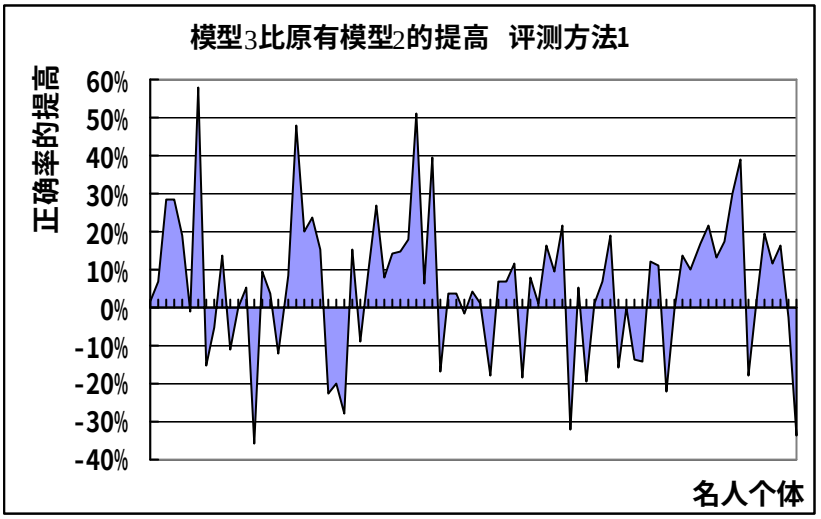

图 3-11　概率 3 比模型 2 相关度评价质量的对比　（评测方法 1）

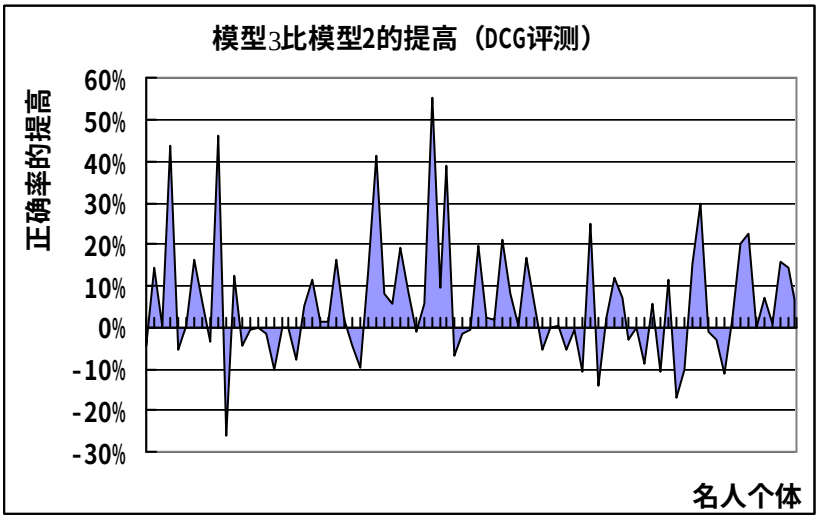

图 3-12　模型 3 比模型 2 相关度评价质量的对比　（DCG 评测）

图 3-11 和图 3-12，均是以名人个人为单位，从这两种指标的评测结果图中可以明显的看到，采用加入 HTML 标记权重系数的模型 3，较明显地提高了系统的性能。大多数实体涨幅为正值。尤其在 DCG 评测指标下，增长趋势更为明显。由此可得，模型 3 在高相关度网页排序方面比 CVSM 的优势更加明显。充分说明首先，在加入补充词典和进行信息提取之后，确保了代表用户信息需求的实体属

性信息的完整性，提高了评价中的准确度；其次，评价过程中综合考虑相关度评价中的多个重要因素，包括词频、文档频、文档长度和 HTML 标记等，提高了系统的相关度评价质量。

3.5 本章小结

本章中实现了基于概率模型的适应 Fame 名人检索特性的相关度评价基本模型——模型 3。首先使用信息提取、补充词典等方法对文档集合进行预处理，并用实体名对网页作首次过滤，然后在 Okapi BM25 检索模型基础上引入 HTML 标记权重系数，改进 Okapi BM25 公式，弥补其没有考虑 HTML 标记的不足。在该基本模型中，综合考虑了影响相关度评价质量的词频、文档频、文档长度及 HTML 标记等重要因素。

在系统中引入了 DCG 评测方法，通过实验为其确定了不同等级相关度文档的权重比例。采用系统原有的评测方法 1 和 DCG 评测方法一起对基本模型进行了相关度评价质量的评测。实验结果表明，HTML 标记权重系数的引入、补充词典的采用都给相关度评价性能带来了提高，模型 3 的评价质量比系统原有模型 2（CVSM）有明显提高。

第四章 支持多层次实体模型的相关度评价

本章将在基本评价模型——模型 3 的基础上，根据系统中各领域实体的具体情况针对领域特点改进相关度评价基本模型，进一步提高系统的相关度评价质量。

4.1 多层次的实体模型

天网知名度系统的原有名人实体模型中，用 8 类属性信息来表示用户的信息需求，分别是：领域、姓名、工作单位、职业、兼职、社会形象、特征词以及代表作。其中特征词一项填写用户对该实体的关注点。以上属性基本涵盖了名人网页的基本特征，系统依据它们对网页进行相关度评价。用户的查询不再像通用搜索引擎一样是个无结构的词串，而是结构化的。原有系统中的实体模型对各领域实体都是通用的，没有对不同领域的实体进行区别。

通过对各领域名人相关网页的人工观察，发现对不同领域的名人，高相关度网页的内容有比较明显的差别，而且有一定规律可循。其中政府类名人的高相关网页往往比较正式，关于其出席某次会议、发表某个讲话等，关于其个人的专门报道比较少，这与其工作性质的要求有关，同时在其高相关网页中，职业、职务的出现率较高；科教类名人的高相关网页内容也比较正规，其职业、职务类属性在其相关网页中出现率较高，而且其高相关网页往往围绕实体的科研领域这个主题，即特征词属性；而媒体、演艺类名人的高相关网页则常常围绕他们的代表作、特征词类属性展开，比如对演艺类名人，主题往往离不开其拍的电影、推出的专辑等。

这 8 类属性信息对不同领域名人相关度评价的贡献有明显差异，因此需要针对不同领域建立不同的名人实体模型，区别各属性信息在相关度评价中的权重。

原有系统中，各类属性并没有根据实体所在的领域进行区别对待。本文中，构建多层次的名人实体模型，对各属性依据领域分配以不同的权重。首先引入领域属性信息权重系数以改进第三章中得到的模型 3，使其支持多层次的名人实体

模型；其次通过针对各个领域分别训练 8 类属性的权重值，以实现对不同领域名人的区别对待，以提高系统相关度评价质量。

4.2 改进的评价模型

把 Okapi 公式中的项 $\frac{(k3+1)qtf}{k3+qtf}$ 记作 M ，该项的主要作用是调整不同查询词对检索结果的影响。qtf 是查询词在查询中的出现频率，k_3 是一个用以增强公式灵活性的参数，Okapi 建议的 k_3 取值范围为 1~1000。对查询中的精华词（也就是对提高相关度评价质量贡献大的词）应附以一个比较高的权重值,使它在相关度评价中起到更大的作用，本章中通过项 M 来实现。

在 Okapi BM25 公式中，用户输入的查询被当作一个无结构的词串，不支持按领域区别属性权重的名人实体模型。作者在基本评价模型——模型 3 的基础上为参数 qtf 引入一个领域属性权重系数 c_{ij} ，i 表示当前实体所属的领域，j 表示当前词所属的属性类。该参数的引入使评价模型支持结构化的名人实体模型，主要用于提高对各领域相关度评价贡献大的“精华”属性的权重，得到模型 4：

$$\sum_{T\in Q\cap D} w^{(1)} \frac{(k_1+1)w_{html}tf}{K+w_{html}tf} \frac{(k_3+1)c_{ij}qtf}{k_3+c_{ij}qtf} + k_2 |Q| \frac{avdl-dl}{avdl+dl} \qquad \text{（模型 4）}$$

4.3 参数的获取

模型 4 中,各领域属性权重系数 c_{ij} 的获取是通过使用训练集进行训练并通过测试集测试的方法而得到的，选取一组使系统性能最优的参数，避免人工赋予的不确定性。

4.3.1 实验设计

k_3的取值范围为 1~1000（根据 Okapi 建议），当固定k_3的值时，M 的值会随c_{ij}*qtf 乘积的增长而增长。通过计算，当k_3值较小时，c_{ij}*qtf 的增长给 M 带来的增长幅度比较小，寻找到理想c_{ij} 的值过程缓慢，当k_3值较大时，c_{ij}*qtf 的增长给 M 带来的增长幅度比较大，也不易于找到理想的c_{ij} 值。因此实验中，选取k_3=10，c_{ij}*qtf 的增长给 M 带来的增长幅度适中，增长幅度量化如下（以 qtf =1 为例）：

$$M=\begin{cases} 1, & c_{ij}=1 \\ 1.8, & c_{ij}=2 \\ 3.1, & c_{ij}=4 \\ 4.1, & c_{ij}=6 \\ 5.5, & c_{ij}=10 \end{cases}$$

分别对每个属性类，依次赋值c_{ij}=2、4、6、10，同时对其它类属性都赋值c_{ij}=1，进行 8 组相关度评价。由于训练集中文学领域和业界领域实体样本偏少，不具备训练条件，所以这次训练剔除了这两个领域，即该领域的实体属性信息的c_{ij} 值仍取 1，不变。

4.3.2 参数的训练

对 8 类属性，依次赋值c_{ij}=2、4、6、10。因为初试探测时，当c_{ij} 增长到 13 时，基本上之前所有的相关度质量涨幅都开始回落，从下图中也可以看到有一些属性当c_{ij} 增长到 10 时，就已经开始回落了，说明不同类属性间的区别对待是有一定限度的，其中任何一类属性信息都不能取代所有类属性的共同作用。说明这几类信息都是需要综合考虑的。因此下面实验中c_{ij} 最大值取到 10 是充分的。

当训练一类属性的权重系数时，固定其它类属性的权重系数 c_{ij} 为 1，于是得到 8 组相关度评价结果。训练集与测试集的划分方法是：每个领域内的名人平均分为两个部分，一部分作为训练集另一部分作为测试集。使用评测方法 1 进行评测，结果如图 4-1 至图 4-7 所示（下面只列出微平均时的情况，宏平均与其稍有差别，但走势基本一致，文中不再一一列出）。

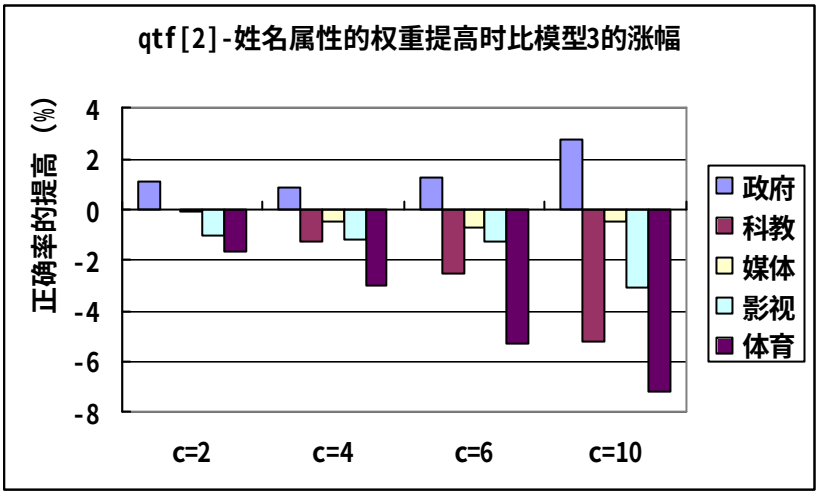

图 4-1　姓名

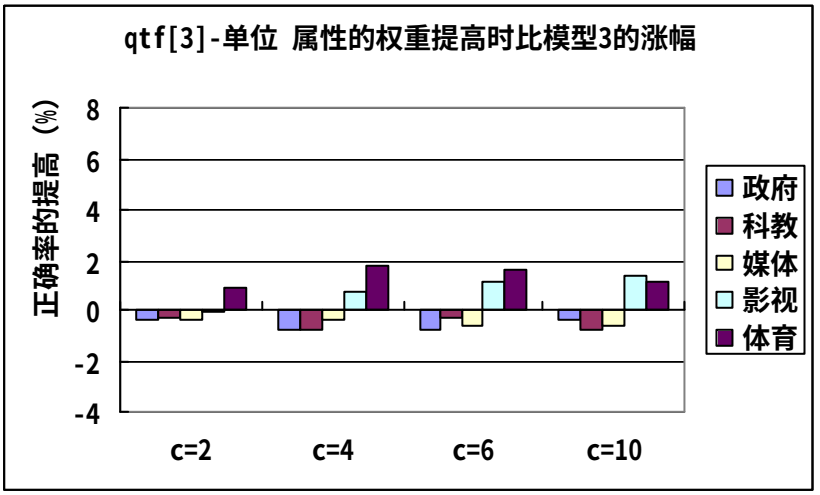

图 4-2　单位

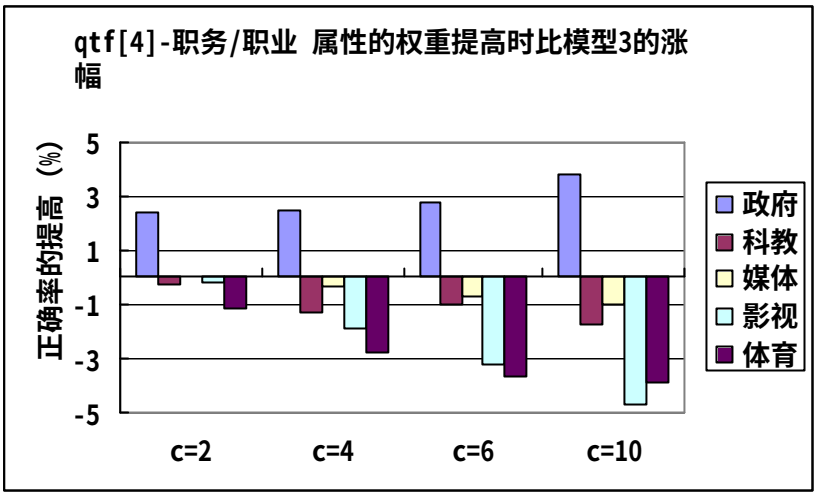

图 4-3　职务

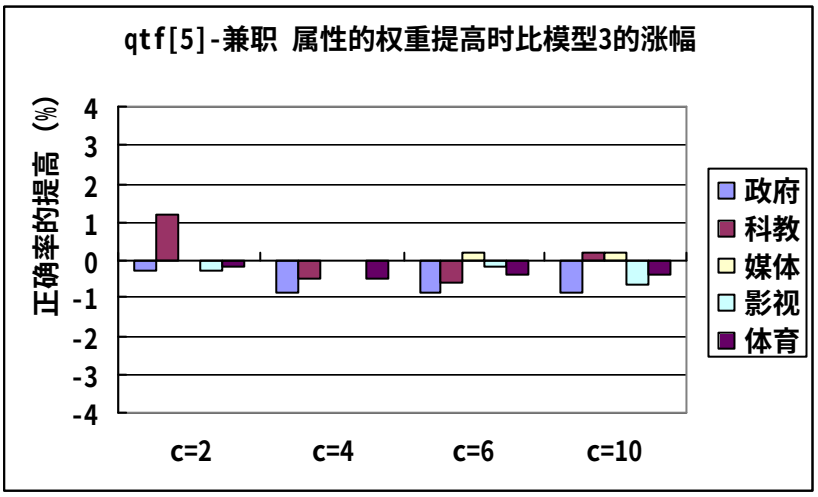

图 4-4　兼职

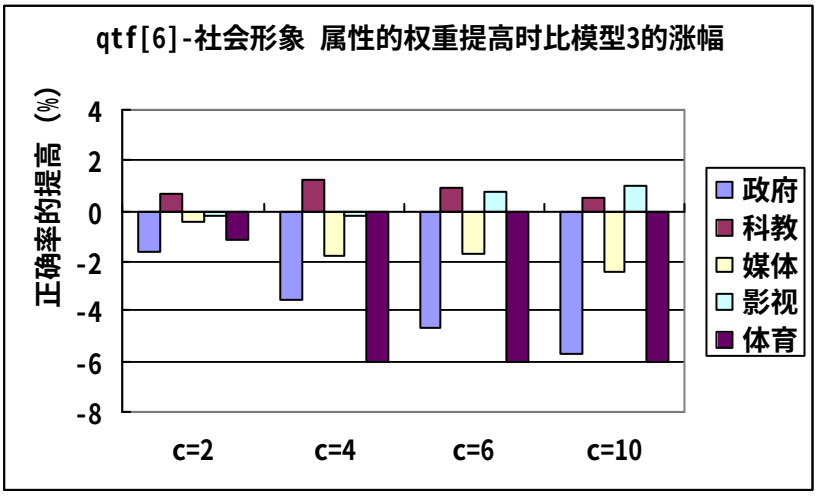

图 4-5　社会形象

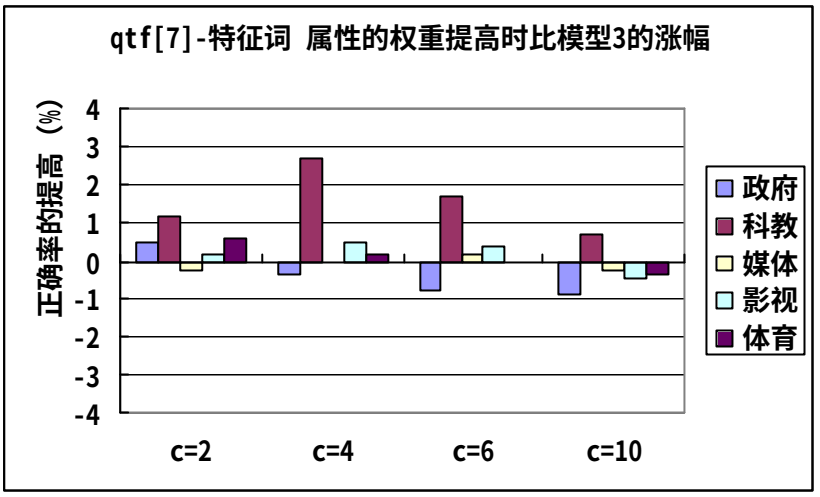

图 4-6　特征词

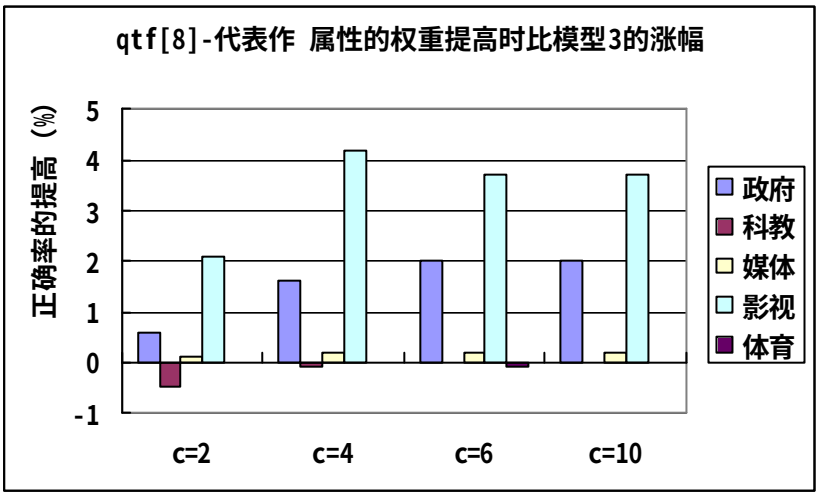

图 4-7　代表作

4.3.3 参数的选定和测试

从上面 8 组评价结果中可以看到，每类属性对相关度评价质量的贡献并不是随着 c_{ij} 的增长而持续增长的，会有一个波峰，那么在为各领域实体确定属性权重系数时，选择出现波峰时的 c_{ij} 值。同时可以看到当某类并非“精华”的属性被提高权重时，对相关度评价质量造成的削减会非常显著，因此在确定具体的 c_{ij} 值时，遵循一个原则：对相关度评价质量带来的涨幅相同的情况下（对正确率保留小数点后两位，采用四舍五入原则）选择小的 c_{ij} 值，这样有利于长远的

评价发展。另外对实验结果还进行了详细分析，由于拥有相关网页多的实体对微平均贡献大，相关度评价结果高的实体对宏平均贡献大，因此综和考虑了微平均和宏平均的情况。

依据以上试验结果和分析，对模型 4 中的领域属性权重系数赋值如下：政府类名人的姓名、职业属性的 c_{ij} 值为 10；科教类名人的兼职属性——4，社会形象、特征词属性——4；媒体类名人的特征词属性——6，代表作属性——4；演艺类名人的特征词、代表作属性——10；体育类名人的单位属性——4，特征词属性——2。对应到 Okapi 公式里面的项 M，c_{ij}=2 对应初始权重（c_{ij}=1）的 1.8 倍，c_{ij}=4 对应 3.1 倍，c_{ij}=1 对应 4.1 倍，c_{ij}=10 对应 5.5 倍。

采用上述训练得出的 c_{ij} 值，采用模型 4 使用测试集，对文档集合进行相关度评价，用评测方法 1 评测结果如下：

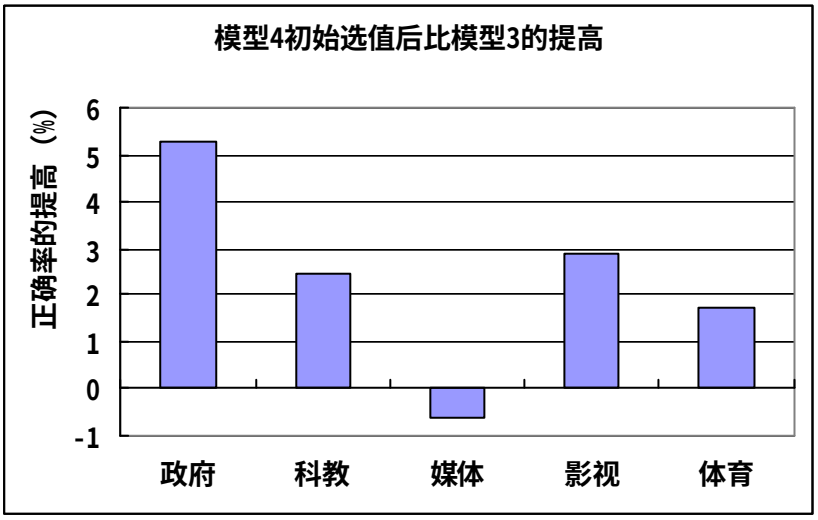

图 4-8　模型 4 初始选值后比模型 3 的提高

上图中是直接把有各属性的波峰值组合起来确定为该领域的实体模型，大多数领域评价结果有了比较大的改进，尤其政府、业界、体育领域的提高大过于单独提高其中任何一类属性带来的提高。但是同时发现，媒体类和演艺类名人使用综合的属性权值后，媒体类略有下降，演艺类虽有提高，但是没有单独提升属性权值时评价结果好。说明多类属性之间的关系对评价结果是有一定影响的。分析其中原因，政府、业界、体育领域，选择增加权重的每个属性都是比初始结果提

高较明显的属性，说明当每个属性都对评价结果有较大的贡献时，在一起使用可以取得更好的结果。而对媒体类名人，特征词和代表作两类属性分别使用时，都是对结果略有改进，程度相对不很明显。演艺类名人代表作类属性对结果贡献较大，特征词属性相对贡献较小。通过实验得出，如果同领域内有贡献比较明显的属性存在时，那些对评价结果有贡献但是比较不明显的属性，不应予以考虑；如果同领域内几个属性都属于有相对不太明显的贡献，则只选择其中一个贡献相对较大的属性提高权重。即：一般< 优+一般< 优，任一个优< 优+优。所以在后面的取值中，去掉了媒体代表作属性的权重增长和演艺类特征词类权重增长。重新定义后，各领域的属性信息权重系数 c_{ij} 值如表 4-1 所示。

表 4-1　c_{ij} 值

	姓名	工作单位	职业/职务	兼职	社会形象	特征词	代表作
政府	10	1	10	1	1	1	4
科教	1	1	1	2	4	4	1
媒体	1	1	1	1	1	6	1
演艺	1	1	1	1	1	1	4
体育	1	4	1	1	1	2	1

采用上述 c_{ij} 值，使用测试集，对文档集合进行相关度评价，并用评测方法 1 进行评测结果如图 4-9 所示（领域内微平均）：

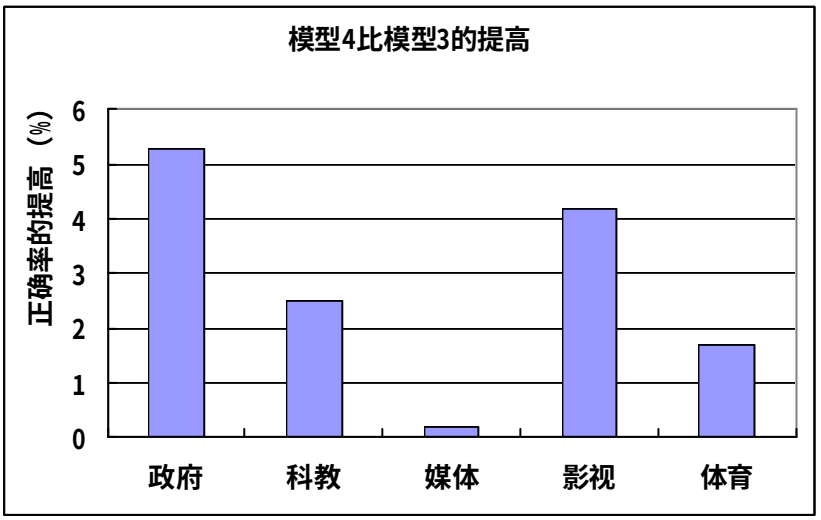

图 4-9 模型 4 比模型 3 的提高

可以看到使用测试集进行评价后结果有较明显的提高。各个领域都是系统现有评价模型中效果最好的。与区别领域前的模型 3 比较，政府类上涨 5.2 个百分点，上涨幅度最大，影视、科教、体育类分别上涨 4.1、2.5、1.8 个百分点，媒体类涨幅最小，0.2 个百分点。因此训练出的属性信息系数是有较好的适用性的。

4.4 本章小结

由于不同领域名人的属性信息对其相关度评价有不同的贡献，本文中构建了区别名人实体领域的多层次实体模型，来更好的描述用户的信息需求。同时在基本模型—模型 3 基础上引入属性信息权重系数 c_{ij} ，对相关度评价模型作进一步改进，从不支持结构化查询需求改进为支持多层次实体模型，得到相关度评价模型 4。各领域的权重系数 c_{ij} 通过训练集训练的方式来获得，避免了人工赋予的不确定因素。最终选取对系统相关度评价质量提高最大的一组权重系数作为模型中的领域参数，该套参数通过测试集的测试，证明有较好的适用性。模型 4 比基本模型——模型 3 提高了系统的相关度评价质量。

第五章 相关性反馈

前面已经提到，因为目前检索系统的召回率无法达到 100%，因此在使用概率模型时完整的相关文档集合和完整的不相关文档集合是不可得的。所以初始检索中，查询词在相关文档和不相关文档集合中出现的概率，是一个经验的估计值。但是相关反馈的方法可以提供一个对这些集合的模拟，提供概率模型中的权重参数的近似估计。

如上一章中所述，天网知名度系统中用户的信息需求是由一个多层次的名人实体模型来表示的。针对不同领域调节各类属性的权重，使整个领域的平均评价结果得到提高。本章中引入相关性反馈，通过反馈的方法，使系统在用户的使用中，实现权重参数的自动调节和相关度评价质量的进一步优化。

5.1 相关性反馈理论

通过对初始信息检索结果的判断，对查询进行重构时，主要有两种方法，一是扩展初始查询，二是调整查询中词的权重。就具体方法而言，主要有三类[17]，一类是基于用户的反馈信息，第二类是基于从初始检索文档中提取到的信息，第三类是基于整个文档集合的一些信息。后面两种方法主要是基于分类算法。在本系统中，查询对应的就是用户属性信息。系统中采用的是调整查询中词的权重的方法，具体方法采用的是第一类方法。

相关反馈是第一类方法中的一个很重要的方法。相关反馈的研究有比较长的历史。1965 年 Rocchio 提出使用相关反馈来进行查询重构[23]。因为用户是可以很容易判断一篇文档是否与他的信息需求相关的，相关反馈在此推动之下发展起来。利用用户进行判别后做出的相关或不相关的判断，系统可以自动的优化查询--向原始的查询中加入一些新的查询词或者调整原来查询中查询词的权重。当使用查询扩展的方法时，基于相关的网页集合根据某种排序公式计算一批新词的排序值序列，并且把排序最高的 n 个词加入到原始查询中。当使用调整原来查询中查询词权重的方法时，如果一个查询词在所有被用户判定为相关的文档中都出

现，那么说明该查询词在提高相关度评价质量方面有贡献，应该提高它在查询中的权重。然后，利用重构后的查询，进行新一轮的检索。早期的实验，包括Salton1971年用SMART系统[24]和1976年Robertson和Sparck Jones用概率模型做的早期实验[15]，在小的测试集上使用用户反馈技术，都发现了明显的提高。在一些后续研究工作的测试集测试中，有有效的表现。

由于有些时候，可能得不到用户的相关性判定，上世纪90年代初，又出现了另一种相关反馈技术，在没有用户参与的情况下有系统自主做出查询重构。其中最著名的就是伪反馈（pseudo-feedback）[25]。伪反馈是用来进行相关反馈的一个广泛使用的方法。它的主要思想是，认为初始检索结果的文档集中，有一小部分是跟用户的查询相关的，比如认为初始检索后的前20篇网页是相关网页，利用查询词在这些文档中出现的情况来近似估计查询词在整个相关文档集中的表现，以此来进行相关性反馈，对初始查询进行重构（查询扩展或调整查询词权重），然后再利用新的查询进行新一轮的检索。

到目前，已经有较多种反馈的方法，大都是基于相关文档集合和不相关文档集合都为已知的假设。 当然，最理想的查询是无法得到的，因为完整的相关文档集和不相关文档集都是不可能得到的。但是相关反馈的判断可以提供一个对这些集合的近似模拟。

5.2 天网知名度系统中的相关性反馈

本系统中采用调整权重的方法。查询扩展的方法，目前仍然是基于无结构查询的，即是把查询当作一个无结构的词串来看待，因此新扩展进来的词不存在区别分类的问题。而在天网知名度系统中，用户属性信息是有结构的，是由区别领域的多层次名人实体模型表示的。如果要做到有效的查询库扩展，就需要把针对不同属性类别扩展新词。然而，现有的信息提取技术比较成熟的是命名实体识别和二元关系识别，针对本系统，只能对姓名和单位、职业/职务属性能够做一定准确性的提取，查询扩展也就只能对这两类属性可行，其他属性目前无法做到达到较高准确性的提取，因而无法做有效的查询扩展。由于多层次名人实体模型表

示的名人属性信息相对普通搜索引擎的平均查询词是比较丰富的，而且已经是一个结构化的表示，因而采用调整查询词权重的方式是可行的。

本系统中采用伪反馈和用户反馈两种方法实现相关性反馈。检索词的权重计算方法均是根据该词在相关文档和不相关文档中的分布概率，如下式[15]所示：

$$w^{(1)} = \log \frac{p_i(1-q_i)}{q_i(1-p_i)} \quad ,$$

其中 p_i （q_i）是第i个索引词在相关（不相关）文档中出现的概率。上式对应于模型 4 中的 $w^{(1)}$。

分别采用三种相关反馈排序算法--经典反馈方法[15]、Pr_cl[26]、Pr_adj[27]。它们的区别在于对查询词在相关文档集合和不相关文档集合中出现概率的计算方法不同：

经典反馈方法： $p_i = \frac{r}{R}, \quad q_i = \frac{n-r}{N-R}$

Pr_cl： $p_i = \frac{r+0.5}{R+1}, \quad q_i = \frac{n-r+0.5}{N-R+1}$

Pr_adj： $p_i = \frac{r+n/N}{R+1}, \quad q_i = \frac{n-r+n/N}{N-R+1}$

其中，r 为相关文档集合中包含检索词 i 的文档个数，R 为相关文档集合中的文档个数，n 为整个文档集合中包含检索词 i 的文档个数，N 为整个文档集合中的文档个数。

实验中，用系统中的评测方法对这三种反馈方法进行评测，Pr_cl 和 Pr_adj 的效果比经典反馈方法更优，因为加入的调节参数更多地避免了 log 后的项取值为 0 的情况。Pr_cl 比 Pr_adj 略优，后面实验中采用 Pr_cl 方法。

相关文档集和的选取是伪反馈和用户反馈的最主要区别。伪反馈方法，在前一章针对不同领域构建多层次的用户属性信息描述基础上进行初始相关度评价，取返回的按相关度从高到低排序的网页队列的前 R 个网页作为相关文档集合，对每个查询词，Pr_cl 公式中的 r 为前 R 个网页中包含该词的网页数量。用户反馈，在本系统中选取训练集中手工标记为“高相关”的网页集合作为相关文档集合，其数量为 R，其中包含该查询词的文档数量为 r。

在这基础之上，通过实验来分析和研究在名人网页的相关度评价系统中，伪反馈和用户反馈两种方法在提高系统评价质量效果上的差别以及影响反馈效果的因素。

5.3 实验与分析

5.3.1 伪反馈

5.3.1.1 反馈效果

本文采用伪反馈方法，主要讨论两个问题：初始检索的质量对伪反馈的影响；现有系统评价模型基础上使用伪反馈方法的可行性。

伪反馈方法，采用 Pr_cl 公式，取返回的按相关度从高到低排序的网页队列的前 n 个网页作为相关文档集合，对每个查询词，前 n 个网页中包含该词的网页数量对应 Pr_cl 公式中的 r 值。实验中，n 的选取从 1 到 19。

第一组实验，由前一章的讨论知道，加入区别领域的属性权重参数后，系统的相关度评价性能有明显的提高。分别以加入该参数前后的两种评价方法作为初始检索方式，进行伪反馈。通过这个实验来验证上述论文中提出的调节初始检索结果对伪反馈质量的影响。两次初始结果之上进行伪反馈，用评价方法 1 进行评价，优化后进行的伪反馈效果比优化前进行伪反馈的提高如图 5-1 所示。

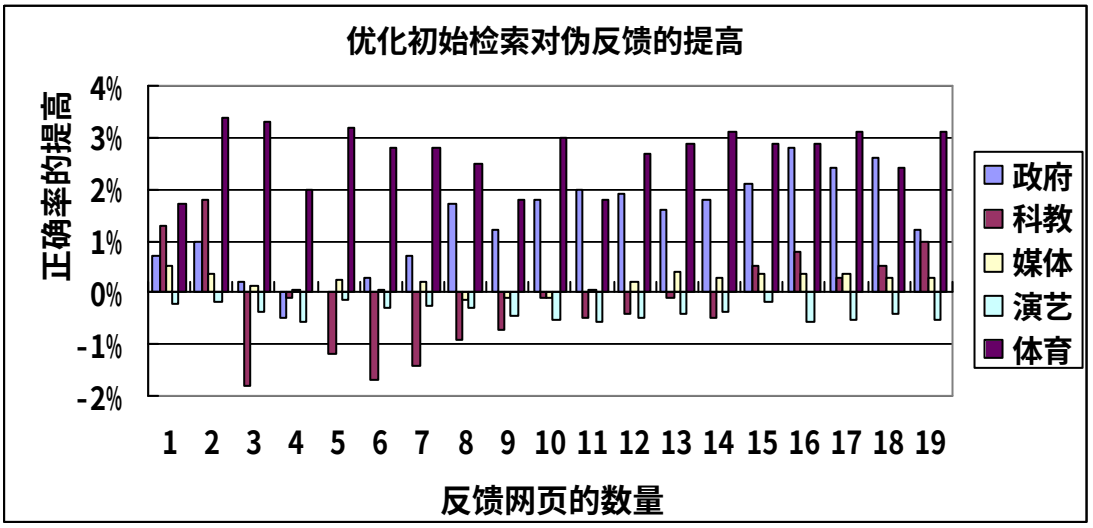

图 5-1 优化初始检索对伪反馈的提高

从图中可以看出，初始检索优化后再进行伪反馈比优化前进行伪反馈的效果有比较显著的提高，证明使用伪反馈方法，初始检索的质量的高低直接决定反馈结果的好坏。要使用伪反馈，就要在使用之前对初始检索进行优化，这个顺序非常重要。

文献[25]中，讨论了如何把相关度排序函数的调整与伪反馈结合起来共同提高检索质量，强调一定要在应用伪反馈之前调整优化排序函数。本文中通过实验得出的结论与其一致。

第二组实验，在前一章针对不同领域构建多层次的用户属性信息描述基础上进行初始相关度评价，对于伪反馈后的评价结果分别用评测方法 1 和 DCG 评测方法分别做了评测。

评测方法 1 得到的曲线，如图 5-2 所示（DCG 评测曲线不再列出）。在选择的反馈网页数量从 1 增长至 19 的过程中，所有领域的评测值都有下降，政府和体育类尤为显著，仅有一个例外，就是当反馈网页数量大于 15 以后，政府领域名人的相关度评价有提高。按降幅从小到大排序，领域序列为：科教、演艺、媒体、体育、政府。

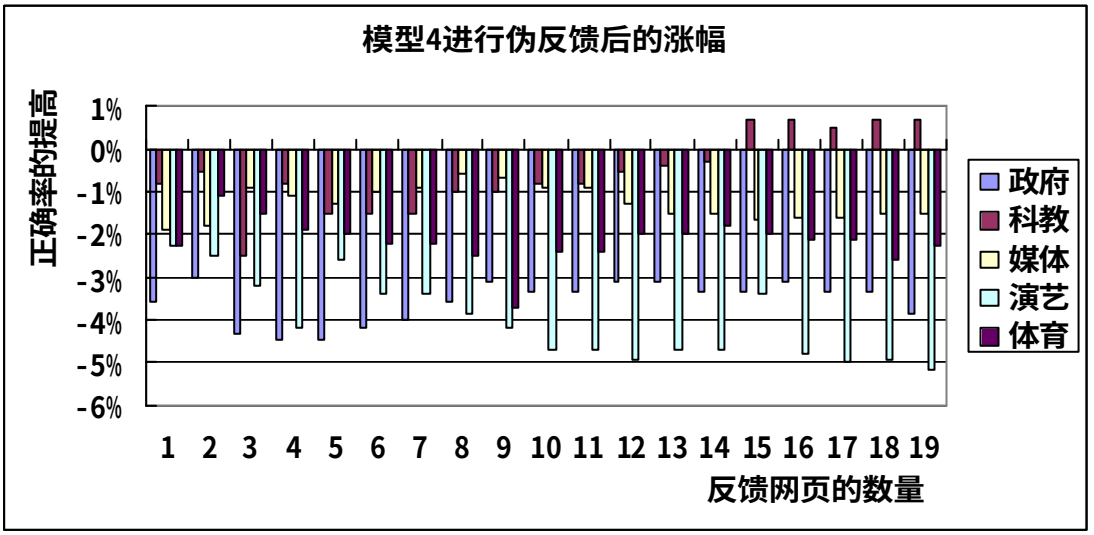

图 5-2 模型 4 伪反馈后的涨幅

5.3.1.2 影响反馈效果的因素

第一组实验结果表明，初始检索的质量很大程度地影响伪反馈的效果。应该先对初始检索模型进行优化，再使用伪反馈，这个顺序很重要。

第二组实验结果表明，在现有系统初始检索的基础上使用伪反馈对相关度评价没有提高，反而有削减。说明伪反馈方法是有一定的适用条件的，反馈前的相关度评价需要有较高的质量，即反馈选用网页的相关性整体上要比较高，以这些网页集合模拟完整的相关文档集合，才有可行性。在当前系统中，初始检索虽然经过了优化，但是返回序列中排在前面的网页，噪音还比较大，还不能达到使用伪反馈方法的条件，直接进行伪反馈，吸纳了很多噪音，导致系统相关度评价性能反而下降。因此，初始检索基础上直接进行伪反馈的方法，在本系统目前的初始检索质量下尚不可行。

5.3.2 用户反馈

5.3.2.1 反馈效果

用户是可以很容易地判断一个网页是否满足他（她）的信息需求的。对每个实体，用语料库中手工判定为高相关度的网页作为相关文档集合，在此基础上仍采用 Prc_l 反馈方法进行反馈。实验结果与反馈前的结果对比，用评测方法 1 进行评测，结果图 5-3 所示。

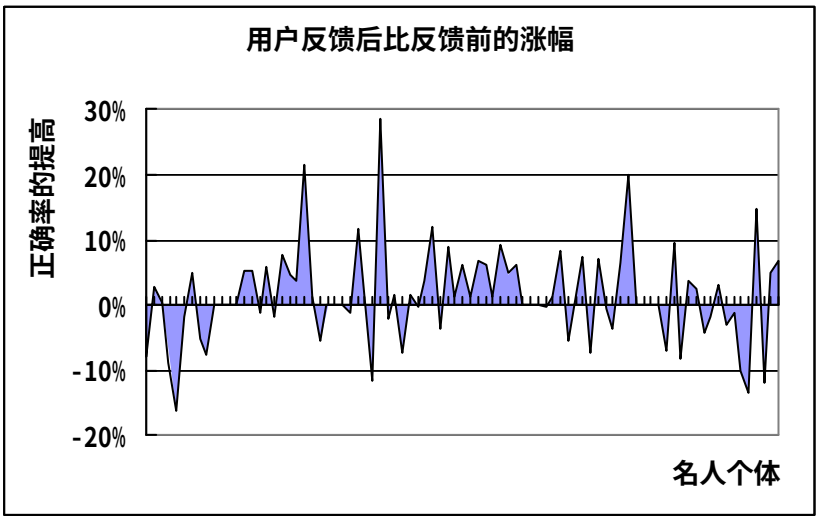

图 5-3 用户反馈后的涨幅

实验数据显示，用户反馈在科教、媒体、演艺领域有较好的表现，但是同时也发现，在政府、体育领域进行反馈后，平均程度上却有下降。就总体而言，用户反馈的文档集合相关度较高，内含噪音较少，用户反馈是目前系统中更可行的相关反馈方案。

5.3.2.2 影响反馈效果的因素

从上文中看到，使用多种不同方法进行反馈时，各领域按涨幅排序得到的序列基本相同，而且科教在几次反馈的各项指标中都有所增长且涨幅明显，领域背后有什么原因呢？查询输入和排序方法是检索中最重要的两个方面，上面讨论了各种排序的方法，那么这次来研究一下查询输入，在本系统里就是用户注册信息。对本系统中各领域名人的注册信息进行统计，数据如表 5-1 所示：

表 5-1 用户属性信息词数

	政府	科教	媒体	演艺	体育
人数	70	39	55	40	24
平均词数（切词后）	28	100	47	40	20

从这个表中可以看出一些规律，进行两种反馈后在各项指标中都表现突出的科教类名人的属性信息词数最多，而且幅度明显。在 DCG 指标中有所提高的媒体和演艺类实体的属性信息词数也较多，而频频下降的政府和体育类名人实体的属性信息词数是最少的。

对于领域中的所有实体，不区分领域，只根据属性信息词数做统计。拥有不同属性词数的用户与他多对应的用户反馈后的增长用 DCG 评测做出的结果，得到结果如下图所示：

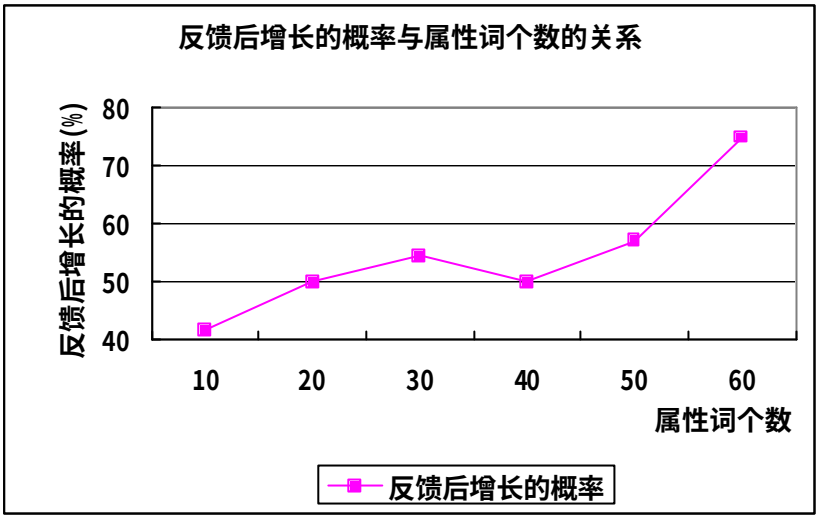

图 5-4 反馈后增长的概率与属性词个数的关系

对于用来描述用户信息需求的实体属性信息而言，它的丰富意味着对该实体的充分描述，即对用户信息需求的充分表达，那么在进行用户反馈的过程中，就可以充分的从相关网页集合中全面的找出真正对相关度评价贡献大的词语，提高其权重。而如果实体属性信息比较贫乏，那么，虽然从相关文档集合中提取出来的以有查询词的分布情况，但是那些对描述该实体非常重要的词的信息就丢失了，所以不能够充分提高相关度检索质量。由于天网知名度的用户属性信息的多层次性，已有的信息提取技术尚无法支持从文档集合中自动分类提取，所以无法使用查询扩展技术，那么，已有实体属性信息的丰富性，就成为用户反馈过程中提高相关度评价的关键。上面的实验中证实了这个想法。

另外，很多以查询扩展实现的用户反馈方法中也发现反馈时加入的词数，用户反馈的文档数都跟反馈后的结果有关系。这和天网知名度系统中的实验结论有相通之处。

上面的实验中，给出的还仅仅是实体属性信息词数与反馈后增长的概率的关系。同时，反馈后相关度评价质量的提高，还收到词的质量的影响。比如，从上面的试验和数据中可以看到，政府类名人的属性信息词数比体育类名人的属性信息词数要多，但是在用户反馈以后，其降幅却比体育类名人更大，通过对政府类名人属性信息的观察，发现政府类名人之间属性信息中的重复率比较高，比如其特征词往往都有“会见”、“报告”、“会议”等，虽然可以体现政府领域的特点，但是在同为政府领域的名人之间，其区分度就会较低。而关于某一次会议召

开的报道网页中，常会出现多个名人，那么这些词就不足以区分该网页对哪为名人更相关了。体育类词虽略少，但是其特征词等能区分他的运动领域，还是有相对较高的区分作用的。因此虽然体育类名人属性信息词数比政府类少，但是跌幅比政府类小。

5.4 本章小结

本章中采用了伪反馈和用户反馈两种相关性反馈方法，通过实验分析两种反馈方法效果的差异以及影响反馈效果的因素，得出三条结论：

1．伪反馈受初始检索质量的影响很大，一定要在伪反馈之前尽量优化初试检索质量。这个顺序很重要；伪反馈需要在初始检索达到一定优化程度之后才能提高系统相关反馈质量，必须保证用以反馈的网页有较高的质量。目前仅用模型3，尚无法达到这种要求，伪反馈方法中噪音太大，不能直接在模型 3 初始检索的基础上进行；

2．用户反馈在总体上提高了系统相关度评价性能；

3．用户反馈的效果与实体属性信息的词数有很大关系。系统试验表明，实体属性信息越丰富，进行用户反馈后相关度评价质量提高的概率越大。

第六章 总结和展望

6.1 总结

天网知名度系统的相关度评价方法是整个系统服务质量的关键所在，本文的主要工作集中在提高系统相关度评价质量部分。

针对系统中原有名人网页相关度评价模型的不足，为提高系统相关度评价质量，本文提出了一种基于概率模型的名人网页相关度评价模型。

1） 首先，针对 Fame 系统中名人网页相关度评价的特点，构建基本相关度评价模型。构建基础是 Okapi BM25 检索模型，在其基础上引入 HTML 标记权重系数，改进 Okapi BM25 公式，弥补其没有考虑 HTML 标记的不足。利用 Fame 系统数据集进行评测，实验结果表明 HTML 标记系数的引入提高了系统相关度评价质量，同时显示该基本模型优于原有系统中的相关度评价模型，提高了系统性能。

2） 其次，由于不同领域名人的属性信息对其相关度评价有不同的贡献，本文中构建了区分领域的多层次实体模型，来更好的描述用户的信息需求。同时在基本模型基础上引入属性信息权重系数，使基本模型从不支持结构化查询需求改进为支持多层次实体模型。各领域的权重系数通过训练集训的方式获得，避免了人工赋予的不确定因素。选取对系统相关度性能提高最大的一组权重系数作为模型中的领域参数，该套参数通过测试集的测试，证明有较好的适用性。

3） 再次，采用了伪反馈和用户反馈两种相关反馈方法，为实体属性信息进行权重的自动调整，以达到系统相关度评价的进一步优化。通过实验得出的结论是：第一，初始检索的质量很大程度地影响伪反馈的效果。应该先对初始检索模型进行优化，再使用伪反馈，这个顺序很重要；同时初始检索的质量需要达到一定高度后，使用伪反馈才能提高系统检索质量，目前系统的初始检索质量仍不适宜直接进行伪反馈。第二，用户反馈在总体上自动优化了属性信息权重，提高了系统相关度评价质量。第三，用户反馈的效果受名人实体属性信息词数的影响，属性信息越丰富，采用用户反馈后评价质量提高的概率越大。

另外，在系统方面，05 年 5 月后，新的系统流程中，改变了原有系统搜集模块的网页搜集机制，从直接进行网页抓取改变为从天网搜索引擎过滤获得网

页。同期，为系统构建了新一批名人实体数据集。加入新名人实体 150 人，人工制作了相应的名人实体属性信息库和标注了相关度高、中、低三个等级的名人实体网页数据集（含网页 3887 篇，平均每位名人实体 26 篇相关网页）。

6.2 展望

进一步工作仍需围绕相关度评价这个核心问题进行，这是提高系统服务质量的关键。包括：

可以在系统现有的 3 个模型基础上尝试进行结果的综合，这几个模型各有侧重，一些研究表明不同方法的评测结果进行综合有可能带来提高。

用户反馈的方法可以尝试变通的查询扩展，通过实验发现了属性信息词数越多，反馈后提高的概率越大，启发我们如果能够适宜地扩展更多的相关词进入属性信息，可以提高系统评价质量。但因为信息提取还无法实现按属性类分别提取，所以扩展进来的词无法分到各个类别中去，可以尝试对于它们进行特殊对待，为其属性信息权重系数赋予一个特殊的值，权重的其它部分可以通过反馈去调整。

可以在系统中扩展新类型的实体，比如机构、产品等。需要根据新的类型建立相应的实体模型。同时目前系统只是针对网页中的文本信息进行检索，多媒体类型数据的相关度评价也是将来的一个研究方向。

另外，天网目前是增量检索，数据更新速度较快，天网知名度系统目前从天网的过滤仍然采用的批处理的方式，定期更新，更新周期要比天网网页长，这是因为，同时系统中网页分析模块的信息提取部分耗时仍较长，天网知名度系统的资源有限，还无法承受与天网同步所带来的数据量扩张和系统消耗。因此在提高信息提取部分的效率和提高系统性能方面还有很多工作需要进一步开展。

参考文献

[1] 李晓明. 对中国曾有过静态网页的一种估计. 北京大学学报，2003，第 39 卷，第 5 期，394-398。

[2] J. Cho. Crawling the Web: Discovery and Maintenance of a Large-Scale Web Data. Ph.D. thesis, Stanford University, 2001.

[3] Brian E. Brewington and George Cybenko. How Dynamic is the Web? Computer Networks, 2000, Vol.33, 257-276.

[4] 中国互联网络信息中心（CNNIC）. 2004 年中国互联网络信息资源数量调查报告，http://www.cnnic.net.cn/download/ 2005/2005041401.pdf，2005。

[5] www.google.com

[6] 咎红英. 基于实体属性的中文网页检索研究. 北京大学博士论文，2004。

[7] K. Sparck Jones. A Statistical Interpretation of Term Specificity and its Application in Retrieval. Journal of Documentation, 1972, Vol. 28, 111-121.

[8] A. Singhal. Modern Information Retrieval: a Brief Overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2001, Vol. 24(4), 35-43.

[9] www.ask.co.uk

[10] D. Yimam and A. Kobsa. Expert Finding Systems for Organizations: Problem and Domain Analysis and the DEMOIR Approach. Journal of Organizational Computing and Electronic Commerce, 2003, Vol. 13(1), 1-24.

[11] D. Mattox, M. Maybury and D. Morey. Enterprise Expert and Knowledge Discovery. In Proceedings of the 8th International Conference on Human-Computer Interaction (HCI International'99), Munich, Germany: 1999, 303-307.

[12] B. Krulwich and C. Burkey. The ContactFinder Agent: Answering Bulletin Board Questions with Referrals. In Proceedings of the 1996 National Conference on Artificial Intelligence, Cambridge, Mass: MIT Press, 1996, Vol. 1, 10-15.

[13] A. S. Vivacqua. Agents for Expertise Location. In Proceedings 1999 AAAI Spring Symposium on Intelligent Agents in Cyberspace, CA, USA: Stanford, 1999, 9-13.

[14] M. Maron and J. Kuhns. On Relevance, Probabilistic Indexing and Information Retrieval. Journal of the ACM, 1960, Vol. 7, 216-244.

[15] C. J. van Rijsbergen. Information Retrieval. London: Butterworths, 1979, 2nd edition.

[16] S. E. Robertson and K. Sparck Jones. Relevance Weighting of Search Terms. Journal of the American Society for Information Sciences, 1976, Vol. 27(3), 129-146.

[17] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval. Addison Wesley, 1999.

[18] S. E. Robertson, S. Walker, and K. Sparck Jones, et al. Okapi at TREC-3. In Proceedings of 3rd Text Retrieval Conference (TREC-3), 1995, 109-126.

[19] S. E. Robertson and K. Sparck Jones. Simple, Proven Approaches to Text

Retrieval. Technical report, University of Cambridge, 1997.
[20] LEI Ming, WANG Jianyong, CHEN Baojue, et al. Improved Relevance Ranking in WebGather. Journal of Computer Science and Technology, 2001, Vol. 16(5), 410-417.
[21] Kalervo Järvelin and Jaana Kekäläinen. IR Evaluation Methods for Retrieving Highly Relevant Documents. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2000, 41-48.
[22] Ellen M. Voorhees. Evaluation by Highly Relevant Documents. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2001, 74-82.
[23] J. J. Rocchio. Relevance Feedback in Information Retrieval. In G. Salton, editor, The SMART Retrieval System: Experiments in Automatic Document Processing, Englewood Cliffs, NJ: Prentice-Hall, 1971, 313-323.
[24] G. Salton. The SMART Retrieval System-Experiments in Automatic Document Processing. Prentice Hall, 1971.
[25] FAN Weiguo, LUO Ming, WANG Li, et al. Tuning Before Feedback: Combining Ranking Discovery and Blind Feedback for Robust Retrieval. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2004, 138-145.
[26] Croft, W.B. and Harper, D.J. Using Probabilistic Models of Document Retrieval without Relevance Information. Journal of Documentation, 1979, Vol. 35, 285-295.
[27] S.E. Robertson, On Relevance Weight Estimation and Query Expansion, Journal of Documentation, 1986, Vol. 42, 182-188.

致 谢

首先感谢我的导师——李晓明教授，李老师在我的学习和研究中给予了很多关键性的指导。他对科研事业的热情给我最强的感染，他踏踏实实做事的工作态度让我最为敬佩。为自己热爱的事业真正踏下心来动脑筋，是让我最受益的事。

感谢闫洪飞老师在日常工作中给予的支持，在天网组的三年里参与了很多方面的工作，拓展了视野，锻炼了能力。

在三年的学习和科研工作中，我有很多的收获和进步，特别感谢给我最多无私帮助和耐心指点的彭波老师！他工作成绩突出，并且非常乐于帮助别人。在本次完成论文过程中，一度出现极大的困难，彭波老师给了我很多指导和帮助，帮我渡过难关，又在后来的论文写作过程中不厌其烦地帮我修改。在此，向彭波老师表达最真挚的谢意！

感谢天网组以及网络实验室其它各组所有的老师和同学们，这是一个团结、奋斗、热情、活泼、积极进取的集体，这是我工作的平台，是我学习的源泉，是我热爱的团队，祝愿网络实验室继续大步向前发展！

感谢相处最久的 Fame 项目组成员咎红英、胡景贺、朱家稷、苏琦、黄连恩、苏玉梅、姚从磊，Fame 项目组是我三年来的“ 主业 ”，虽然 Fame 组的一些成员已经毕业或到别的地方求学，但是在一起的时光永远那么让人留恋，工作中我们建立了深厚的友谊，大家给了我很多的帮助，祝愿大家再未来能实现更多理想，取得更多进步！

感谢我挚爱的父母！多年来的外地求学，少了很多团聚的日子，但你们对我的牵挂我感受得真真切切，你们给我的鼓励激发我向前。每当我遇到困难，你们总是我最坚强的后盾，让我知道永远有一个温馨的港湾等候着我，于是我勇敢、坚强、无所畏惧！我深爱的父母，你们的健康和快乐就是我最大的幸福！

感谢三年来与我朝夕相伴的好伙伴邱海艳、朱亚莉、李翔鹰。我们天天“ 腻 ”在一起整整三年了，从万柳到燕园，为了再次住到同一屋檐下还曾颇费周折。三年里我们一起分享开心和郁闷，共着同一个节拍，还制造了众多的经典语录。平静的日子里有大家陪伴，我每天都过得很快活；遇到困难的日子里，有大家的出谋划策和大力支持，我终于挺了过来！爱你们，感谢你们，祝福你们！

最后要感谢参加 05 年 5 月份‘ 天网志愿者 ’活动来自校内外的 26 位志愿者，他们为 Fame 系统制作了 150 位名人的数据集，对 Fame 系统的发展做出了很大的贡献，感谢他们的辛勤劳动和对北大天网的支持！

燕园三年的求学生活匆匆又匆匆，回首来时路，看到自己的成长，无论曾经的顺利和坎坷，都很欣慰，真诚感谢一路上所有给过我关心和帮助的人们！

北京大学学位论文原创性声明和使用授权说明

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，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。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。本声明的法律结果由本人承担。

论文作者签名： 日期： 年 月 日

学位论文使用授权说明

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

按照学校要求提交学位论文的印刷本和电子版本；

学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务；

学校可以采用影印、缩印、数字化或其它复制手段保存论文；

在不以赢利为目的的前提下，学校可以公布论文的部分或全部内容。

（保密论文在解密后遵守此规定）

论文作者签名： 导师签名：

日期： 年 月 日