分类号 TP391 密级

U D C

Web 数据挖掘用户兴趣建模技术研究

赵欣欣

导师姓名(职称) 刘玉树（教授） 答辩委员会主席 涂序彦

申请学科门类 工 学 论文答辩日期 2006-8-27

申请学位专业 计算机应用技术

2006 年 8 月 1 日

摘 要

随着 Internet 的飞速发展，Web 已经发展成为一个全球的、巨大的、分布和共享的信息空间。Web 数据具有分布、异质、动态、半结构或非结构等特征，这无疑给 Web 上的数据挖掘提出了挑战。Web 个性化服务，既是一种个性化服务，又是一种信息服务，它能够满足用户的个体信息需求。而用户模型是个性化服务的基础和核心。

本文围绕着如何建立用户个性化兴趣模型，分析了解用户兴趣，以便更好的为用户提供个性化服务进行了研究，并且研究了构建用户个性化兴趣模型中的关键技术，同时对用户个性化兴趣模型的应用进行了研究，包括个性化信息检索、个性化信息推荐及个性化信息推送。主要取得了以下几个方面的成果。

（1）提出了基于标记窗的网页正文信息提取方法

提出了标记窗的概念和基于标记窗的网页正文信息提取方法。该方法能够解决网页正文存放在多个 td 中的情况，能够解决正文文字短的网页的正文提取问题，尤其重要的是，它能够处理非 table 结构的网页正文提取问题。本方法无须将网页表示成一棵树，只需利用正则表达式，就可以直接提取出网页中标记对之间的正文，大大降低了算法的复杂度。

（2）提出了基于概念关系的用户兴趣模型的构建方法及基于用户兴趣森林模型的网页预取方法

用户访问一个页面后，一般会随着页面的链接来访问其它页面，可以对用户即将访问的链接进行预测，预先下载用户即将访问的页面，从而加快用户的浏览速度。根据用户的访问历史，利用“知网”建立了基于概念关系的用户兴趣森林模型，在此基础上，提出了基于用户兴趣森林模型的预取方法。通过计算超链接描述文字的平均带权语义距离，得到每个超链接描述文字的评价函数，进行网页预取。实验结果表明，该方法的预取命中率在 61%左右，具有较高的系统性能。

（3）提出了结合用户相对停留时间的用户兴趣模型的构建方法及基于泊松分布模型的网页推荐方法

用户兴趣模型的构建是进行个性化服务的基础，是进行网页推荐的基础。Web 推荐是基于数据挖掘和机器学习的方法，对用户可能感兴趣的网页进行推荐，它根据用户的爱好兴趣对用户可能访问的网页进行预测，并提供给用户进行选择。利用泊松分布对用户进行建模，当网页的归一化时间在 1-2 之间时，对用户可能访问的网页进行

预测，提供一系列推荐网页供用户选择，实验结果表明，推荐点击率在52%左右。

（4）提出了基于改进的汉宁窗函数的信息检索算法

传统的基于关键词匹配的检索方法检索时间长，检索结果质量差，无法适应用户群体的多样性。基于改进的汉宁窗函数的信息检索模型从词密度角度对用户的查询进行分析建模，对检索关键词进行概念扩展后进行概念检索，突破了机械式字面匹配局限于表面形式的缺陷，从词所表达的概念意义层次上来认识和处理用户的检索请求，根据用户访问 Web 的兴趣、爱好和使用目的，最大限度的满足用户的个性化检索需求。实验表明，该算法可以使得检索的查准率和召回率有大幅提高，很好的改善了检索的性能，提高了信息检索的有效性。

（5）提出了基于 RSS 的个性化信息推送方法

将 RSS 技术和个性化技术相结合的基于 RSS 技术的个性化信息推送方法将“推”技术和“拉”技术有效结合，打破了传统的信息获取方式，减少了用户上网搜索的工作量，将个性化的信息直接送给用户，提高了用户获取信息的效率。实例表明此方法是可行、有效的，可以为用户提供方便的个性化服务。

关键词：Web 数据挖掘；个性化；信息提取；用户兴趣模型；预取；推荐；信息检索；信息推送；

Abstract

With the rapid development of the Internet, Web has become a global, tremendous, distributed and shared information space. Special characteristics of the data in the Web, including distributed, heterogeneous, dynamic, semi-structured or non-structured, challenge the research work of Web mining. Web personalized service is not only a personalized service, but also an information service, which can meet the user's individual information requirement. And the user model is the foundation and the core of the personalized service.

This paper discussed how to construct user's personalized interest model, how to analyze and comprehend user's interests to provide better personalized service for user. And also did some research work on the key techniques in constructing user's personalized interest model. At the same time, some applications of the user's personalized interest model are mentioned in the paper, including personalized information retrieval, personalized information recommendation and personalized information push. The main innovative achievements are as follows.

（1）Web page information extraction algorithm based on tag window is proposed

The concept, tag window, and the Web page information extraction algorithm are brought forward in the paper. This method can deal with some special circumstances, all the web content information was put into several tds, and the character numbers of web content information were at most equal to that of the other information, navigation bars, advertisement, and the copyright, etc. Especially, it can extract the web content information which is not existed as the table format. And using this method, it was no need to express the web page as a tree, and only using regular expression can extract the context information existing between the tag pairs of the web page, which reduced the complexity of the algorithm greatly.

（2）The construction method of the user interest model based on the correlative relations of the concepts and the web pages prefetching method based on weighted semantic distance are proposed

Commonly speaking, after user visited a web page, he would visit another web page following the hyperlinks in the current web page. So, predicting the hyperlinks user will

visit soon and pre-sending them can increase the speed of user browsing the Web. Depending on the user's navigation history and the concept correlations living the HowNet, a user interest forest model based on the correlative relations of the concepts is constructed. And the web prefetching method based on weighted semantic distance is designed on the foundation of the user interest forest model. By computing the average weighted semantic distance of the texts in the hyperlinks can get the evaluation function of them, and the web pages a user may visit in the future are prefetched according to it. Experiments showed that the average hit-ratio of this method was about 61 percent.

（3）The construction method of the user interest model combining the relative time user spent on the web pages and the web page recommendation method based on Poisson distribution are proposed

The construction of the user's interest model is the base of providing personalized service and the web page recommendation. Web recommendation is the method based on data mining and machine learning, which recommends the web pages user maybe interest and predicts the web pages user may visit according to the user's interests and favorites and presents them for the user to choose. When the normalized time was between 1 and 2, utilizing the Poisson distribution to construct the user's interest model, and predict the web pages user may visit, and offer a series of the web pages for user to select, can get a better hit-ratio and an acceptable computational complexity.

（4）An information retrieval algorithm based on improved Hanning Window is proposed

The retrieval method based on keyword matching has a long retrieval response time and a bad retrieval quality, and cannot meet the different people's needs. An information retrieval algorithm based on improved Hanning Window analyzed the user's query considering the density of the words and did concept expanding on the query keywords and concept retrieval, which came over the disadvantages of the mechanically and outwardly matching and recognized and dealt the user's query requirement in the concept level of the words. It accords to the user's interests, favorites and purposes of visiting web to satisfy the user's personalized retrieval requirements by great extend. Experiments showed this method had improved the precision and recall greatly and ameliorated the retrieval

performance and boosted the efficiency of the information retrieval.

（5）a personalized information pushing method based on RSS is proposed

A personalized information pushing method, combining the RSS techniques and personalized techniques, integrates the “Push” and “Pull”, which break the traditional information getting method and reduced the work load of the user searching the web. It sends the personalized information to the user directly and increased the efficiency of getting the information. An example showed this method was feasible and effective, and could provide a convenient personalized service for the user.

Keywords： Web mining; Personalization; Information Extraction; User Interest Model; Prefetch; Recommendation; Information Retrieval; Information Push;

目 录

图 索 引

表 索 引

第一章 绪论

1.1 Web 数据挖掘

1.1.1 Web 数据挖掘的含义

Web知识发现[1]是一项综合技术，涉及web、数据挖掘、计算语言学、信息学等多个领域。Web数据挖掘是指从大量web文档的集合C中发现隐含的模式p。如果将C看作输入，将p看作输出，那么web数据挖掘的过程就是从输入到输出的一个映射$\xi: C \to p$。

Web数据挖掘是从数据挖掘发展而来，但是，web数据挖掘与传统的数据挖掘相比有许多独特之处[2]。首先，web数据挖掘的对象是海量的、异构的、分布的web文档。其次，web在逻辑上是一个由文档节点和超链构成的图，因此，web数据挖掘所得到的模式可能是关于web内容的，也可能是关于web结构的。此外， web文档本身是半结构化和无结构的，且缺乏机器可理解的语义，而数据挖掘的对象在一定程度上局限于数据库中的结构化数据。

Web数据挖掘包括三种完全不同的行为——结构挖掘、使用挖掘、内容挖掘，所有的这些行为都有数据挖掘的特性并且都被包括在网络中[3]。图 1.1 给出了Web数据挖掘的分类图。

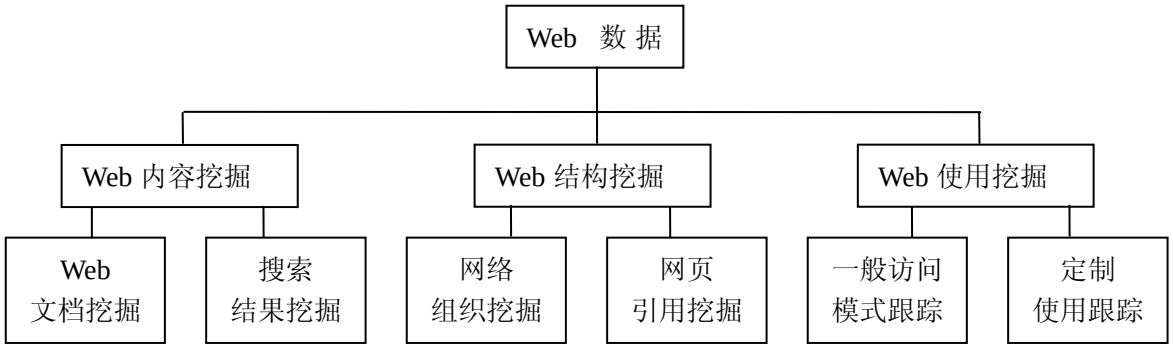

图 1.1 Web 数据挖掘的分类

Fig.1.1 the classification of the web data mining

Web结构挖掘是从WWW的组织结构、Web文档结构及其链接关系中推导知识[4-6]。由于Web文档之间的关联关系使得WWW不仅揭示了Web文档所包含的信息，也揭示了文档间的关联关系所代表的信息，反映了文档之间的某种联系，同时能体现某个页

面的重要程度。Web结构挖掘的目的是：发现Web的结构和页面的结构及蕴含在这些结构中的有用模式[7-8]；对页面及其链接进行分类和聚类[9]，找出权威页面（Authority Page）。主要的算法有：PageRank、HITS及改进的HITS、Hub/Authority算法[9][20]。

Web 内容挖掘是一种基于网页内容的 Web 数据挖掘，是从大量的 Web 数据中发现信息、抽取知识的过程。这些数据有文本数据，也有图像、视频、音频等多媒体数据，有来自于数据库的结构化数据（structural data），也有用 HTML 标记的半结构化数据（semi-structural data）和无结构（nonstructural data）的自由文本。

Web使用挖掘[10-11]是从Web访问日志中发现用户的访问模式，预测用户的浏览行为。通过对不同的Web站点的Web访问日志文件的挖掘分析，人们可以获取访问模式信息，帮助理解用户的意图和行为,为用户提供个性化的服务；可以了解Web结构，分析系统性能，改进Web站点的结构及其服务质量；可以改进Web系统设计；可以对日志数据进行多种统计，包括频繁访问页、单位时间访问频度、访问量的时间分布等。

1.1.2 Web 数据挖掘研究现状

1995 年，Lieberman 的 Letizia 是最早利用访问者的浏览信息进行推荐的系统。客户端的代理 Agent 可以监视用户的浏览过程，并搜索潜在的兴趣页面作为推荐，该 Agent 采用启发式搜索策略来推断用户兴趣。

1996 年，Yan等人根据访问者的访问模式自动对访问者进行分类，他们提出的模型有两个模块：离线模块（对Web Log进行聚类分析）和在线模块（动态连接）[12]。

1997 年，Joachims设计了WebWatcher，它是在早期Web使用挖掘中广泛流传的系统。其思想是创建一个导航Agent，基于用户兴趣，给用户浏览网站的时候提供导航[13]。系统由概括用户开始，查询他们的兴趣，每次用户访问一个页面，信息将通过代理Server，以便能够很容易跟踪用户在浏览WebSite时的会话，凡是用户可能感兴趣的链接都将被高亮度显示。

1999 年，Mladenic提出的Personal WebWatcher系统，可以为特定用户定制结构，仅记录用户访问的URL，并对兴趣链接进行高亮度标识，但不同于WebWatcher的是它没有学习过程，而且不询问关键字[14]。

1996 年Chen等人提出最大向前参考（Maximal Forward Reference）概念，主要是基于关联规则发现[15]。

1998 年Wu等人的Speed Tracer Project构架在Chen等人基础上，使用了参考页面和

请求页的URL作为浏览步骤和重建浏览过程对会话进行识别[16-17]。

1998 年，Zaiane 的 WebLogMining 系统结合在线分析处理 OLAP、数据挖掘技术、多维数据立方体对交互隐式知识进行抽取，系统过滤 Web Log 数据后，将它们转化成相关 DB。在第二阶段创建数据立方体，其中每个维代表了相邻域中所有可能的取值。此后，OLAP 技术被用来结合数据挖掘技术进行 Log 数据的预测、分类、以及时序分析。

2001 年 Huang 等人提出使用立方体模型精确识别 Web 访问会话，维护会话的顺序，并使用多个属性来描述被访问的页面。

1997 年Shahabi提出使用Client端的代理Agent捕捉客户的行为，并可以根据相似兴趣对用户进行分类[17]。

2000 年 Joshi，2001 年 Krishrapuran，2000 年 Nasravi 等人针对 Web 使用挖掘中的不确定性，通过使用鲁棒（Robust）算法进行用户会话聚集。将用户和页面分成多个集群，对 Log 数据预处理后，创建一个非对称矩阵来聚集典型会话。

1999 年Cooley，2000 年Srivastava将Web使用挖掘定义为 3 阶段：数据预处理、模式发现、模式分析。原型系统为WebSIFT。首先是智能清洗[18]、预处理，然后识别用户、Server会话，并依据参考域推断缓冲页面，同时还可以进行结构和内容的预处理[19-20]。

1999 年，Masseglia提出原型系统WebTool，将关联规则和时序模式发现应用于Web Log Files，从而动态定制超文本结构，并将整个Web使用挖掘过程分为两个阶段：数据预处理阶段和数据挖掘阶段。并于 2000 年使用ISEWUM方法，处理挖掘用户模式问题[21]。

1998 年 Mulvenna 和 Buchner 提出知识发现处理以便发现市场智能体，并建议创建一个能够进行在线分析的环境，定义了 Web Log Data 数据立方体（Hypercube）。1999 年 Buchner 提出 MIDAS 算法从 Log 文件中提取时序模式。

2000 年Spiliopoulon设计了MINT挖掘语言，通过识别出浏览模式的不同，并开发了可视化站点语义，用来扩展web使用挖掘（Web Usage Mining）的性能。在这个方法中，概念层次被用作Page分组的基本方法[22]。

2000 年 Coenen 等人提出适应 Web 站点的框架。

2000 年 Mobasher 的 WebPersonalizer 被认为是当前最为先进的系统，提供了 Web Log Files 的挖掘框架以便发现知识，并根据当前用户与以往用户相似的浏览操作，对

其进行推荐页面。目前又增加了加权处理，并提供实时推荐引擎。

目前国际上有影响的典型Web数据挖掘系统有[23]：

（1） Net Perceotion 公司的 Net Perception：采用了“实时建议”技术，让网站能够根据用户以往的浏览行为，在其他用户中找出与他相类似的浏览行为，从而为用户提供个性化的浏览建议。这种技术利用了网站用户的浏览行为有相似的一面，因此其预言准确性较高，并且它是实时运行，随着浏览量的增加，它会变得越来越聪明。

（2） Accrue 公司的Accrue Insight和Accrue Hit List：Accrue Insight主要是帮助顾客解决电子商务方面的问题，它是一个综合性的Web分析工具，能够对网站的运行状况有深入、细致和准确的分析[24]。它的设计是以顾客为中心的，通过分析顾客的行为模式，帮助网站采取措施来提高顾客对网站的忠诚度，从而建立长期的顾客关系。它利用了多种Web数据收集方法，包括Advanced Network Collectors, Server Collectors和Web Server Log Files，而不是像许多网站那样，仅仅分析Log文件。Advanced Network Collectors以其能收集到最大量的数据而著称，它能够收集到Web Server日志里所得不到的信息，但是对于加密的Session或者与它不适用部分，则用到另外两种方法。根据原始数据，Accrue Insight运用了一种叫做“Server Collector”的分析方法，它支持镜像服务器和负载平衡、路由器和一些其他网络结构设备，能够将一些加密的地址转化为可分析的形式。Accrue Hit List 是一个功能强大的Web报表分析工具，主要是用于完成网站流量分析，适合于中型网站，主要运用在市场分析，搜索引擎，广告等方面。

（3） Web Trends 的 Commerce Trends 3.0：它能够让电子商务网站更好地理解其网站访问者的行为，帮助网站采取一些行为来将这些访问者变为顾客，将一次性的顾客变为长期的忠实顾客。另外，它还提供了完全的“browser-based”方法，使得不同部门能够在任何时间得到他所想得到的个性化的报表。

（4） IBM 的产品是 IBM SurfAid ：通过一些技术和服务对网站用户的行为进行分析和理解。

（5） SAS 的产品 SAS e-Discovery 是一个电子商务分析包。

Web 数据挖掘的产品还有：XML Miner 挖掘 XML 代码,并使用模糊规则从中找到

关联和有价值的预测的 XML Miner；电子商务数据分析和数据挖掘专家 Web Usage Mining Consulting 等。

国内关于Web数据挖掘的研究也十分活跃，如南京大学计算机科学与技术系软件新技术国家重点实验室研发的网络信息挖掘系统——IDGS系统，采用向量空间模型和基于词频统计的权值评价技术，根据用户提交的挖掘目标样本，在Web上自动查找用户所需的信息，并利用改进的Robert技术，进行中英文技术资料的收集等工作[10]。

1.1.3 Web 数据挖掘面临的问题

WWW是一个巨大、分布广泛、全球性的信息服务中心，涉及经济、文化、教育、新闻、广告、消费、娱乐、金融、保险、销售、电子商务等信息服务，内容极其丰富。对Web进行有效的信息抽取和知识发现具有极大的挑战性，面临着很多具体问题[1]。

（1）目前，对感兴趣的信息仅限于利用各种搜索引擎进行查找[26]。尽管业界开发了很多的搜索引擎，但其检索性能和服务质量并不令人满意。主要表现在：

a) 检索方式单一，检索时间长，检索结果质量差，难以精确表达用户需求，无法适应用户群体的多样性。
b) 检索召回率和精度低。低查准率（精度）导致引擎返回的检索结果中往往含有大量无关信息。有用信息匮乏，用户难以得到真正感兴趣或有用的信息。低查全率（召回率）导致很多相关的文档查不到[27]。
c) 搜索引擎的更新周期较长，无法适应信息的快速增长[28]。
d) 缺乏检索导航信息[29]。用户无法顺利、快速地从巨大的信息网络中找到目标信息。
e) 定制服务能力差。不能根据用户多样化的需求，自动地、最大程度地满足用户的需求。
f) 主动服务和个性化服务能力差[30]。

（2）Web 页面以某种格式（HTML或XML）呈现的半结构化数据（semi-structured data），数据结构不规则（irregular）或不完整（incomplete）[31]，复杂程度高于普通的文本文档；数据结构隐含、模式信息量大、模式变化快；大量的文档无任何排列次序，无分类索引。

（3）Web 是一个异质、分布、动态的信息源。Web 及其数据的更新速度、增长速度极快，也无固定的模式。同时，其上的信息几乎都是隐藏的、潜在的、未知的，

从 Web 上发现这些未知的信息和有用的模式，仅用传统的基于关键字的检索方式很难实现。

（4）目前 Web 上的数据以 TB 数量级计算，且在迅速地增长，能否或如何构建一个庞大的数据仓库把 Web 上所有分布和异质的数据集成在一起。最近，有些研究工作在致力于存储和集成 Web 上的所有数据。

（5）不同的用户访问Web的兴趣、爱好和使用目的千差万别，面对一个非常广泛的形形色色的用户群体，能否使用户根据自己的爱好兴趣定制网页，甚至Web server能否根据发现的用户描述文件（profile） 自动为用户定制网页，从而提供个性化的信息检索和查询服务，已经成为了研究的焦点问题[32]。

（6）网络上信息储备量极大且信息内容十分丰富，但信息的利用率很低。Web 上的信息对用户个人而言，被使用到的只是极小的一部分，其余信息对用户来说是不感兴趣的[33]。据统计,99%的web信息对于 99%的用户是无用的[1]。

1.2 个性化服务的发展

1.2.1 个性化服务的兴起

目前，国内外学者对个性化有多种不同的说法。

IBM认为：个性化是搜集、存储站点访问者的信息，在对这些信息进行分析的基础上，在合适的时间将合适的信息传递给每一位用户的过程[34]。

美国学者Richard Dean认为：个性化是指让Web站点更多地回应每个用户的个人需要[35]。

国内学者认为：个性化的含义是使事物具有个性，或使其个性凸现，它包含两层含义，其一，个性是需要经过培养而逐步形成，其二，个体总是具有一定个性的，让这种个性得到认可、了解，并在一定空间中得以体现、展示[36]。

个性化服务[37]是一种能够满足用户个体需求的服务方式。它根据用户提出的明确要求提供准确的信息服务；或通过对用户专业特征、使用偏好的分析而主动地向用户推荐其可能需要的信息。同时个性化服务是一种培养用户个性、引导需求、引发需求的服务。

Web个性化服务[38]是指服务提供商根据以往用户在访问网站过程中的行为和兴趣，为相似行为用户提供相应的Web对象（如网页、文本、图像、声音、视频等等）

服务，从而简化用户查询过程，使得用户得到快速、准确的信息结果。这种个性化服务应尽可能使得每个用户在浏览该商业网站时都能产生这样的感觉——他/她是该网站唯一的用户，尽可能地迎合每个用户的浏览兴趣并通过动态快速调整页面内容来适应用户浏览兴趣的变化[39]。这样，Web数据挖掘技术成为了Web个性化服务的主要实现方式。

个性化服务是Internet信息增长的必然结果，是满足用户需求的服务，是培养个性、表现个性的信息服务。1995 年至 1997 年，美国人工智能协会春季会议（AAAI）、国际人工智能联合大会（IJCAI）、ACM智能用户接口会议（ACMIUI）和国际WWW大会等重要会议发表了多篇个性化服务原型系统的论文，标志着个性化服务研究的开始。1997 年 3 月，《Communications of the ACM》组织了个性化推荐系统的专题报道，个性化服务已经受到相当的重视。2000 年 8 月，《Communications of the ACM》再次组织了个性化服务的专刊，个性化服务的研究已经进入快速发展阶段。2000 年，美国NSF基金开始支持有关个性化服务的研究[40]。同年，以美国为主的多国个性化研究机构和网络公司成立了个性化协会，旨在推动个性化服务的发展。近几年，我国学术界也开始了个性化服务的研究，已经有相当一批有实力的科研机构投入到这个领域的研究中，也取得了丰富的研究成果。

1.2.2 个性化服务系统

随着电子商务的不断发展，个性化服务显得越来越重要。尽管已经存在许多个性化服务系统，但个性化服务技术仍有很多值得研究和探讨的领域和方向，主要有以下几个方面[41]。

（1） 用户兴趣和行为的表达。由于用户兴趣是多样的、动态变化的，跟踪、学习和表达用户兴趣是一个最基本和难以解决的问题。

（2） 分类和聚类技术

（3） 个性化推荐技术。如何克服现有推荐技术的缺点是进一步研究的方向。

（4） 安全技术。

SmartPush系统[42]根据用户的个性化信息向用户递送电子数据。

Alxa 系统收集用户的使用方式，作为对站点的质量评估的原始资料，以此为基础，确定相关的链接，提供推荐服务。

Leitzia[43-44]系统在用户浏览时向用户提供与其当前访问页面内容相关的其可能感

兴趣的链接。

1.2.3 个性化预取

缓存技术是将用户最近和经常访问的页面保存在本地机器上，当用户再次访问这些页面时，首先检查 cache 中是否存在对应的页面，如果有，则检查 Web 服务器上对应的页面有没有更新，如果没有更新，那么就从 cache 中取出该页面，发送给用户；否则就从对应的 web 服务器上取出页面给用户。当服务器中的 web 页被修改而缓存的页没有被修改时，就出现了给用户发送的网页内容和服务器的内容不一致的现象，这就是数据不一致问题。Web 缓存必须保持数据的新鲜度，也就是要保证 web 缓存的一致性。有四种熟为人知的保持缓存一致性的技术：浏览器轮询机制（client polling），失效性复查机制（invalidation callbacks），生命周期机制（time to live），是否曾经被修改机制（if-modified-since）。

Web 缓存的基本思想是以存储空间换取Internet 带宽，这样可以获得更好的响应时间，减少所占用的Internet 带宽。Web 缓存主要有四种模式：代理缓存、Web 缓存重定向、透明代理缓存、反向代理缓存。典型的web缓存模式如图 1.2 所示[45]。

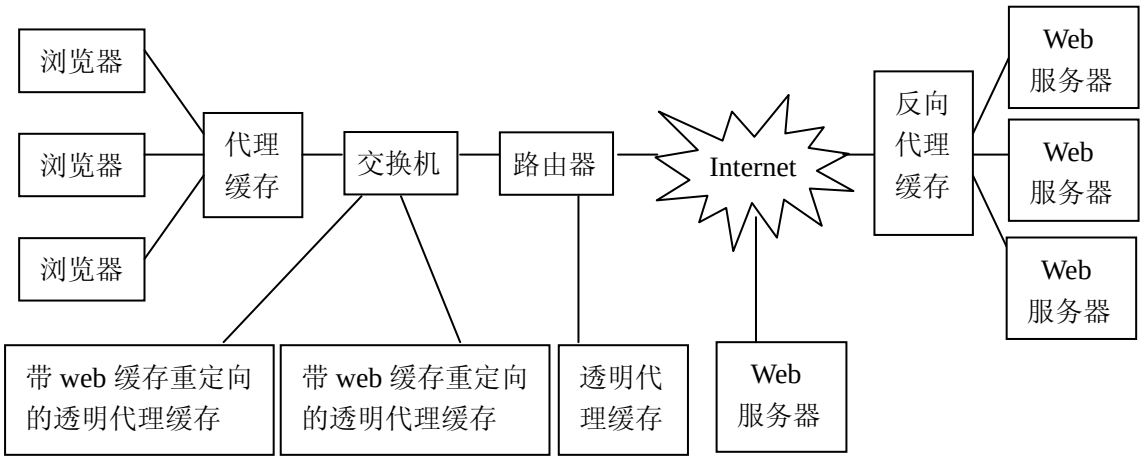

图 1.2 web 缓存的典型模式

Fig.1.2 the typical pattern of web caching

Web缓存有三个研究领域：缓存替代策略，缓存一致性和预取。缓存替代策略决定当web缓存满了的时候，web缓存中的哪些文档将被替代。替代策略的目标是最优化web缓存的使用空间，减少用户的等待时间。目前有几种重要的缓存替代策略：先进先出策略（FIFO），最近最少使用策略（LRU），最少使用频率策略（LFU），对象大

小策略（Size），最佳替换法策略（OR）等。通常，人们使用文档命中率，文档字节命中率，文档分组命中率[46]用来评价缓存替代策略，公式如下：

$$\text{文档命中率}=\frac{\text{文档的命中次数}}{\text{用户访问文档的总次数}},$$

$$\text{文档字节命中率}=\frac{\text{命中的字节数}}{\text{用户访问文档的总字节数}},$$

$$\text{文档分组命中率}=\frac{\text{命中的分组数}}{\text{用户访问文档的总分组数}}。$$

一个好的缓存替代策略应该有较高的文档命中率，较高的字节命中率和文档分组命中率。找到这三个评价指标的最佳结合点的代替算法是最佳的替代算法。

Web 预取的基本依据是，在用户发出的 2 次 HTTP 请求之间，常常有一段空闲时间，称为用户思考时间。这段空闲时间的长度可以从几秒钟到几分钟。预取就是利用这段空闲时间提前把用户不久将要访问的网页下载，并存放到缓存中，以减少用户实际访问时的等待时间。其基本思想是将用户不久可能访问的某些页面，在用户还没有请求时就取到用户的缓存中。当用户请求了预取页面中的某个页面时，由于该页面已经在本地的缓存中，因此能够直接从本地读取，从而减少了用户等待时间。预取方法利用了网络的相对空闲时间，是缓存机制的有效补充手段。如果预取足够准确，则缓存的性能将得到明显改进。

预取可以用来提高计算机系统功能，如：操作系统，文件系统和基于 web 的系统的性能的一种技术。一个好的预取系统是能够准确预测用户或者系统的下一个请求，仅仅使用空闲的系统资源预先执行这些请求的系统。

随着 Internet 的迅猛增长，万维网上的文档也随之迅速增长。大量的信息通过 Internet 传送，并且被周期性地修改更新。如今，在网络上传送的大多数内容是多媒体数据，如图像和视频，很大程度上的影响了网络的性能。网页预取的目的在于通过分析用户的访问历史记录，预测用户可能将会浏览的网页，预先取出存放于 Cache 中，以备用户访问，从而减少访问延迟。

预取是一种主动的 cache，它一直是 web 缓存技术的一个重要话题。Web 预取的主要思想是预测用户将要使用的文档，然后将他们放入缓存中。如果预取的内容是用户需求的内容，那么用户的访问是没有延迟的。预取技术在 web 中的应用大大减少了用户的等待时间。预取需要解决三个问题：第一，如何尽可能多的发现用户将来可能

的访问请求；第二，如何从这些请求中准确的预存用户的请求；第三，如何让预测算法普遍试用。

现有的预取方法有：基于历史的预取（Histroy-based-prefetch）和基于链接的预取（Link-based-prefetch）等。基于历史的预取方法仅仅能预测到曾经访问过的网页，对未曾访问过的网页不能作出预测。基于链接的方法对预测的结果不加取舍，将所有的链接和图像都进行缓存，故用户需要在预测的结果中进行选择，有可能更加浪费用户的时间。

预测模型是缓存和预取的核心。近年来，许多学者对网页预取技术进行了研究，提出了多种模型。

Albrecht[47]等人建立了一种结合四种马尔可夫模型进行预取的混合的马尔可夫模型。他们假设用户已经访问的网页序列是一个马尔可夫链并在马尔可夫模型中考虑并应用了时间因素，基于时间间隔信息和文档序列信息对web服务器log进行训练，学习马尔可夫模型。

Lau和Horvitz建立了仅根据查询和浏览信息来预测用户下一个查询的贝叶斯网络。他们假设下一个查询仅仅依靠前一个查询和时间间隔，对其他因素独立。在这一前提下，对用户查询进行分类，构建Baysain模型来预测用户的下一个查询目标或者查询意图[48]。

Su[49]等人在网页预取系统中应用n-gram语言模型，提出了一种利用服务器日志文件，运用n-gram预测模型对用户未来可能进行的Web 访问请求进行预测。把一系列的n长度的网页看作是一个n-gram，通过计算每一个n-gram的出现次数，基于最大计数的给出网页的预测结果。

Azer提出了一种基于概率模型的预取方法。根据服务器Log 数据，服务器计算出在一定时间间隔内，网页被连续访问的概率，并建立条件概率矩阵。据此，预测用户的访问请求[50]。

Schechter 等构造用户访问路径树,采用最长匹配方法,寻找与当前用户访问路径匹配的历史路径,以此预测用户接下来的访问请求[51]。

徐宝文[52]等提出了一种基于数据挖掘的预取模型。在这个模型中，用户兴趣表现为对词条的兴趣，兴趣关联规则表示从一个词条转向其他词条的可能性。其预取思想是将页面中的链接根据一定的策略进行排序，将认为最有可能被用户访问的页面预取过来。考虑到系统的响应时间及系统实现的难易程度，其利用兴趣间简化的关联规则

来实现对用户行为的预测。

许欢庆等[53]提出了基于用户访问路径分析的服务器端预取模型。模型通过分析用户的访问路径，挖掘其中蕴含的用户信息需求，据此预测用户的下一步访问请求。为了实现用户访问序列中潜在意图的挖掘，模型引入了隐马尔可夫模型（hidden Markov model,简称HMM）。

[54]中，将用户请求文档中的部分链接或者全部链接发送到客户端，这个模式可以减少客户的相对反映时间，但是，却没有进行有效的预测。[55]和[56]中，客户端从服务器端搜集最受欢迎的链接信息，然后决定预取哪些文档。

基于概念联想网络的网页预取模型是孙强等人提出的[99]。概念联想网络是通过连接概念间的联想关系形成的，由节点和连接节点的有向边构成。节点存放概念，通过定义串连规则和并联规则将节点联系起来。基于构建的概念联想网络实现网页预取评价。

Krogear，Long 和 Mogul 研究了缓存和预取对终端用户反映时间的影响，他们发现，仅使用缓存技术可以节省 26%的反映时间，而将缓存和预取相结合，构建如图 1.3 所示的将 web 缓存和预取相结合的模型，可以节省用户 60%的时间。

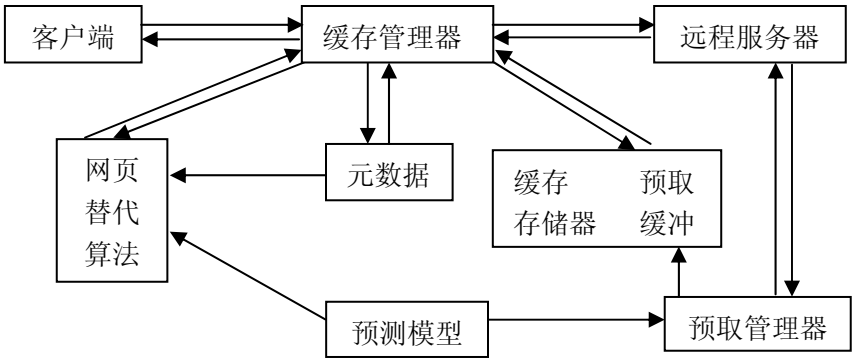

图 1.3　Web 缓存和预取相结合的模型

Fig.1.3 the model combining the web cache and web prefetching

1.2.4 个性化推荐

个性化推荐是指根据用户的兴趣特点，向用户推荐其感兴趣的信息。个性化推荐的原理是根据用户模型寻找与其匹配的信息，或者寻找具有相近兴趣的用户群而后相互推荐浏览过的信息。它的实质是“信息找人”的服务模式，可以减少用户寻找信息的时间，提高浏览效率。根据推荐所采用的技术，个性化推荐可分为基于内容的推荐、

协同推荐和混合推荐。

基于内容的推荐通过比较资源与用户模型的相似度来推荐信息。基于内容的推荐是目前个性化推荐服务的主流。Stanford 大学的 Balabanovic 和 Shoham 推出了个性化的推荐智能体 LIRA。它自动搜索页面，通过用户的反馈学习和更新用户模型，将满足用户兴趣的页面推荐给用户。

MIT开发的个性化浏览辅助智能体Letizia[57]通过跟踪用户的浏览行为学习用户模型，然后主动进行宽度优先搜索，将用户可能感兴趣的页面推荐给用户。

California大学推出的Syskill&Webert[58]通过用户对浏览页面的标注学习用户模型推荐满足用户兴趣的页面。

何军和周明天[59]讨论了基于内容的推荐中的关键技术。欧洁等[60]对基于内容的推荐中用户兴趣的发现进行了研究。

基于内容的推荐实现机制简单，但是此方法向用户推荐与用户模型匹配的信息，只能推荐与用户已有兴趣相似的信息，不能发现用户感兴趣的新信息。

协同推荐不比较资源与用户模型的相似性，它通过比较用户间的相似性推荐信息。具有相似兴趣的用户属于同一用户类。当用户对某信息感兴趣时，将此信息推荐给同一用户类中的其他用户。

目前，提供协同推荐服务的系统主要有Let's Browse[61]、PHOAKS[62]、Referral Web[63]、SiteSeer[64]和GroupLens[65]。国内已经对协同推荐的相关技术进行了研究。路海明等[66]基于多Agent混合只能实现合作推荐；赵亮等[67]提出了一种高校的个性化推荐算法；林鸿飞和王剑峰[68]就协同推荐的文本过滤模型进行了探讨。

协同个性化推荐的优点是可以发现用户可能潜在感兴趣的新信息，缺点是不能推荐那些从来没有被同类中其他用户访问过的信息[85]。

混合推荐是将基于内容的推荐和协同推荐结合起来的一种推荐方法。既比较资源与用户模型的相似度，又寻找具有相近兴趣的用户类。可以更好的进行推荐。Fab[69]是Stanford大学推出的推荐系统，它根据用户对浏览页面的标注构建用户模型，并根据用户模型的相似性寻找具有相似兴趣的用户。清华大学推出的OpenBookmark[70]通过集中管理用户群的Bookmark来实现混合推荐。

1.2.5 个性化信息检索

信息检索是用户寻找、定位感兴趣信息的主要途径，Internet 信息检索服务的质量

决定了用户使用 Internet 信息的效率。现有的 Internet 检索服务没有考虑用户的差异，对于任何用户，只要输入的关键词相同，返回的检索结果就完全相同。而实际上，不同用户由于知识背景、兴趣爱好等方面不同，需要的信息往往不同。当人们利用搜索引擎搜索信息时，往往会得到大量无用的信息，而真正满足人们需要的信息则淹没在搜索返回的信息海洋中。这种现象已经受到用户越来越多的不满。

互联网的飞速发展给自然语言处理的研究带来了新的机遇和挑战。把自然语言处理技术应用到网页处理中，对网络中的信息进行深层次的加工处理，有效的从浩瀚的信息海洋中挖掘可以为人所用的各种知识，提取出人们所需的信息，已经成为许多研究人员的研究目标。传统的信息检索不能很好地理解人的查询需求，没有考虑关键词所处的特殊环境，无法针对用户的个性化需求提供很好的个性化服务。智能化、个性化信息检索成为当今信息检索技术的研究热点。

个性化信息检索是指根据用户的兴趣和特点进行检索，返回与用户需求相关的检索结果。目前，个性化信息检索还处于研究阶段，NEC 研究院提出了个性化元搜索引擎原型系统 Inquirus2，国内南京大学推出了个性化信息检索智能体 DOLTRI-Agent，浙江大学提出了个性化检索系统 NetLooker。

由于在检索中考虑了用户的差异，个性化信息检索可以大大提高检索的效率。个性化信息检索尚处于研究阶段。Glover等[71]提出的个性化元搜索引擎原型系统 Inquirus2 根据用户输入的偏好优化查询关键词，对搜索引擎返回的结果进行排序。DOLTRI-Agent[72]可以学习用户兴趣，并根据得到的用户模型提供个性化信息。徐振宁等[73]提出了基于本体论实现个性化信息检索的技术。

自然语言理解是自然语言处理的高级阶段，它研究如何能让计算机理解人们日常使用的语言，使得计算机懂得自然语言的含义，并用自然语言回答人们提出的问题。目前，由于自然语言理解技术自身发展还不成熟，因此，将自然语言理解与信息检索技术相结合，进一步提高检索的整体性能仍是一个有待挖掘的课题。个性化信息检索是信息检索的发展趋势，以自然语言为基础的信息检索技术研究已经成为信息检索领域研究的新方向。

1.2.6 个性化信息推送

信息推送服务是基于推送技术发展而出现的一种新型服务，突出的是信息的主动服务。信息推送服务方式可以分为两大类：一类是由智能软件完成的全自动化的信息

推送服务；另一类是借助于电子邮箱并依赖于人工参与的信息推送服务。信息推送服务的基本过程是：用户信息需求了解、专题信息搜索、信息定期反馈。

一般来说，首先由用户向系统输入自己的信息需求，然后由系统或人工在网上进行针对性的搜索，最后定期将有关信息推送到用户主机上，通过邮件、“频道”推送、预留网页等多种途径将信息送给用户，改“人找信息”为“信息找人”。信息推送服务打破了传统的信息获取方式，减少了用户上网搜索的工作量，将个性化的信息直接送给用户，提高了用户获取信息的效率。

1.3 面向个性化服务的用户兴趣建模

用户兴趣建模是个性化服务的基础和核心，用户兴趣建模就是从用户信息中构建用户模型。

1.3.1 用户兴趣建模概念

从有关用户兴趣和行为的信息（如：浏览内容、浏览行为等）中归纳出可计算的用户模型的过程，即用户兴趣建模，是个性化服务的核心和关键技术，也是个性化服务中最重要的一环。作为个性化服务的基础和核心，用户模型的质量直接关系到个性化服务的质量，只有在高质量的用户兴趣建模的基础上，只有当用户的兴趣、偏好和访问模式等用户信息可以很好地被系统“理解”的时候，才可能实现理想的个性化服务，才能实现个性化服务系统所追求的各种目标。下图描述了用户模型与个性化服务之间的关系。

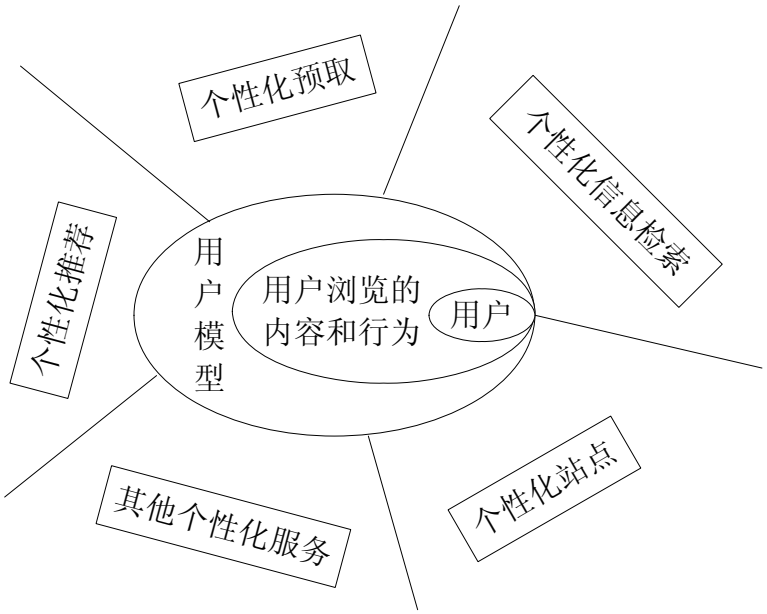

图 1.4　用户模型与个性化服务之间的关系

Fig.1.4 the relation of the user model and the personalized service

用户模型的表示决定了用户模型反映用户真实信息的能力，同时也在一定程度上限制了用户兴趣建模方法的选取。用户模型的表示目前还没有一个统一的标准，常用的表示方法有以下几种。

（1） 主题表示法

用户模型的主题表示法是指以用户感兴趣的信息的主题来表示用户模型的方法。如果用户对体育和音乐感兴趣，那么用户模型表示为{体育，音乐}。用户的兴趣可以由用户自己提供，也可以由系统按照一定规则通过学习算法，从用户访问的页面中抽取反映用户兴趣的部分关键词组合来表达用户兴趣。

（2） 用户 Bookmark 表示法

Bookmark法是指用用户保存过的重要站点或页面的Bookmark来表示用户模型，在用户浏览web 的过程中，如果用户遇到其很感兴趣得或者有价值的网页或者网站，一般会将相应的URL保存在Bookmark中，以方便以后的浏览。用户的Bookmark可以反映用户很关注的兴趣主题。典型的采用用户的Bookmark来表示用户模型的个性化系统如Siteseer、PEA[74]。这两个系统中，Bookmark的不同的目录表示兴趣的不同类别，各个目录中的文档（URLs）表示成加权关键词向量的形式，用户兴趣由URLs的列表和它们的结构组成。国内清华大学的卢增祥等也提出在网络信息过滤时，采用Bookmark表示用户兴趣[75]。Bookmark表示方法仍然是基于关键词的表示方法。

（3） 关键词列表表示法

用户模型的关键词列表表示法是指以用户感兴趣的信息的关键词来表示用户模型的方法。如果用户对乒乓球感兴趣，则用户模型可以表示为{乒超，邓亚萍，王励勤，波尔，梅兹}等。关键词可以由用户指定，也可以通过学习算法得到。通过学习算法得到表示用户模型的关键词在本质上与文本分类中的特征选取问题相似，都是通过训练样本得到一个较小的特征集合，只不过文本分类的目的是为了减少分类器的计算量，提高分类器的精度。

（4） 基于向量空间模型的表示方法

基于向量空间模型的表示方法是指用关键词向量空间中的向量来表示用户模型的方法。向量空间模型（Vector Space Model）是 60 年代末由GeraldSalton等人提出的，它是一种知识表示方法[76]，是一个文档表示的常用方法，也是最早最出名的一个数学模型。在该模型中，文档被看作是由一组正交词条向量所组成的向量空间，每个文档 D 表示为其中的一个范化特征向量D=(t_1,w_1(D)；t_2,w_2(D)；…；t_n,w_n(D))，其中 t_i 为词条项，$w_i(D)$ 为 t_i 在 D 中的权值。t_i 可以是 D 中出现的所有单词，也可以是 D 中出现的部分单词，$w_i(D)$ 一般被定义为 t_i 在 D 中的出现频率的函数。基于向量空间模型的表示法能够很好的表示用户兴趣模型，但是随着用户兴趣的增加，用户模型会不断地增大，因而基于VSM的用户模型表示方法需要大量的空间和计算开销。

（5） 基于本体的表示法

本体是从哲学领域借鉴过来的术语，一种存在的系统化解释，是共享概念模型的明确的形式化规范说明。在知识工程领域中，有关研究者将本体定义为对概念化对象的明确表示和描述。由于本体对特定领域对象的表示与描述具有规范性、可重用性、可靠性等特点，本体被应用于信息检索领域，对文档、用户模型进行描述，以提高系统的精确性。本体从不同层次的形式化模式，给出这些词汇（术语）和词汇间相互关系的明确定义,通过概念之间的关系来描述概念的语义。

（6） 粗兴趣粒度表示和细兴趣粒度表示法

用户模型的粗、细兴趣粒度表示是根据用户模型表示用户兴趣的信息粒度来区分的，粗兴趣粒度表示是指在用户兴趣仅仅被分为用户感兴趣类和用户不感兴趣类，用户兴趣建模就被视为一个二类归纳学习问题，需要正、反例集对学习算法进行训练。这种方法简单而且直观，但是存在这样的两个问题：

①将用户所有感兴趣的信息归为用户感兴趣类、所有不感兴趣的信息归为用户不感兴趣类，会降低用户模型的精度。用户感兴趣的信息在主题上可能差别很大，如：

乒乓球和医学，将这些内容迥异的兴趣类混杂在一起使得用户兴趣模型容易受各兴趣类样本分布与数量的影响，导致样本较少的兴趣类在用户模型中得不到反映。用户不感兴趣类涉及的内容更为广泛，有限的反例数量根本不足以覆盖用户不感兴趣类的空间。采用有限的反例集训练出来的系统在面对新的用户不感兴趣类的信息时难以处理。因此，将用户兴趣类别简单的划分为用户感兴趣类和用户不感兴趣类两类容易导致构建出的用户模型与真实模型相差很远，降低个性化服务的质量。

②正、反例集的获取是以牺牲用户的正常浏览和引入噪声样本为代价。正、反例集的获取通常采用强交互法和简单推测法。强交互法是指每个正例和反例的获取都需要用户与系统的交互才能得到。简单推测法是指根据用户对超链接的选择与否推测用户对页面是否感兴趣。

细粒度表示则要在用户模型中区分用户的兴趣主题。在现有的个性化服务系统中，用户模型大多采用兴趣粒度表示法，这是因为粗兴趣粒度对用户兴趣建模实现起来较为简单，但是细粒度用户模型更能细致地刻画用户的兴趣和偏好，可以很好的为用户提供个性化服务。

（7） 语义网表示方法

语义网表示方法是通过带弧线的节点表示用户兴趣和其上下文，节点表示词，节点之间具有弧线表示这两个节点在相同的文档中出现，不同于知识表示领域的语义网表示方法。语义网络表示方法不仅包含了用户感兴趣的关键词，还包含关键词之间的同现关系，同现关系的出现在一定程度上表达了概念之间的某种关系。

典型的系统有SiteIF[77]、ifWeb，在这两个系统中，对应文档结构化的部分（如主机、文档大小、图片数目等）由属性-值对来表示，用户模型由属性-值对的集合以及加权的语义网络组成，语义网络的节点对应文档中的术语（概念），术语之间的弧线表示这两个术语在相同的文档中同现。但是语义网仅仅考虑了在相同文档中的“同现”这种关系，对于描述概念之间的真实关系是远远不够的。

（8） n-grams 表示方法

n-grams方法不再是单纯地通过独立的词来描述兴趣，而是考虑词与词之间的相关性，通过词与词之间的同现关系来表达兴趣，能够表达一定的语义信息，并在一定程度上体现兴趣间的相关性，但由于其是基于统计的思想，因此需要分析大量的用户兴趣的文本才能获得用户兴趣的n-grams表示。PSUN是典型的n-grams表示方法[77]，通过给出一些用户感兴趣的文档来向系统提供初始的用户概貌，重复出现的词通过上述

n-grams方法存储，如果n个词多次同时出现，则可提供一些上下文关系。n-grams存储在一个相互吸引或排斥的词的网络上，用同时出现的程度来确定相互吸引的程度。

（9） 固定文章集的表示方法

用固定文章集来表示用户的兴趣[78]是由清华大学卢增祥提出的，固定文章集是指从近似总体文章中集中选择最具有代表性的固定子集，该子集能够充分反映某一领域中的各种用户的需求。在这种方法中，用户可以通过评价一些专门选择的文章来表达自己的信息需求。该方法的实质也是关键词的表示法，处理时要将用户感兴趣的文章集转换成用户的兴趣关键词集合。

1.3.2 用户兴趣建模技术的分类

根据建模过程中用户参与程度的不同，用户兴趣建模技术可以分为用户手工定制建模、示例用户兴趣建模和自动用户兴趣建模。

（1） 用户手工定制建模

用户手工定制建模是指用户模型由用户自己手工输入或选择的一种用户兴趣建模方法，如用户输入其感兴趣的关键词，或者选择其感兴趣的列表。在个性化服务发展的早期，用户手工定制模式是用户兴趣建模的主要方法。其实现方法简单，但是存在如下问题：

①用户难以全面、准确的表达自己感兴趣的关键词，从而导致构建的用户模型不够准确。

网站设计者按照自己的理解组织网站结构，有些栏目可能包含了用户感兴趣的信息，但是用户根据自己的理解认为该栏目并不包含自己感兴趣的信息，这样导致了用户不能准确地定制用户模型。当用户自己输入其感兴趣的关键词作为用户模型的时候，用户列出一系列用户感兴趣的关键词，但是却未必能够详尽准确的表达。

②用户兴趣发生改变时，需要用户重新输入。

用户的兴趣是随时间的推移而发生变化的，一些用户曾经感兴趣的主题会被渐渐遗忘，新的兴趣会不断产生。用户兴趣的动态变化不能在静态的用户手工定制的用户模型中实时的更新。时间越长，手工定制的用户兴趣模型就越不能正确反映用户的真实兴趣，也就越不能为用户提供很好的个性化服务。

③ 手工定制方法完全依赖用户，容易降低用户使用系统的积极性。

心理学研究表明，用户不愿意提供明确的反馈信息。对用户来说，系统的易用性

是用户衡量服务质量的重要标准，任何一种服务，不管其性能如何，如果用户需要付出很多努力才能享用这种服务的话，用户一般就会放弃这种服务。

（2） 示例用户兴趣建模

示例建模是指由用户提供与自己兴趣相关的实例集和类别属性来建立用户模型的一种建模方法。示例一般通过用户在浏览过程中对浏览过的页面标注感兴趣，不感兴趣的程度得到。浏览过的页面及相应的标注就是用户兴趣建模的示例。这样，在用户的浏览过程中，就需要用户对浏览过的页面标注感兴趣的程度，严重的干扰了用户的正常浏览，降低了为用户提供个性化服务的易用性。理想的用户兴趣建模方法应该是无需用户主动提供任何信息，系统根据用户的浏览内容和用户的浏览行为自动地为用户构建用户兴趣模型。

（3） 自动用户兴趣建模

自动用户兴趣建模是指建模过程中无需用户输入其感兴趣的关键词，也无需用户标注其对浏览过的页面的感兴趣程度，根据用户的浏览内容和浏览行为自动地为用户构建用户兴趣模型。在现有的个性化服务系统中，采用自动用户兴趣建模方法构建用户模型的系统主要有卡内基-梅隆大学的 Personal WebWatcher、德国国家研究中心的 ELFI、麻省理工学院的 Letizia 等。总的说来，自动用户兴趣建模有利于提高个性化服务系统的易用性，促进个性化服务的发展。

1.3.3 用户兴趣建模技术的国内外研究现状

MyYahoo 是采用手工定制建模方法建立用户模型的典型代表。在用户登录 MyYahoo 站点后，用户须从成百上千的栏目中手工选择其感兴趣的栏目。

麻省理工学院的Lieberman[79]通过用户的行为推测用户对页面的兴趣。比如，如果用户打印某个页面，则推测用户对该页面感兴趣；如果用户经常访问某个页面，则推测用户对该页面感兴趣；如果用户在某个页面上的停留时间很短，那么推测用户对该超链接对应的网页不感兴趣，等等。这样，将用户感兴趣的页面中的一些关键词构成用户模型。

WebWatcher[80]也是采用手工定制方法对用户兴趣建模。它要求用户输入感兴趣的关键词，使用用户提供的对超链接的建议获得用户的评价，帮助用户确定信息的位置。该系统将输入的关键词作为用户模型，然后进行个性化推荐。它不断地给用户推荐一系列站点并建立超链接，伴随用户浏览网络。如果用户知识某次检索结果成功，就用

代表用户兴趣的关键词为每一个超链接加上注释，存入知识库。

Pazzani和Billsus通过用户对其浏览页面的标注的用户感兴趣和不感兴趣的页面作为训练样本集，设计的Syskill&Webert[81]要求用户对每一个浏览过的页面标注“感兴趣”、“不感兴趣”或者“一般”，而后计算词的期望信息增益，选择期望信息增益大的 128 个词构成用户模型，该系统通过计算页面中单词与类别的互信息找出反映用户兴趣的关键词，构建用户兴趣模型。

1996 年，卡内基－梅隆大学[82]推出了Personal WebWatcher，它记录用户浏览的页面，观察用户对页面中超链接的选择，推断用户浏览过的页面属于感兴趣类还是不感兴趣类，分别作为训练例集的正例和反例。通过计算单词与类别的互信息选择用户模型的关键词，构建用户模型。

Adomavicius 和 Tuzhilin 采用数据挖掘的方法对用户个体的访问记录进行挖掘，得到关联规则，结合用户登记的个人信息，为用户构建用户模型。

Chan 通过观察用户对页面中超连接的选择获取用户感兴趣与不感兴趣的页面作为训练样本，而后计算单字间的期望互信息，选择期望信息大的 250 个单字构成用户模型。

Schwab 等通过观察用户对页面的选择获取用户感兴趣的页面作为训练样本集，用出现在用户感兴趣页面中指定位置的词构成用户模型。

林鸿飞和杨元生根据用户提供的各类示例文档，通过考察特征、段落和类别的表达能力构建用户模型。

总的来说，现存的用户兴趣建模方法存在这样的问题：基于向量空间模型的表示法能够很好的表示用户兴趣模型，但是随着用户兴趣的增加，用户模型会不断地增大，因而基于 VSM 的用户模型表示方法需要大量的空间和计算开销。语义网仅仅考虑了在相同文档中的“同现”这种关系，对于描述概念之间的真实关系是远远不够的。从本质上说，以上的方法大多数是在词层次上对用户兴趣建模，需要很大的空间开销，容易造成数据稀疏。虽然有的方法是在概念层次上对用户兴趣建模，但是却没有考虑到用户在网页上的停留时间这一表现用户兴趣的关键要素。

1.4 论文的研究内容和安排

本文在现有技术的基础上，对 web 个性化服务中的几个关键技术进行了研究。下面是研究的主要内容及相关章节内容的安排：

第一章是绪论。本章对选题的研究背景、意义和当前的发展状况进行了描述，讨论了国内外相关研究的研究现状，给出了本文的主要研究内容和论文的组织结构。

第二章是基于标记窗的网页正文信息提取。信息技术的飞速发展，网络信息的急剧增长，给信息的获取、存取、传递以及使用带来了一系列的问题。人们已经开始探讨一种新的资源描述方式，关于数据的数据——元数据（Metadata）。它是对万维网信息的一种描述方式，是机器可理解的信息。万维网中的海量信息是获取元数据十分重要的来源。本章分析了当前人们如何对 web 数据进行预处理的方法，提出了标记窗的概念，而后又提出了一种可以准确提取网页正文信息的基于标记窗的网页正文提取方法。

第三章是基于概念关系的用户兴趣建模。大多数用户兴趣模型的构建方法仅仅从词形、词频角度分析用户浏览的网页，没有考虑词与词之间存在的概念关系。为了反映词词之间的关系，归纳用户兴趣，本章采用“知网”作为概念知识库，对用户兴趣词条进行概括与抽象，阐述了一种基于概念关系及其权重的用户兴趣森林模型的具体构建过程及更新过程，并提出了在语义网络下，基于该用户兴趣森林模型的网页预取方法，将用户最可能访问的网页链接预先取回，以方便用户访问。

第四章是基于泊松分布的用户兴趣建模。本章给出了基于泊松分布，结合用户相对停留时间的用户兴趣模型的具体构建过程，并提出了基于该模型的网页推荐的方法，对用户可能访问的网页进行预测，并提供给用户进行选择。通过一系列实验的结果表明，此推荐方法可以产生很好的点击率。

第五章是基于改进的汉宁窗函数的信息检索模型。通过分析布尔模型、向量空间模型及概率模型的缺陷，本章以自然语言作为检索语言进行语义检索，结合用户兴趣模型，考虑检索词的含义、顺序及词密度，提出了基于改进的汉宁窗函数的信息检索模型，并给出了基于改进的汉宁窗函数的信息检索算法。此算法提高了信息检索结果的精度和检索有效性，为用户提供更好的个性化检索服务奠定了基础。

第六章是基于 RSS 的个性化信息推送。本章提出了基于 RSS 的个性化信息推送，通过推拉技术的结合，将用户需要的信息主动推送给用户。使得信息流不再是用户单一的“拉”，还包括反方向的“推”，这样，用户避免了网上漫无边际的查找与长时间的等待。通过一个用户需求为将标题为”model”的信息组织起来的实例证明了，将 RSS 技术和个性化技术相结合的基于 RSS 技术的个性化信息推送方法是可行、有效的，可以为用户提供方便的个性化服务。

最后，对本文的工作进行了全面的总结，并讨论了今后需要进一步研究的内容。

第二章 基于标记窗的网页正文信息提取

2.1 信息抽取技术

虽然万维网信息量巨大、信息的表达方式千差万别、内部信息之间关系错综复杂，但基本上都是用标记语言 HTML 和 XML 书写的。当前的静态万维网信息多以超文本标记语言（HyperText Markup Language, HTML）来书写与发布，它是通过在普通文本的基础上加上特殊的标记（tag）完成的。所有的标记都是由“<” 和“>”括起，如<a>，在开始标记的标记名前加上符号“/”既是其终结标记，如。

一个普通的 web 页面就是一个 HTML 文本文件，它被组织成一棵树的结构，其树根是一对<HTML>和</HTML>标记，所有内容都在这对标记之内。网页中出现的表格是使用了 TABLE 标记，显示一个表需要 3 个重要的标记：

（1） <TABLE>：容器标记，用以指明这是表格而且其余表格标记只能在这个标记的范围内才适用。

（2） <TR>：用以表示表格的行。

（3） <TD>：用以表示表格行中的单元。

信息抽取（Information Extraction）技术是近十几年来发展起来的新领域，起源于文本理解，属自然语言处理研究的领域。信息抽取是直接从自然语言文本中抽取事实信息，并以结构化的形式描述信息，适用于信息查询、文本深层挖掘、问题自动回答等方面的应用。Web 信息抽取（Web Information Extraction，简称为 WebIE）是将 Web 作为信息源的信息抽取，其核心是抽取分散在 Internet 上的半结构化的 HTML 页面中的隐含信息。

随着万维网的飞速发展，其上的信息源随之日益丰富。然而，Web 页面中经常含有广告链接、导航条、版权等非网页主题信息的内容，页面所要表达的主要信息经常被隐藏在无关的内容和结构中，为 web 信息的提取带来了很大困难，限制了 web 信息的可利用性。正确提取网页正文信息，实际上就是提取出页面要表达的主要内容，是信息搜索（Information Search）等 Web 信息处理的基础。

传统的网页数据抽取方法通常是由包装器（Wrapper）完成的。但是，获取包装器中信息模式识别的知识是一个瓶颈问题。采用半自动化方法获取知识规则的XWRAP系统在进行网页抽取之前，对网页进行检查并进行预处理，最后将网页表示成一棵树

[83]。

在 Web 信息提取领域，已经有大量的研究工作，包括 HTML 结构分析方法、基于自然语言处理的方法、机器学习和 ONTOLOGY 等。

王琦等[84]基于DOM规范，提出了基于语义信息的STU-DOM树模型，将HTML 文档转换为STU-DOM 树，并对其进行基于结构的过滤和基于语义的剪枝完成对网页的主题信息的提取。

Kristina Lerman等[86]提出了通过对行和列的分组，从list和table中自动提取web数据。但是，此方法在一些假定条件成立的情况下才能进行，而且需要分析许多网页之后，才能从单一的一个list中提出信息。

崔继馨等[87]提出了基于DOM的web信息抽取方法，采用人工方式对样本页面附加语义信息，然后对样本页面中的样本记录进行标记，通过机器学习的方法产生信息抽取规则，利用这些规则完成对相似结构网页的信息抽取。由于此方法需要人工参与，故使得系统的可用性降低。

文献[88]提出了基于统计的网页正文信息提取方法，它根据网页中的 HTML 标记将网页表示成一棵树，然后利用树中的每个结点包含的中文字节数从中选择包含正文信息的结点。此方法适用于网页中的所有正文信息都放在一个 td 中的情况下的网页正文提取。

Finn等人[89]将HTML文档看作字符和标签组成的序列，在字符集中的区域提取文字。这种方法仅适合主题文字集中的网页，不能有效处理段落间有表格或链接等标签丰富的结构。

Kaasinen等[90]提出了Desk-Card模型，将网页（Desk）分为若干Card，每次显示一个Card，减少了页面大小，但是没有提取出信息，用户需要阅读多个Card才能确定主题。

Buyukkokten等[91-92]提出了STU（Semantic textual unit）模型，STU对应网页中的块（block），将网页分割为平行的STU。此方法改变了源网页的结构和内容，而且没有提取出主题信息，保留了无关的文字和链接。

Gupta等[93]从网页中删除无关部分，但是在删除链接时没有考虑上下文的语义信息，使得提取的网页正文信息不完整。

通过分析可知，现存的网页正文提取方法不能处理网页的正文部分被存放在多个 td 中的情况；不能处理一个 td 中含有不同内容的情况，就是说，不能处理一个 td 中

存放的不仅仅是网页正文的情况，不能处理网页正文文字长度很短，短到和网页其余部分文字 （如：广告，导航条）长度相当的特殊情况。另外，现存方法都限于提取存放在 td 中的网页正文信息，但经过统计，存在大量未采用 table 结构存放正文信息的网页，对于这种情况，上述方法都无能为力。基于此，本章提出了基于标记窗（Tag Window）的网页正文信息提取方法来解决上述问题。

2.2 基于标记窗的网页正文信息提取方法

一般来说，用户的登录文件中包含：用户的 IP 地址；访问的日期和时间；请求方法（GET，POST）；访问网页的 URL；协议（HTTP1.0，HTTP1.1）；返回码；传送的字节数等。如表 2.1 所示。

表 2.1 用户登录文件格式

Tab.2.1 the form of the user profile

请求源	用户 ID	请求日期及时间	方法，URL，HTTP 协议	状态码	传送的字节数
211.68.2.8	-	[23/Oct/2004:10:16:23-0500]	“GET/HTTP/1.0”	200	3897
211.68.2.8	-	[23/Oct/2004:10:16:45-0500]	“GET A.gif HTTP/ 1.1 ”	200	5639
59.73.75.253	-	[23/Oct/2004:10:17:11-0500]	“GET B.html HTTP/ 1.1 ”	200	29816
211.68.2.8	-	[23/Oct/2004:10:17:36-0500]	“GET C.html HTTP/ 1.1 ”	200	31786
59.73.75.253	-	[23/Oct/2004:10:17:57-0500]	“GET A.gif HTTP/ 1.1 ”	200	5639

将服务器日志转换成记录，存放在数据库中。也就是说，用户的每个访问记录对应于数据库中的一条记录。（将原始的 web 服务器日志文件转化为一系列分用户的会话过程，一个会话过程就是同一个用户的 IP 在一段时间内的连续 web 请求。）

为了确定访问网站的用户，登录文件应当包含登录到服务器或者用户自己的计算机的用户的 ID。然而，大多数网站，用户无须登录就可以访问，大多数 web 服务器也不需要用户在自己的计算机上提出验证用户登录身份的请求。这样，根据 HTTP 标准获得的可用信息不足以区分出来自相同主机或代理的不同用户。由 Internet 服务提供者（ISP）分配的 IP 地址或者代理服务器也给用户的唯一识别带来了难题。对于这个问题，最好的解决方法就是利用 cookies。cookie 是和 web 站点相关的代码。只要用户使用同一个浏览器，这个识别方法足够识别出发出每一个 URL 请求的用户。由于在本文的数据集中，这些 cookies 是不可用的，所以，本文进行了假定，一个 IP 对应一个用户。因为在以后的清洗步骤中，无须 HTTP 协议版本信息，所以，在此步移除了这一信息。

将服务器登录文件 $L=L_1,L_2,\cdots,L_{|L|}$，其中 $L_i=(IP_i,TIME_i,URL_i,PROT_i,STATUSCODE_i,BYTES_i)$，($L_i\in L$，$i\in[1,\cdots,|L|]$) 转化成一系列事务 $T=T_1,T_2,\cdots,T_{|L|}$，其中，$T_i=(UID_i,TIME_i,URL_i,STATUSCODE_i,BYTES_i)$，($T_i\in T$，$i\in[1,\cdots,|L|]$)。$|L|$是登录文件$L$中的登录记录数，$UID_i$是用户识别号。对于每一个相同的 IP 地址，分配一个唯一的用户识别号。也就是说，T 中的一些用户请求可能拥有同一个用户识别号。

在用户请求浏览某一特定的网页的时候，除了 HTML 文件外，图形和脚本也随之被下载到到客户端，这样就可能导致存在多个登录网页的入口。由于只想对用户的浏览行为进行分析，所以，那些用户没有明确请求的文件请求毫无意义，移除了扩展名为.gif，.bmp，.jpg，.ico，.png，.css，.swf，.cab，.rar，.js 的文件。此外，本文进行了规范化登录文件中的 URL 的操作。大多数 web 服务器将目录请求视为象“index.html”或者“home.html”一样的默认文件请求。目录请求可以跟随也可以不跟随斜线。意思是说，请求“www.163.com”，“www.163.com/”，“www.163.com/index.html”，“www.163.com/home.html” 的内容是一样的。

服务器返回的状态码是对用户请求的反映，400 系列的状态码意味着失败，500 系列的状态码意味着服务器错误。基于此，移除了状态码是 400 系列和 500 系列的登录文件。下一步进行的是提取 HTML 文件，仅保留“GET…HTML”格式的 URL 网页请求。本文应用基于标记窗的网页正文提取算法对这部分文件进行网页正文的提取。

基于标记窗提取网页正文信息的方法不仅适合于处理一个网页中所有的正文信息都放在一个 td 中的情况，也适合于处理网页正文放在多个 td 中的情况，尤其重要的是，它能够解决非 table 结构的网页正文提取问题，实现简单，通用性好。

定义 2.1 标记窗（Tag Window）：称 HTML 中成对出现的标记为标记对，称 HTML 格式的网页中出现在 title 之后的显示内容非空的标记对为标记窗。

Internet 上的许多页面都是 HTML 格式的，这些页面由一系列 HTML 标记内容组成。W3C 组织将 HTML 页面定义为一层层标记的嵌套体，如：<html><head><title></title></head><body></body></html>的形式，也就是说，有一个开始标记就有一个结束标记和它对应。但是，实际中的 HTML 并不完全遵守这样的规

范化格式，只有开始标记没有相应结束标记的 HTML 往往也能显示网页的正确内容。因特网中存在着的这些大量不规范的网页给网页内容分析工作带来了极大的困难。因此，在进行网页内容提取工作之前必须对网页进行规范化。

基于标记窗提取网页正文信息的方法流程图如图 2.2 所示，具体步骤如下：

第一步，对网页进行规范化处理

一个网页是规范化的网页，如果它满足如下条件[94]：

（1） 在除了网页标记 tag 的地方出现的“<”和“>”用&lt；和&gt；代替；

（2） 所有标记的属性值放在引号中，如；

（3） 所有的标记都是匹配的，也就是说，每个开始标记都对应着一个结束标记，如：<body>，</body>；

（4） 所有的标记都是正确嵌套的，如：<a>………。

第二步，获取网页的title及其各级标题<h_1>，…，<h_g>的标题内容，对它们应用中科院的ICTCLAS分词系统进行分词处理，去掉停用词、虚词，得到只包含实词的标题词序列 $S_{title}=\{W_{t_1},W_{t_2},\cdots,W_{t_m}\}$ 及各级标题词序列 $S_{h_i}=\{W_{h_{i_1}},W_{h_{i_2}},\cdots,W_{h_{i_n}}\}$(i=1, …, g)。

第三步，提取网页正文

（1） 找出所有的标记窗 tw，对每一个标记窗 tw_i(i=1, …, N)，去掉其中的 HTML 标记，得到不含任何 HTML 标记的字符串；

（2） 对每个标记窗 tw_i 内的字符串分词，得到字符串序列。为了简化计算，仅取出 S_{tw_i} 中的实词，得到 $S_{tw_i}=\{W_{tw_{i_1}},W_{tw_{i_2}},\cdots,W_{tw_{i_q}}\}$。用Levenshtein Distance公式[95] (如公式 2.2 所示)，计算标题词序列 S_{title} 与字符串词序列 S_{tw_i} 的距离。

$$L\left(\left(x^1,\cdots;x^p\right),\left(y^1,\cdots;y^q\right)\right)=\begin{cases}p & q=0\\ q & p=0\\ \min\left(L\left(\left(x^1,\cdots;x^{p-1}\right),\left(y^1,\cdots;y^q\right)\right)+1,L\left(\left(x^1,\cdots;x^p\right),\left(y^1,\cdots;y^{q-1}\right)\right)+1,\right. & \text{否则}\\ \left.L\left(\left(x^1,\cdots;x^{p-1}\right),\left(y^1,\cdots;y^q-1\right)\right)+Z\left(x^p,y^q\right)\right)\end{cases} \tag{2.2}$$

其中，$Z(x,y)=\begin{cases} 0 & \text{如果} x=y \\ 1 & \text{否则} \end{cases}$

（3） 比较 L 和 q 的大小，如果 $L<q$，则标记窗 tw_i 中的文字是正文信息，将其提取，否则，舍弃标记窗 tw_i 中的文字。

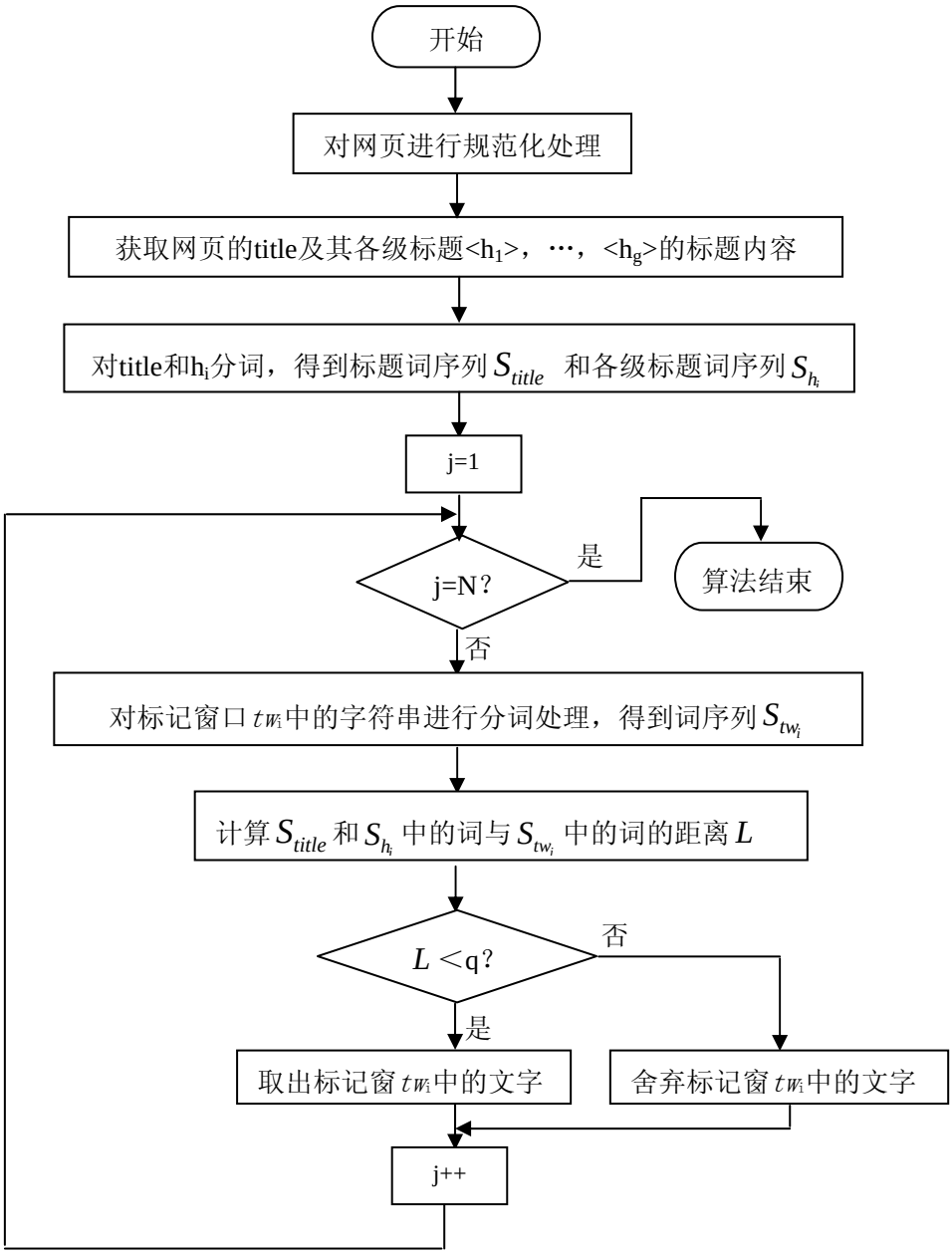

图 2.1 基于标记窗的网页正文信息提取方法流程图

Fig 2.1 The flowchart of the web content information extraction method based on Tag Window

对于标记嵌套的情况，先处理最里层的标记对，提取其中的正文，并且清除此标记对及其中间的正文，然后，处理外一层的标记对，依此类推。

下面用一个实际的例子来说明基于标记窗提取网页正文内容的方法。该例子所用的网页是http://weather.sina.com.cn/news/2005/1216/12339.html，此网页的标题是“新疆发布大雾黄色预警信号”。

过滤掉该网页对应的 HTML 文件中显示内容为空的标记对之后，用得到的标记窗中的 4 个标记窗为例说明。对每个标记窗分词之后得到词序列（仅保留实词）、标题词序列以及他们之间的距离如表 2.2。按照前述算法，应该提取标记窗 1，2，3 中的文字信息，舍弃标记窗 4 中的信息。

表 2.2 一个网页实例

Tab.2.2 A webpage instance

标记窗序号 No. tw	标题词序列 S_{title}	对标记窗内的文字进行分词处理得到的词序列 S_{tw_i}	S_{title} 与 S_{tw_i} 的距离	q
1	新疆 发布 大雾 黄色 预警 信号	新疆 发布 大雾 黄色 预警 信号	0	6
2		新疆 气象台 今日 15 时 30 分 发布 大雾 黄色 预警 信号	4	10
3		今天 中午 乌鲁木齐 市区 开始 出现 浓雾 能见度 为 300 米 预计 浓雾 将 维持到 明天 上午 希 各 有关 单位 做 好 防 雾 工作 注意 交通 安全	29	30
4		责任 编辑 漾 绕	5	4

2.3 实验与结果分析

2.3.1 基于标记窗的网页正文信息提取方法的准确率

为了考察本章提出的方法的实际效果，随机选择来自www.sina.com.cn，www.sohu.com，www.bit.edu.cn，www.people.com和www.mop.gov.cn网站的 788 个网页进行了实验，实验结果如表 2.3 所示。正确提取的网页数是指将网页的全部正文信息正确提取出来的网页的个数，错提、少提正文信息的网页都是错误提取的网页。

$$准确率=\frac{正确提取的网页数}{网页总数}。$$

表 2.3　基于标记窗的网页正文信息提取方法的准确率

Tab 2.3 the accuracy of the web page content information extraction method based on Tag window

网页来源	网页总数	正确提取的网页数	错误提取的网页数	准确率
www.sina.com.cn	198	183	15	92.4%
www.sohu.com	186	178	8	95.7%
www.bit.edu.cn	80	78	2	97.5%
www.people.com.cn	211	195	16	92.4%
www.mop.gov.cn	113	109	4	96.5%
合计	788	743	45	94.9%

通过对结果的分析发现，之所以会出现对网页正文信息的错提、少提，是因为网页设计者的想法的不同，导致他们可能使用一些修辞手法，如：比喻，拟人等手法吸引 web 访问者的浏览注意力。这样就导致了标题词序列中的词根本没有在网页正文中出现，造成了对网页正文信息的错误提取，影响了网页正文信息提取的准确率。

2.3.2、基于 DOM 的网页主题信息自动提取方法、基于统计的网页正文信息提取方法与基于标记窗的网页正文信息提取方法的准确率比较

为了说明本章提出方法的有效性，仍使用上述来自www.sina.com.cn，www.sohu.com，www.bit.edu.cn，www.people.com和www.mop.gov.cn网站的随机选择的 788 个网页，与基于统计的网页正文信息提取方法和基于DOM的网页主题信息自动提取方法进行了准确率的比较实验，实验结果如表 2.4 所示。

表 2.4　网页正文信息提取方法的准确率比较

Tab 2.4　Accuracy comparision of the web page content information extraction methods

网页来源	网页总数	基于统计的网页正文信息提取方法准确率	基于 DOM 的网页主题信息自动提取方法准确率	基于标记窗的网页正文信息提取方法准确率
www.sina.com.cn	198	92.8%	95.1%	92.4%
www.sohu.com	186	93.5%	94.6%	95.7%
www.bit.edu.cn	80	97.4%	92.8%	97.5%
www.people.com.cn	211	93.7%	94.7%	92.4%
www.mop.gov.cn	113	95.8%	96.1%	96.5%
合计	788	94.64%	94.66%	94.9%

2.4 小结

Web 上的数据抽取技术是目前热点的研究方向，虽然国内外的研究在一些技术上较为成熟和完善，但仍没有一个产品或系统在各个方面符合人们对 Web 信息抽取的要求。随着新技术和新思想的介入，Web 信息抽取技术处于不断地更新和发展中。正确提取网页正文信息，实际上就是提取出页面要表达的主要内容，是信息搜索（Information Search）等 Web 信息处理的基础。本章提出了标记窗的概念和基于标记窗的网页正文信息提取方法。基于标记窗的网页正文信息提取方法解决了网页正文存放在多个 td 中的情况，解决了正文文字短的网页的正文提取问题，尤其重要的是，它能够处理非 table 结构的网页正文提取问题。本方法无须将网页表示成一棵树，只需利用正则表达式，就可以直接提取出网页中标记对之间的正文，这大大降低了算法的复杂度。实验表明，此方法性能好，适用性强。

第三章 基于概念关系的用户兴趣建模

3.1 词语语义相似度的计算

词语相似度是一个主观性相当强的概念。词语之间的关系非常复杂，其相似或差异之处很难用一个简单的数值来进行度量，所以脱离具体的应用去谈论词语相似度，很难得到一个统一的定义。从某一角度看非常相似的词语，从另一个角度看，很可能差异非常大。相似度这个概念，涉及到词语的词法、句法、语义甚至语用等方方面面的特点。其中，对词语相似度影响最大的应该是词的语义。

相似度是一个数值，一般取值范围在[0,1]之间。一个词语与其本身的语义相似度为 1。如果两个词语在任何上下文中都不可替换，那么其相似度为 0。

3.1.1 知网

知网（英文名称为HowNet）是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库，由多个数据文件组成。在知网中，义原是最基本的、不易于再分割的意义的最小单位。知网的基本思想是：设想所有的概念都可以被各种各样的有限义原集合表示。如果能够把握这一有限的义原集合，并用它来描述概念之间的关系以及属性与属性之间的关系，就有可能建立所设想的知识系统[96]。尽管被人们称为知识词典的常识性知识库是知网的最基本的数据库，但是它不是一部语义词典，它的全部的主要文件包括知识词典构成了一个有机结合的知识系统。例如，主要特征文件、次要特征文件、同义、反义以及对义组的形成，以及事件关系和角色转换等都是系统的重要组成部分，而不仅仅是标注的规格文件。

《知网》中有两个主要的概念：“概念”与“义原”。

“概念”是对词汇语义的一种描述。每一个词可以表达为几个概念。“概念”是用一种“知识表示语言”来描述的，这种“知识表示语言”所用的“词汇”叫做“义原”。

“义原”是用于描述一个“概念”的最小意义单位，是最基本的、意思不能再分割的最小语义单位。将它们组合可以生成更大的语义单位，如词语的概念。义原之间存在相互关系，在《知网》中，一共描述了义原之间的八种关系：上下文关系、同义关系、反义关系、对义关系、属性- 宿主关系、部件- 整体关系、材料- 成品关系、

事件- 角色关系，它们组成了一个复杂的网状结构。

《知网》一共采用了 1500 义原，这些义原分为以下几个大类：Event|事件、entity|实体、attribute|属性值、aValue|属性值、quantity|数量、qValue|数量值、SecondaryFeature|次要特征、syntax|语法、EventRole|动态角色、EventFeatures|动态属性。

对于这些义原，把它们归为三组：第一组，包括第 1 到 7 类的义原，称之为“基本义原”，用来描述单个概念的语义特征；第二组，只包括第 8 类义原，称之为“语法义原”，用于描述词语的语法特征，主要是词性（Part of Speech）；第三组，包括第 9 和第 10 类的义原，称之为“关系义原”，用于描述概念和概念之间的关系。

除了义原以外，《知网》中还用了一些符号来对概念的语义进行描述，如表 3.1 所示：

,	多个属性之间，表示“和”的关系
#	表示“与其相关”
%	表示“是其部分”
$	表示“可以被该‘V’处置，或是该“V”的受事，对象，领有物，或者内容
*	表示“会‘V’或主要用于‘V’，即施事或工具
+	对 V 类，它表示它所标记的角色是一种隐性的，几乎在实际语言中不会出现
&	表示指向
~	表示多半是，多半有，很可能的
@	表示可以做“V”的空间或时间
?	表示可以是“N”的材料，如对于布匹，标以“?衣服”表示布匹可以是“衣服”的材料
{}	(1) 对于 V 类，置于［］中的是该类 V 所有的“必备角色”。如对于“购买”类，一旦它发生了，必然会在实际上有如下角色参与：施事，占有物，来源，工具。尽管在多数情况下，一个句子并不把全部的角色都交代出来 (2) 表示动态角色，如介词的定义
()	置于其中的应该是一个词表记，例如，(China|中国)
^	表示不存在，或没有，或不能
!	表示某一属性为一种敏感的属性，例如：“味道”对于“食物”，“高度”对于“山脉”，“温度” 对于“天象”等
[]	标识概念的共性属性

表 3.1 《知网》知识描述语言中的符号及其含义

Tab.3.1 the symbol and its meaning of the knowledge describing language in Hownet

知网又把这些符号分为几类：一类是用来表示语义描述式之间的逻辑关系，包括以下几个符号：, ~ ^ ，另一类用来表示概念之间的关系，包括以下几个符号： # % $ * + & @ ? ! ，第三类包括几个无法归入以上两类的特殊符号：{} () [] 。

概念之间的关系有两种表示方式：一种是用“关系义原”来表示，一种是用表示概念关系的符号来表示。按照理解，前者类似于一种格关系，后者大部分是一种格关系的“反关系”，例如“$”就可以理解为“施事、对象、领有、内容”的反关系，也就是说，该词可以充当另一个词的“施事、对象、领有、内容”。

本文利用了上下位关系及同义关系，进行后面的概念语义相似度的计算。

3.1.2 基于知网的词语语义相似度的计算

3.1.2.1 词语的语义相似度计算

对于两个汉语词语W_1和W_2，如果W_1有n个义项（概念）：C_{11}，C_{12}，……，C_{1n}，W_2有m个义项（概念）：C_{21}，C_{22}，……，C_{2m}，规定，W_1和W_2的相似度是各个概念相似度的最大值，也就是说：

$$Sim(W_1,W_2)=\max_{\substack{i=1,\cdots,n\\ j=1,\cdots,m}} Sim(C_{1i},C_{2j}) \qquad (3.1)$$

这样，就把两个词语之间的相似度计算问题归结到了两个概念之间的相似度计算问题。

3.1.2.2 义原的语义相似度计算

义原的相似度计算[97]是概念相似度计算的基础，因为所有的概念最终都归结于用义原来表示。

由于在知网中，所有的义原根据上下位关系构成了一个树状的义原层次体系，形如图 3.1。假设两个义原 p_1和 p_2在这个层次体系中的路径距离为 d，那么 p_1和 p_2之间的语义距离：

$$Sim(p_1,p_2)=\frac{\alpha}{d+\alpha} \qquad (3.2)$$

其中，α是一个可调节的参数。

- entity|实体
  ├ thing|万物
  … ├ physical|物质
  … ├ animate|生物
  … ├ AnimalHuman|动物
  … ├ human|人
  | └ humanized|拟人
  └ animal|兽
  ├ beast|走兽
  …

图 3.1 树状的义原层次结构

Fig.3.1 hierachical structure of the sememe

3.1.2.3 实词的语义相似度计算

对于实词的语义相似度计算，将其分成四个部分来计算：

（1） 第一独立义原描述式：记为 $Sim_1(C_1,C_2)$；

（2） 其他独立义原描述式：除第一独立义原以外的，所有其他独立义原（或具体词），记为 $Sim_2(C_1,C_2)$；

（3） 关系义原描述式：所有的用关系义原的描述式，记为 $Sim_3(C_1,C_2)$；

（4） 符号义原描述式：所有的用符号义原的描述式，记为 $Sim_4(C_1,C_2)$。

那么，两个实词的语义相似度：

$$Sim(C_1,C_2)=\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}Sim_j(C_1,C_2) \qquad (3.3)$$

其中，β_i（1≤i≤4）是可调节的参数，且 $\beta_1+\beta_2+\beta_3+\beta_4=1$，$\beta_1\geq\beta_2\geq\beta_3\geq\beta_4$。后者反映了 Sim_1 到 Sim_4 对词语总体相似度所起到的作用依次递减。第一独立义原描述式反映了一个概念最主要的特征，所以其权值定义一般在 0.5 以上。式 3.3 也反映了主要部分的相似度值对于次要部分的相似度值起的制约作用，也就是说，如果主要部分相似度比较低，那么次要部分的相似度对于整体相似度所起到的作用也要降低。

3.2 基于概念关系的用户兴趣建模方法

概念是关于具有共同属性的一组对象、事件或符号的知识，它可能是具体的，也可能是抽象地刻画。概念是对事物本质特征的概括和抽象[98]，概念并不是孤立存在的，

一个概念总是与其他概念之间存在着各种各样的关系，但是它并不受词汇语种、多义性和歧义性的影响，用概念表示用户兴趣，在词的概念含义层次上建立联系，不但可以准确的表示用户兴趣的本质内容，而且，利用概念的抽象性可以将多个术语归结为一个概念，有效的降低向量的维数。

3.2.1 基于概念关系的用户兴趣森林模型的构建

首先，根据第二章提出的基于标记窗提取网页正文的方法提取用户浏览过的所有网页的正文信息。

然后，对提取出来的网页正文信息进行分词处理，得到一系列网页正文的词集 $S_{text}=\{W_{text1},W_{text2},\cdots,W_{textn}\}$，计算 S_{text} 中的词 W_{texti}（i=1,2,⋯,n）出现的频率。

最后，根据概念的上下位关系及同义关系，构建含有语义关系的用户兴趣森林，流程图如图 3.2 示。

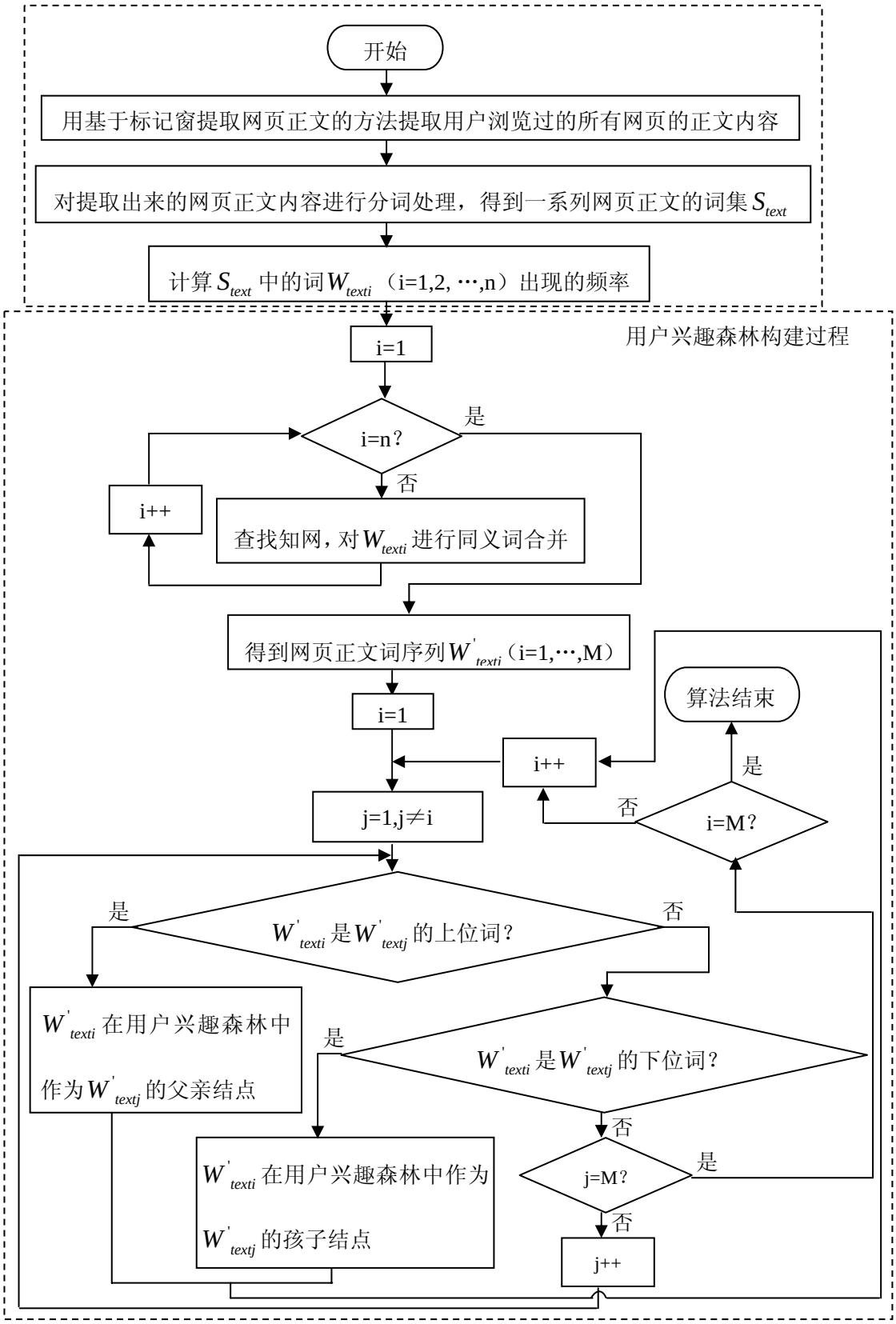

图 3.2 构建用户兴趣森林模型的流程图

Fig.3.2 the flowchart of the user interest model construction

具体的构建方法如下。

对S_{text}中的词查找知网，进行同义词合并。对于存在多个相近概念的词，将其进行同义词合并，并用序列表示，权重为这些概念相近词的词频的均值。如，计算机，电脑，微机是一组概念相似的词，它们在某一用户浏览过的网页正文中出现的频率是0.5、0.43、0.27，那么在用户兴趣森林中用（{计算机，电脑，微机}，0.4）表示一个结点。同时，查找S_{text}中的词词之间的上下位关系，如果W_{texti}是W_{textj}的上位词，那么W_{texti}在用户兴趣森林中作为W_{textj}的父亲结点，如果W_{texti}是W_{textj}的下位词，那么W_{texti}在用户兴趣森林中作为W_{textj}的孩子结点。图 3.3 即是一棵表示词的上下位关系和同义关系的树，蔬菜是植物的下位词，是豆角的上位词，包菜、甘蓝和大头菜是同义词。

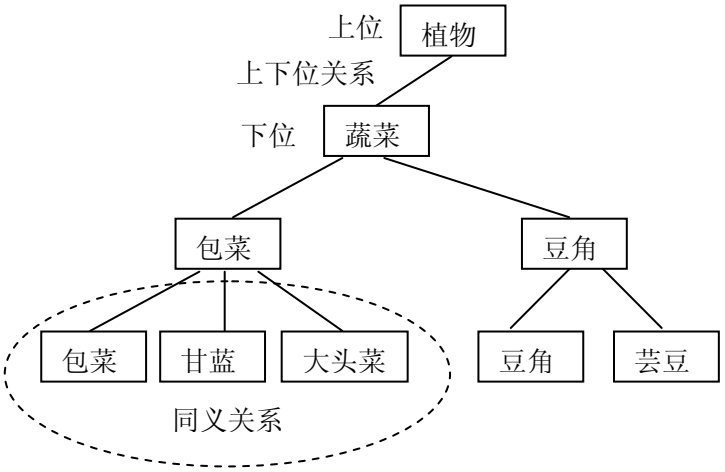

图 3.3 表示上下位关系和同义关系的树

Fig.3.3 a tree showing the hyponymy and the synonymy

这样，构建出来的用户兴趣森林的结点，如第i棵兴趣子树上的第j个结点$W_{\mathrm{int}erest_tree_{ij}}$就是用户浏览过的网页中的词集，结点的权重是这些词集出现的频率的均值。如对某用户浏览的网页进行上述步骤，得到 23 个词，根据概念的上下位关系及同义关系，构建了如图 3.4 所示的用户兴趣森林。此表征用户兴趣的用户兴趣森林由两棵兴趣树组成，一棵是体育方面的，一棵是计算机方面的。由此可知，此用户对体育和计算机方面的信息感兴趣。

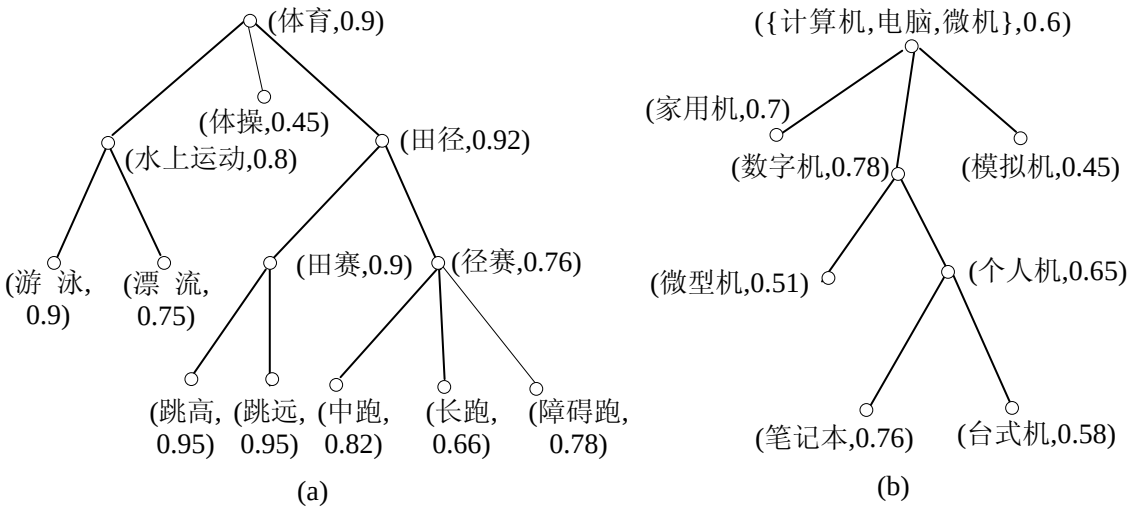

图 3.4 用户兴趣森林

Fig.3.4 user interest forest

3.2.2 用户兴趣森林的更新

由于用户在不同时期有不同的兴趣爱好，用户的知识、能力都在不断变化，为了使构建的用户模型能适应用户的这些变化，就需要根据获得的用户新信息，随时的更新用户模型。

对用户每次访问的新网页进行标题和正文提取，然后对其进行分词，得到词序列$\{W_i\}$(i=1,…,N)。调整预先构建的用户兴趣森林中结点的权重，得到表征用户兴趣的新的用户兴趣森林。

设$Weight_{pri}$表示结点的初始权重，fre表示用户浏览的当前网页的中该结点的词频，$Weight_{cur}$表示结点的当前权重，那么，调整结点权重的方法如下（流程图如图3.5 所示）。

Procedure（）

{

if（词W_i 未在用户兴趣森林中出现）

then

{

在知网中查找W_i的同义词W_{syn_i}，

if （W_{syn_i}在用户兴趣森林中出现）

then

{

将W_i作为W_{syn_i}序列中的一员加入到用户兴趣森林中；

按照公式 3.4 调整W_{syn_i}序列结点的权重

$$Weight_{cur\_W_{syn\_i}} = \frac{1}{2}\left(Weight_{pri\_W_{syn\_i}} + fre_{W_i}\right); \quad (3.4)$$

}

else if （W_{syn_i}也未在用户兴趣森林中出现）

then

{

查找W_i的上下位关系，将其作为一个新的结点插入到兴趣森林的合适位置；

$Weight_{cur\_W_i} = fre_{W_i}$；

}

}

else if （词W_i在用户兴趣森林中出现）

then

按照公式 3.5 调整结点W_i的在用户兴趣森林中的权重，

$$Weight_{cur\_W_i} = \frac{1}{2}\left(Weight_{pri\_W_i} + fre_{W_i}\right) \quad (3.5)$$

}

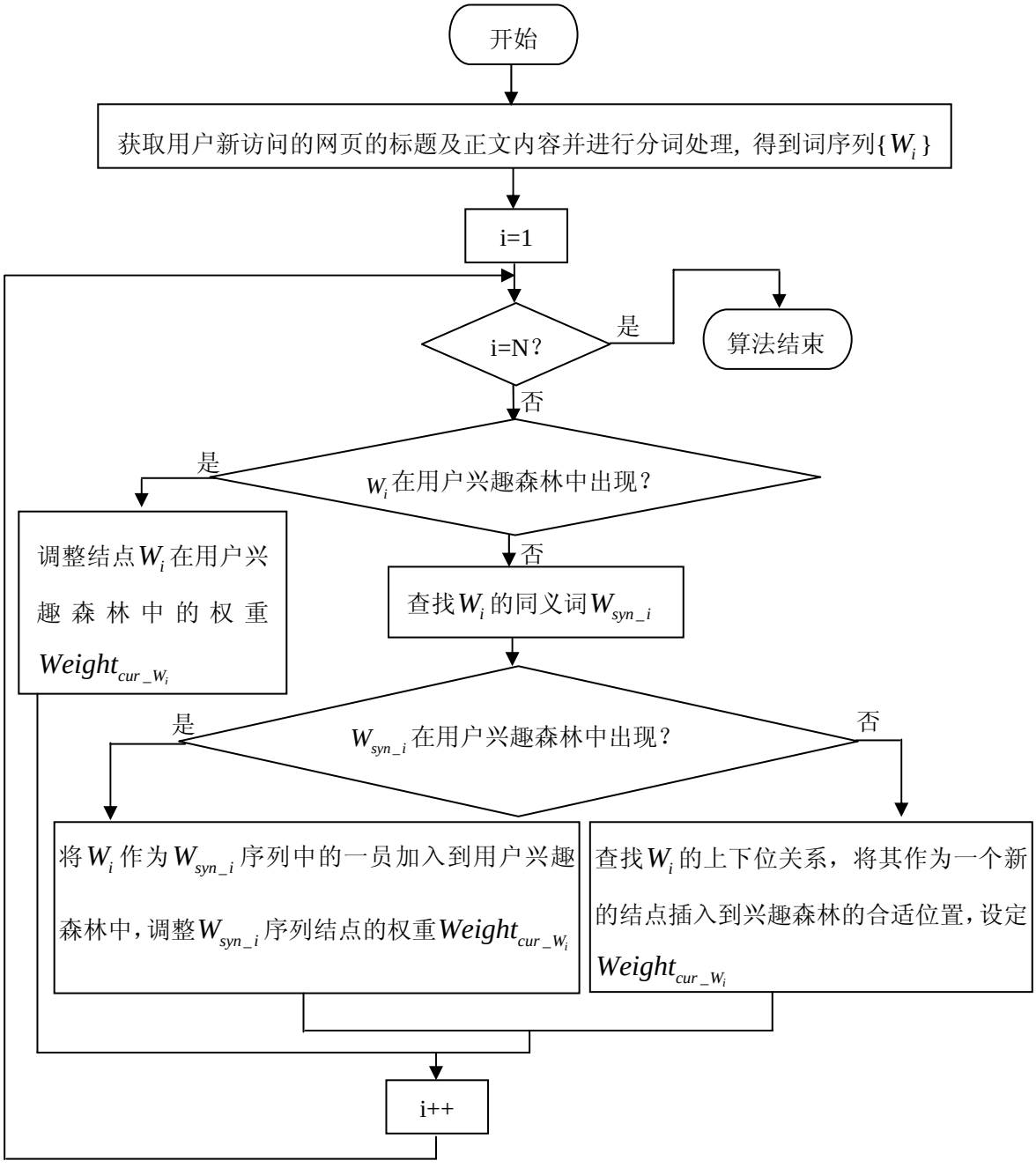

图 3.5 更新用户兴趣森林模型的流程图

Fig 3.5 The flowchart of the user interest forest model updating

3.3 基于用户兴趣森林模型的网页预取方法

现存的预取方法有的未实现在用户浏览过程中的在线学习，有的需要模型对预取网页的范围预先采取分类处理，有的方法不能实现对未被访问过的文档进行预取分析。本章提出了一种语义网络下的用户兴趣建模思想，挖掘用户访问历史中蕴涵的语义相关关系，建立基于用户的语义网络下的用户兴趣模型，并基于语义网络下的用户

兴趣模型设计了基于用户兴趣森林模型的网页预取方案，将用户最可能访问的网页链接预先取回，以方便用户访问。特别的，本方法计算了词序列$S_{url_text_i}$中词对之间的语义相似度，主要是得到超链接描述文字url_text_i的平均带权语义距离，目的是考察每个超链接描述文字url_text_i表意的确定程度。如果超链接描述文字url_text_i描述的不是一个确定信息，那么预取这样的超链接毫无意义。

定义 3.1 结点间的路径长度l_{pq}：结点p,q之间的最短路径所经过的边的数目，也就是连接p,q之间的线段的个数称为结点p,q间的路径长度。

定义 3.2 结点的深度H_p：根结点的深度是 1，其余结点的深度H_p等于其父结点深度H_{par}加 1。

定义 3.3 带权语义距离D_{pq}：称结点p,q的语义距离与他们之间的路径长度的比值及结点p,q最近公共父结点在树中的深度的加权表达式为结点p,q之间的带权语义距离，即：

$$D_{pq}=\alpha^{Sim\left(W_p,W_q\right)\Big/ l_{pq}}+\beta H_{par}\ , \tag{3.6}$$

其中，α和β为常量（文中设$\alpha=0.55$，$\beta=0.45$），H_{par}是p,q的最近公共父结点在树中的深度。

下面具体给出基于用户兴趣森林模型的网页预取方法的步骤。

步骤 1 获取用户当前浏览网页中包含的每个超链接描述文字url_text_i的内容，并进行分词处理，得到url_text_i的词序列$S_{url_text_i}=\left\{W_{url_text_{i1}},W_{url_text_{i2}},\cdots,W_{url_text_{in}}\right\}$。

步骤 2 对每个$S_{url_text_i}$，进行如下操作：

首先，计算$S_{url_text_i}$中的词$W_{url_text_{ij}}$与构建的用户兴趣森林中每个结点$W_{interest_tree_j}$的语义相似度，将用户兴趣森林中与$W_{url_text_{ij}}$具有最大语义相似度的词组成与$S_{url_text_i}$词集相对应的词集$W_{interest_treei}=\left\{W_{interest_treei_j}\right\}$(j=1,⋯,n)，并按照$W_{interest_treei_j}$的权值$Weight_{interest_treei_j}$大小降序排列$W_{interest_treei_j}$，设定排序之后的词集为$W'_{interest_treei}$；

然后，顺序计算$W'_{interest_treei_j}$中的词$W'_{interest_treei_p}$与词$W'_{interest_treei_q}$$\left(1\le p\le n,1\le q\le n,p\ne q\right)$之间的路径长度$l_{pq}$，如果词$W'_{interest_treei_p}$与词$W'_{interest_treei_q}$不在一棵用户兴趣树上，那么设定词对之间的距离是常量M (文中设其为 1000)。

最后，计算$S_{url_text_i}$中词对之间的语义相似度，进而得到$S_{url_text_i}$中词对之间的带

权语义距离 $D_{i_{pq}}$，得到超链接描述文字 url_text_i 的平均带权语义距离，如公式 3.7 示：

$$D_i = \frac{\sum_{p=1,q=1且p\neq q}^{n} 2D_{i_{pq}}}{n(n-1)} \tag{3.7}$$

步骤 3 按照公式 3.8，计算 url_text_i 的评价函数 f_i 的值，

$$f_i = aD_i + \frac{b\sum_{j=1}^{n} Weight_{\text{interest\_tree}i_j}}{n}, \tag{3.8}$$

其中，a,b 为常量(文中设 a=0.45，b=0.55)。

步骤 4 降序排列 f_i 的值，预取前 20%的 f_i 对应超链接。

3.4 实验结果与分析

3.4.1 基于用户兴趣森林模型预取方法的平均命中率和平均漏取率

为了验证基于用户兴趣森林模型预取方法的可行性和有效性，对 10 个用户从 2005 年 12 月 1 日至 2005 年 12 月 31 日连续 1 个月内的网页访问记录进行分析，构建了相应的用户兴趣森林模型。当这 10 个用户访问网页时，根据构建的用户兴趣森林，进行网页预取。定义命中率和漏取率作为预取模型的性能评价指标。命中率（hit_ratio）是指用户点击同时也被预取的链接数与预取的链接总数的比值，漏取率（miss_ratio）是指用户点击但是未被预取的链接数与预取的链接总数的比值。设 C_{user} 是用户点击的链接， P 是预取的链接，则

$$hit\_ratio = \frac{|C_{user} \cap P|}{|P|}, \tag{3.9}$$

$$miss\_ratio = \frac{|C_{user} - P|}{|P|}。 \tag{3.10}$$

实验结果如图 3.6 所示。本预取方法的平均命中率在 61%左右，不低于现存方法的命中率，但仍存在对某一用户的预取命中率较低的情况。之所以出现这种情况，是因为发生了用户“兴趣漂移”现象，也就是说随着时间的推移，用户的兴趣发生了变化，用户原本感兴趣的主题渐渐被遗忘，新的兴趣主题不断产生，导致先前构建的用户兴趣森林已经不能准确的表达用户的兴趣。

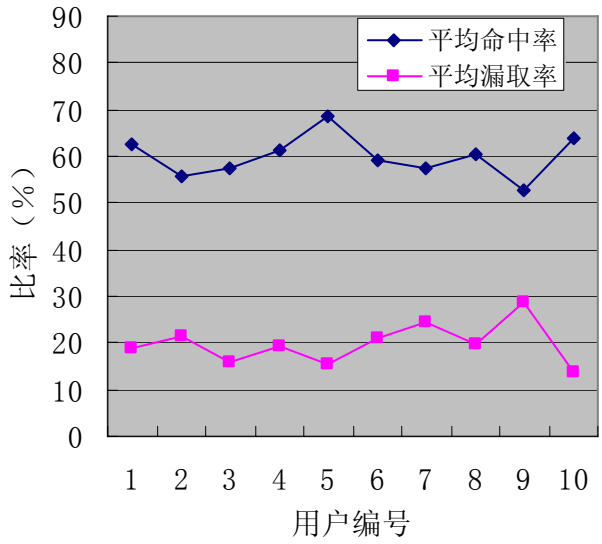

图 3.6 基于用户兴趣森林模型预取方法的平均命中率和平均漏取率

Fig.3.6 the average hit-ratio and the average miss-ratio of the web prefetching method based on user interest forest model

3.4.2 简单兴趣模型、概念联想网络模型、兴趣森林模型预取命中率比较

为了验证本章提出的方法，选择了 10 个用户计算机中的缓存内容进行对比分析。分别用简单兴趣模型(词条，权重)，概念联想网络模型和本文的概念兴趣森林模型对用户的兴趣建模。然后根据模型对网页进行预取，并对用户点击情况进行跟踪记录，分别统计它们的命中率并进行比较。命中率是指用户点击同时也被预取的链接数与预取的链接总数的比值。实验结果如图 3.7 所示。从命中率对比来看，采用兴趣森林模型可以有效地提高网页预取的准确性。

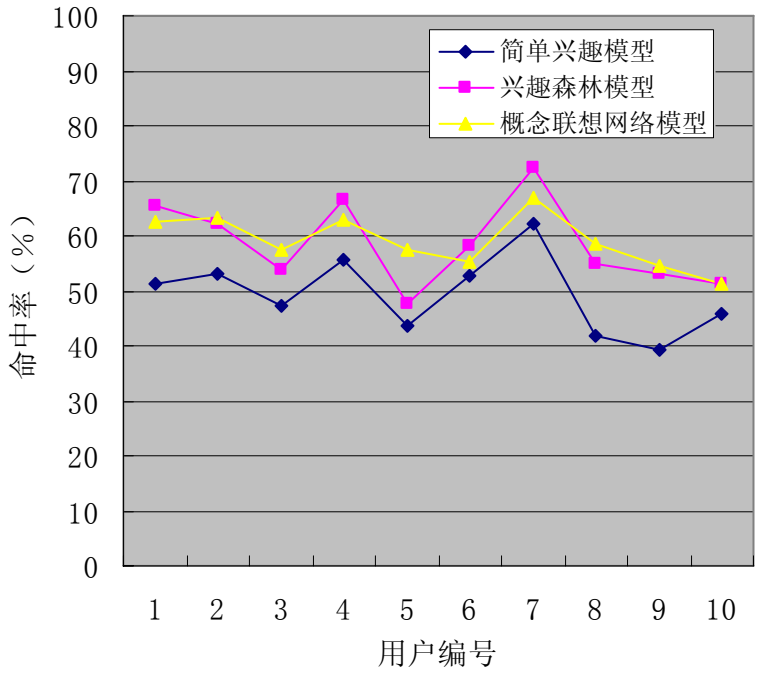

图 3.7　简单兴趣模型、概念联想网络模型与兴趣森林模型预取命中率的比较

Fig.3.7 the hit-ratio comparison of three models: the simple interest model, the concept association network model and the interest forest model

3.5 小结

随着 Internet 上信息的日益丰富，用户访问网络时的延迟也随之增加。预取技术通过预取用户将来可能访问的 web 页，来减少网络延迟，改善系统性能。从能够反映用户兴趣和行为的信息中归纳出可计算的用户模型的过程，即用户兴趣建模，是网页预取技术的核心和关键。用户模型的质量直接关系到个性化服务的质量，只有在高质量的用户兴趣建模的基础上，只有当用户的兴趣、偏好和访问模式等用户信息可以很好地被系统“理解”的时候，才可能实现理想的个性化服务，才能实现个性化服务系统所追求的各种目标。本章分析了现存用户兴趣模型的优缺点，利用“知网”建立了基于概念关系的用户兴趣森林模型，给出了用户兴趣森林的具体构建和更新过程。在构建的用户兴趣森林的基础上，计算出链接描述文字的平均带权语义距离，最后得到每个超链接描述文字的评价函数，为网页的预取提供了量化指标。实验结果表明，该方法能够提高预取的命中率，降低漏取率，具有较高的系统性能。

第四章 基于泊松分布的用户兴趣建模

Web推荐已经成为许多web站点的不可分割的部分。大、中、小公司都通过使用web推荐增加他们网站的实用性和客户满意度。现在，人们考虑了各种各样的信息（如：产品或者用户特征，访问历史，等等），已经开发了许多技术，使用不同的统计方法或者数据挖掘方法，来解决产生推荐的任务[101]。

[102]中给出了他们对 web 推荐算法研究的详细描述。根据分类，他们提出的系统属于“混合”推荐系统。[103]提出了基于给定 web 页的链接度来选择推荐内容的算法。[104]是基于用户的反馈进行推荐。其中的一个作者已经开始研究另一种面向反馈的动态产生 web 页的推荐方法。

Web 推荐系统就是基于数据挖掘和机器学习的方法，对用户可能感兴趣的网页进行推荐，它根据用户的爱好兴趣对用户可能访问的网页进行预测，并提供给用户进行选择。

许多研究者已经发现马尔可夫模型以及其变型，或者基于序列模式挖掘的模型很适合用来预测 web 用户的下一个请求。但是，较高阶的马尔可夫模型由于大量状态的存在而变得非常复杂，而低阶的马尔可夫模型不能捕捉到一个会话中用户的整个行为。基于序列模式挖掘的模型仅仅考虑了数据集中的频繁序列，很难跟随不在序列模式中的网页对用户的请求进行预测。再者，很难找到能够在一个会话中，挖掘两种不同类型信息的模型。

用户在网页上的停留时间是表现用户兴趣的关键要素。用户在网页上的停留时间越长，说明用户对该网页越感兴趣，相反，若用户在网页上仅仅是短暂停留，表明，用户对此网页的感兴趣程度很低，或者，用户对该网页根本不感兴趣。本章将用户在网页上的停留时间这一要素考虑在内，提出了基于泊松分布的用户兴趣建模方法。

4.1 归一化网页访问时间的计算

为了从用户浏览过的网页中理解用户的兴趣，需要从服务器中提取下列信息。

（1） 哪个用户访问了网站？识别用户的目的是获取他的访问路径。

（2） 用户浏览网页的路径。得到用户浏览的每个网页和访问网页的顺序，可以识别用户是如何访问这些网页的。

（3） 用户在每个网页上花费的时间。通过获取用户在网页上花费的时间，可以知道用户是否对其浏览的网页感兴趣。

（4） 用户从哪里离开的网站。用户离开网站前，浏览的最有一个网页可能是服务器会话结束的逻辑位置。

一旦识别了用户，就将每个用户的点击流，也就是用户的访问网页序列，分成一系列的会话。

定义 4.1 会话：一个会话就是从用户访问网站的时间开始到其离开网站为止，由用户执行的一组行为。

用户的会话行为可能中止于下述不同的原因：

（1） 用户已经达到了他访问网站的目的；

（2） 用户发现访问行为已经不再有意思；

（3） 会话时间超过了设定的时间限制。

访问日志中不存在用户登录、离开网站和使用网站的信息。并且，不是任何人任何时刻都能获取网站服务器中用户访问网站的登录文件内容。所有的这些都给确定用户的会话带来困难，导致很难清楚地确定一个会话的开始和结束。再者，来自其他服务器上的网页请求也不是总可以利用的，一个用户可能多次访问一个网站，web 服务器登录文件可能为每个用户记录了多个会话。会话识别的目标就是将每个用户对网页的访问分成独立的会话，通常采用超时方法识别用户会话。超时阈值的设定有两种方法（Δt 是固定的时间限制）。

一种方法是设定相邻请求之间的超时时间：如果两个网页间的请求时间的差值超过一定的界限，就认为用户开始了一个新的会话，也就是说，如果

$$TIME_i - TIME_{i-1} \le \Delta t, i \in \left[1, \cdots, |L|\right],$$

则认为开始了一个新的会话。

另一种方法是设定整个用户会话的超时时间，也就是说，如果

$$TIME_i - TIME_{|L|} \le \Delta t, i \in \left[1, \cdots, |L|\right]$$

则认为开始了一个新的会话。

本文中，采用后一种超时阈值的设定方法。1994 年，Catledge和Pitkow通过试验发现时间间隔设定为 25.5 分钟比较合适[100]。本文的实验中，设定 $\Delta t = 30$分钟 。当出现一个新的IP地址，或者同一个IP地址的网页访问时间超过了 30 分钟，就认为开始了

一个新的会话。

将识别出来的用户会话号 SID_i 加入到事务 T_i 中，也就是说，$T_i=(UID_i,TIME_i,URL_i,STATUSCODE_i,BYTES_i,SID_i)$，其中，$SID_i$ 是当同一用户每次进行新的会话时，创建的会话识别号。表 4.1 显示了从服务器登录文件中提取的，经过用户识别和会话识别之后形成的带有 UID 和 SID 的用户事务格式。

表 4.1 用户识别和会话识别之后事务的格式

Tab.4.1 the form after identifying the user and the session

用户 ID	请求日期及时间	方法，URL	状态码	传送的字节数	会话 ID
1	[23/Oct/2004:10:16:23-0500]	“GET”	200	3897	1
1	[23/Oct/2004:10:16:45-0500]	“GET A.gif”	200	5639	1
1	[23/Oct/2004:10:17:36-0500]	“GET C.html”	200	31786	1
2	[23/Oct/2004:10:17:11-0500]	“GET B.html”	200	29816	1
2	[23/Oct/2004:10:17:57-0500]	“GET A.gif”	200	5639	1

定义 4.2 时间间隔：用户两个行为之间的时间差值。

定义 4.3 网页访问时间：在一个用户的同一会话中，服务器日志中的两个连续的网页请求之间的时间间隔，就是用户浏览网页的时间。也就是说，数据库中相邻两条记录在时间字段的差值就是用户浏览网页的时间，对于最后一个网页的浏览时间，本文取该用户在当前会话中浏览过的所有网页时间的均值。

上述定义受限于大量的噪声数据，比如，不能准确的确定用户的行为。用户打开网页之后，可能在仔细地浏览网页，也可能去喝咖啡，打电话而根本没有浏览网页内容。本文假定，用户浏览网页的时候，类似上述的间断情况不会经常发生。所以，本文对这些情况不进行处理。

即便做了上述假定，仍存在另一个问题，就是，在服务器登录文件中记录的网页的浏览时间经常比客户端用户的实际浏览时间长。如图 4.1 所示，网页 A 和网页 B 间的请求时间是 t_6-t_1，但是网页 A 的实际浏览时间仅仅是 t_5-t_4。通常来说，一个网页由多个文件组成，如：页面，图片和脚本。用户浏览了一个网页，但是他不明确地将网页或者图片载入到他自己的浏览器中。由于客户端的连接速度和网页文件的大小不同，登录文件中记录的浏览时间和用户的实际浏览时间之间存在一定的差值，从一秒钟到几分钟不等。为了削减上述影响，本文应用启发式方法计算网页的访问时间。按照下面所述算法 4.1 计算一个 HTML 页的网页访问时间，即：从用户上一次发送完 HTML 请求，服务器发送了最后一个非 HTML 页到用户请求下一个 HTML 页之间的时间差值。下面介绍计算网页访问时间的算法 4.1。

假定用户在时间t_0请求了第一个 HTML 网页，上一个用户会话的结束时间和新用户会话开始的时间设定为t_0。在请求完网页 A 之后的最后一个非 HTML 页的时间是t_4，用户在时间t_6请求了 HTML 网页 B，那么网页 A 的访问时间为$t_6 - t_4$。对于用户会话的最后一个网页请求，本文设定其访问时间为所有会话中的非最后一个网页请求的所有网页访问时间的平均值。

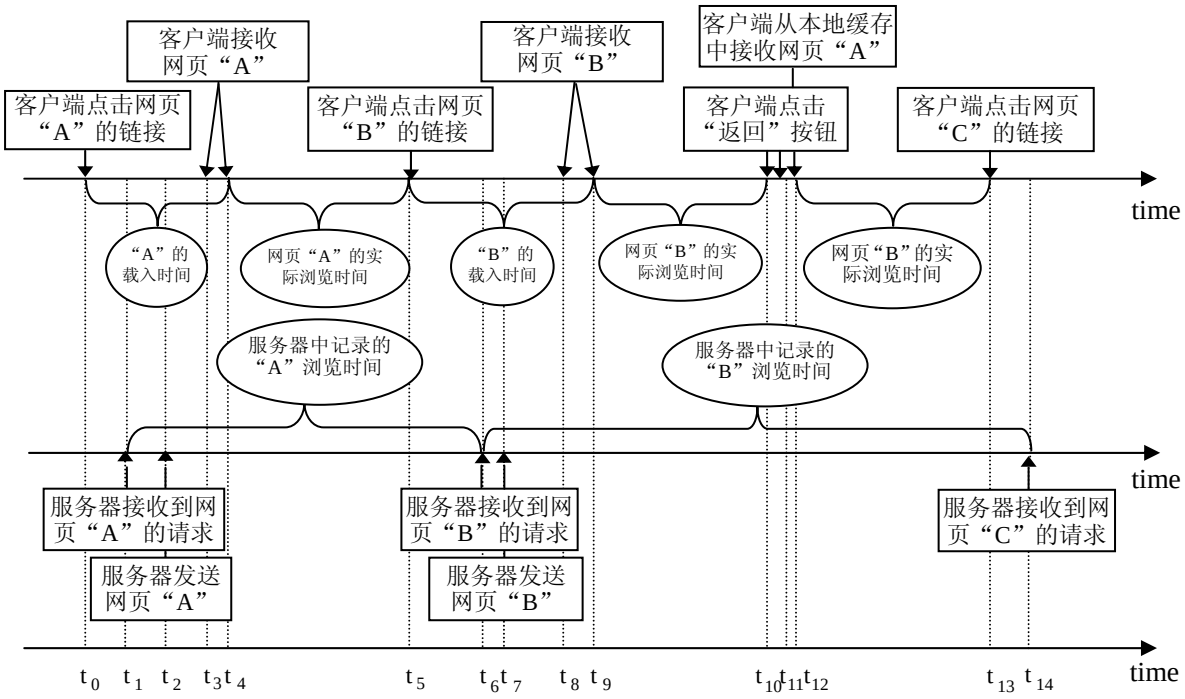

图 4.1 网页请求及响应的时间序列

Fig.4.1 Timeline for the webpage request and response

算法 4.1 计算网页的访问时间

输入：用户事务T， Δt

输出：网页的访问时间

Begin

{

按照UID和UID对T排序；

for (i=1;i<|L|;i++)

对于所有的UID_i和SID_i对

{

OPEN SESSION={Φ}；

for (j=1； j<|L|； j++)

对于包含UID_i和SID_i的所有T_j

{

if（URL_j 是 HTML 页）

then

{

if（存在 S_k 属于 *OPEN SESSION*）

then

if（$TIME_j$-*END TIME*(S_k)<= Δt ）

then

URL_j 的网页访问时间= $TIME_j$-*END TIME*(S_k)

else

{

关闭 S_k；

打开一个新的会话 S_k；

END TIME(S_k)=$TIME_j$;

}

else

if（存在 S_k 属于 *OPEN SESSION*）

then

if（$TIME_j$-*END TIME*(S_k)<= Δt ）

then

URL_i的网页访问时间= $TIME_j$- *END TIME*(S_k)

else

{

打开一个新的会话 S_k；

END TIME(S_k)=$TIME_j$;

}

}

else

if（存在 S_k 属于 *OPEN SESSION*）

then

if（$TIME_j$-*END TIME*(S_k)<= Δt ）

then

END TIME(S_k)=*TIME*$_j$;

else

{

关闭 S_k；

打开一个新的会话 S_k；

END TIME(S_k)=*TIME*$_j$;

}

else

{

打开一个新的会话 S_k；

END TIME(S_k)=*TIME*$_j$;

}

}

}

}

由于网页结构、长度、网页类型、用户对特定主题的兴趣度、网络连接速度等大量因素都可能影响到用户在网页上的花费时间，所以，仅仅使用用户在网页上的花费时间来衡量用户对网页的感兴趣程度是不准确的。本文通过对时间进行归一化来消减这些因素对时间的影响。由于想要捕捉到某一特定用户在一个会话周期内访问的某一网页比其访问的其他网页的相对重要性，本文通过同一会话S_k中的网页访问次数归一化网页访问时间。如式 4.1 示。

$$norm_{URL_i} = [(TIME_i - \min(T(S_k)))/(\max(T(S_k)) - \min(T(S_k)))] * (\max(norm) - \min(norm)) + \min(norm) \quad (4.1)$$

其中，$\max(T(S_k))$和$\min(T(S_k))$是用户在同一会话S_k中单一网页访问时间的最大值和最小值。$\max(norm)$和$\min(norm)$分别是归一化网页访问时间的最大值和最小值。本文，设定$\max(norm)$的值为 10，$\min(norm)$的值为 1。

经过处理之后，事务T_i的形式化描述变为：

$T_i = (UID_i, norm_{URL_i}, URL_i, STATUSCODE_i, BYTES_i, SID_i)$。

定义 4.4 用户在网页上的相对停留时间：用户在网页上的停留时间与此网页长度

的比值称为用户在网页上的相对停留时间，即：

$$\text{用户在网页上的相对停留时间}=\frac{\text{用户在网页上的停留时间}}{\text{网页长度}}。$$

构建基于泊松分布的用户兴趣模型的根据是：一个用户在一个页面上停留的时间表示他对该页面的感兴趣程度。一个 session 中，用户在一系列页面上花费的时间分配比率形成了该用户在这个 session 中的兴趣集合。首先将用户的所有 session 进行分类，使得具有相似兴趣集合的 session 聚集到同一个类中。该工作最关键的理念就是用户的 session 可以根据用户 session 中在一个公共网页上花费的时间进行聚类。具体的方式是通过下面的模型进行的：

(1) 当一个用户到达一个 web 站点时，他当前的 session 被分配到其中的一个聚类中。

(2) 根据访问时间，该用户在这个 session 中的浏览行为通过那个类的访问时间的泊松分布模型产生。

由于不知道实际的用户session应该被聚到哪一个类中，因此，本文使用一个标准的学习算法——EM算法[105]学习聚类分配以及每一个泊松分布的参数。得到的聚类集合中包括那些具有类似兴趣的session，而且，每一个聚类都有自己的表示这些兴趣的参数。根据用户被分配的聚类以及该聚类的参数，可以预测该用户后面可能访问的页面。这个模型根据预测产生一系列的推荐。接下来将给出详细的模型。

4.2 基于泊松分布的用户兴趣模型

基于模型的聚类方法使得给定的数据和一些数学方法能够得到较好的符合。这些方法通常是根据一个假设的概率分布得到的[106]。给定一个被考查的数据集合 $D=\{x_1,\cdots,x_M\}$，其中每一个被考查的数据 x_i 是由一系列参数定义的概率分布生成的，记这些参数为 Φ。这种概率分布由 $c_i\in C=\{c_1,\cdots,c_H\}$ 的混合模型组件组成。每一个组件的参数 Φ_h 是 Φ 不相交的子集，其中 $\Phi_h\left(h\in[1,\cdots,H]\right)$ 是一个向量，表示第 h 个组件的概率分布函数。

一个被考查数据 x_i 是按照如下过程生成的：首先，根据混合权重或者聚类先验概率分布 $p(c_h\mid\Phi)=\xi_h$，其中 $\sum_{h=1}^{H}\xi_h=1$，选择一个混合组件；然后，让这个被选中的混

合组件根据它自己的参数按照 $p\left(x_i \left| c_h;\Phi_h \right.\right)$ 分布生成该考察对象。这样，一个数据 x_i 的相似度就可以被特征化为所有混合组件的全部可能的总和：

$$p\left(x_i \mid \Phi\right)=\sum_{h=1}^{H} p\left(c_h \mid \Phi\right) p\left(x_i \mid c_h,\Phi_h\right)=\sum_{h=1}^{H} \xi_h p\left(x_i \mid c_h,\Phi_h\right) \tag{4.2}$$

统计学称该模型为有 H 组件的混合模型。这样，基于模型的聚类问题包括发现模型，也就是说，寻找符合数据的模型结构以及该结构的参数。可以通过最大似然估计方法和最大后验概率方法选择这些参数。其中，最大似然估计方法最大化：

$$\ell_{ML}\left(\Phi_1,\cdots,\Phi_H;\xi_1,\cdots,\xi_H \left| D\right.\right)=\prod_{i=1}^{M}\sum_{h=1}^{H} \xi_h p\left(x_i \mid c_h,\Phi_h\right) \tag{4.3}$$

而最大后验概率方法最大化 Φ 的后验概率：

$$\ell_{MAP}\left(\Phi_1,\cdots,\Phi_H;\xi_1,\cdots,\xi_H \left| D\right.\right)=\prod_{i=1}^{M}\sum_{h=1}^{H} \frac{\xi_h p\left(x_i \left| c_h,\Phi_h\right.\right) p\left(\Phi\right)}{p\left(D\right)} \tag{4.4}$$

其中，由于 $p\left(D\right)$ 不是 Φ 的函数，故在式 4.4 的计算中忽略它。

实际上，在实际的应用过程中，使用上式的对数(log)式，式 4.3 及式 4.4 的对数表示如下：

$$L\left(\Phi_1,\cdots,\Phi_H;\xi_1,\cdots,\xi_H \left| D\right.\right)=\sum_{i=1}^{M}\ln\left(\sum_{h=1}^{H} \xi_h p\left(x_i \left| c_h,\Phi_h\right.\right)\right) \tag{4.5}$$

$$L\left(\Phi_1,\cdots,\Phi_H;\xi_1,\cdots,\xi_H \left| D\right.\right)=\sum_{i=1}^{M}\ln\left(\sum_{h=1}^{H} \xi_h p\left(x_i \left| c_h,\Phi_h\right.\right)\right)+\ln p\left(\Phi\right) \tag{4.6}$$

模型的参数集合 Φ 包括表示聚类先验概率分布(ξ_h)的表明选择不同混合组件的概率混合权重，和为数据假定的概率分布的参数集合：

$$\Phi=\left\{\Phi_1,\cdots,\Phi_H,\xi_1,\cdots,\xi_H\right\},\sum_{h=1}^{H}\xi_h=1 \tag{4.7}$$

使用期望最大化(EM)算法训练模型的参数。EM 算法是一个对参数进行估计的通用的迭代算法，当部分随机变量缺失或者不完整时，通过最大似然对他们进行参数估计。在期望步(E-step)中，根据可见变量值和现有的参数值，通过计算缺失变量的可能性，计算这些变量被填充的可能性。在最大化步(M-step)中，根据填充的变量调整参数。

假定 $D=\left\{x_1,\cdots,x_M\right\}$ 是包括 M 个可变量的集合，$G=\left\{y_1,\cdots,y_M\right\}$ 代表隐藏变量 Y

的M个值的集合，使得每一个y_i可以表示为$y_i=\{y_{1_i},\cdots,y_{H_i}\}$的形式，并对应一个数据点$x_i$。假定$Y$是离散的，并且用下面的可能的值来表示分类标签：

$$y_{ji}=\begin{cases}1 & \text{如果}x_i\text{在聚类}j\text{中}\\ 0 & x_i\text{不在聚类}j\text{中}\end{cases}$$

如果Y是可见的，那么，ML 评估问题可以在最大化下面式子的基础上得到解决：

$$L_c(\Phi;D,G)\triangleq\ln p(D,G|\Phi) \tag{4.8}$$

为了表示缺失的数据，在可见数据和当前参数评估(如式 4.9 示)的条件下，计算全部数据可能性的条件期望：

$$S(\Phi,\Phi')=E\left[L_c(D,G|\Phi)\middle|D,\Phi'\right] \tag{4.9}$$

其中，$L_c(D,H|\Phi)$定义如下：

$$L_c(D,G|\Phi)=\sum_{i=1}^{M}\ln p(x_i,y_i|\Phi) \tag{4.10}$$

将等式 4.9 中的S函数扩展如下：

$$\begin{aligned}S(\Phi,\Phi')&=E\left[\sum_{i=1}^{M}\ln p(x_i,y_i|\Phi)\middle|D,\Phi'\right]\\&=\sum_{k=1}^{H}\sum_{i=1}^{M}\ln p(x_i,y_i|\Phi)\prod_{j=1}^{M}p(y_{kj}|x_j,\Phi')\\&=\sum_{i=1}^{M}\sum_{k=1}^{H}\left(\ln p(x_i,y_i|\Phi)\,p(y_{ki}|x_i,\Phi')\right)\prod_{j\neq i}\sum_{k=1}^{H}p(y_{kj}|x_j,\Phi')\\&=\sum_{i=1}^{M}\sum_{k=1}^{H}\ln p(x_i,y_i|\Phi)\,p(y_{ki}|x_i,\Phi')\\&=\sum_{i=1}^{M}\sum_{y_i}\ln p(x_i,y_i|\Phi)\,p(y_i|x_i,\Phi')\\&=\sum_{i=1}^{M}\sum_{y_i}p(y_i|x_i,\Phi')\ln\left[p(x_i|y_i,\Phi)\,p(y_i|\Phi)\right]\\&=\sum_{i=1}^{M}\sum_{y_i}p(y_i|x_i,\Phi')\ln\left[p(x_i|y_i,\Phi)+\ln p(y_i|\Phi)\right]\end{aligned} \tag{4.11}$$

在每一次 EM 迭代过程中，使用当前的参数Φ'，最大化包含参数集合Φ的S函数。在每一次递归结束时，新的优化参数集合Φ变成了下一次递归的参数Φ'。使用这个

过程，EM 算法可以通过下面的过程来实现。

（1） 为参数集合$\Phi^{'}$选择一个初始的估计，设定$n=0$；

（2） E 步((E)xpectation Step)：对n，使用公式 4.9 计算$S\left(\Phi,\Phi^{'}\left(n\right)\right)$；

（3） M 步((M)aximization Step)：利用新的估计$\Phi^{'}(n+1)$替换当前的估计$\Phi^{'}(n)$，其中，$\Phi^{'}(n+1)=\arg\max_{\Phi} S(\Phi,\Phi^{'}(n))$；

（4） 设定$n=n+1$，重复步 2 和步 3，直到收敛

通过重复地应用 E 步和 M 步，参数集合Φ将至少收敛到 log 可能性函数的局部最大值。

4.3 基于泊松分布的用户兴趣模型的构建

根据前面章节所述，将 web 服务器的日志转化为一系列的用户会话的集合。根据在公共网页上的花费时间的相似度对用户会话进行聚类。

例 4.1 表 4.2 中给出了一个包含十个网页$P=\left\{p_1,p_2,\cdots,p_{10}\right\}$的 web 站点的用户会话的样例集，其中网页集*PAGES*对应*P*中的页的子集，归一化时间*NORMS*对应着*P*中网页的归一化访问时间。

表 4.2 一个用户会话的例子

Tab.4.2 an instance of a user session

会话号	网页集	归一化时间
1	{p_3, p_2, p_1, p_{10}, p_6, p_5}	{0,0,0,1,0,0,8,0,7,9}
2	{p_7, p_9, p_2, p_1, p_6, p_5, p_8}	{9,9,2,0,10,5,0,0,10,3}
3	{p_7, p_6, p_4, p_8, p_1, p_9, p_2}	{10,10,6,0,2,9,0,10,0,10}
4	{ p_5, p_6, p_2, p_9, p_8, p_7}	{9,8,1,0,0,5,0,0,10,3}
5	{p_7, p_1, p_2, p_9}	{10,0,7,0,3,9,0,0,0,0}
6	{ p_7, p_9, p_2, p_1, p_6 }	{10,0,7,0,2,10,0,0,0,10}
7	{p_{10}, p_5, p_2, p_6, p_3}	{0,0,0,1,0,10,9,0,7,9}

4.3.1 聚类 Web 日志数据中的用户 Session

本节，首先描述用来对Web日志数据中的用户Session进行聚类的特定混合模型。然后，给出使用EM算法训练泊松分布混合模型时的更新参数。根据用户在每一个session中的兴趣，使用基于模型的技术来对用户的session进行分组。假定数据是通过以下的方式生成的[108]：

（1） 当一个用户到达一个 web 站点时，他的会话以一定的概率分配到*H*类中的

某个类中。

（2） 如果一个用户的 session 属于某一个类，他接下来的请求根据该类中的特定可能性分布产生。

由于假定数据是通过一个混合模型来产生的，所以，每一个用户的 session 都是通过模型参数子集Φ_h定义的概率分布来生成的。假定$X=\{x_1,x_2,\cdots,x_M\}$是M个用户的会话集，C是离散有值的变量集合，变量的值分别为：$c_1,c_2,\cdots,c_H$，对应于用户会话的位置聚类分配。在这种情况下，一个用户 session 的混合模型是：

$$p(X=x_i|\Phi)=\sum_{h=1}^{H}p(C=c_h|\Phi)p(X=x_i|C=c_h,\Phi_h)$$
$$=\sum_{h=1}^{H}\xi_h p(X=x_i|c_h,\Phi_h) \quad (4.12)$$

其中，ξ_h是选择聚类c_h的概率。一个用户 session，不妨假设为x_i，被看作是一个访问页面时间的n维向量$(x_{i1},x_{i2},\cdots,x_{in})$，其中，$x_{ij}$对应于用户 session 中*NORM*字段中的$norm_{pi}$，每个p_j是页面集合$P=\{p_1,p_2,\cdots,p_n\}$中的一个。P中的每一个页面对应模型中的一个维。n维向量表示了用户的聚合兴趣。

在本章的例子中，混合模型被认为是类标号缺失的一个分布。虽然使用频繁模式挖掘可以减少输入数据的维数，但是，如何估计概率仍然是一个问题。处理这个问题的一个关键思想是为底层分布设定一个结构，例如，可以假设维度的独立性：

$$p(x_i)=\prod_{j=1}^{n}p_j(x_{ij}) \quad (4.13)$$

因为用户 session 是归一化访问时间的一个n维向量，所以可以很容易的将这种假设运用到本章提出的模型中。尽管两个用户 session 中访问页面的顺序可能不同，但是，如果在每一个会话中对应于同一网页的归一化网页访问时间相同的话，这两个 session 可以被表示成为一个相同的向量。

例 4.2　为了解释本章构建的模型中的独立性假设，考虑表 4.2 中的第 2、5、6 这三个用户会话，在会话 5 和会话 6 中，页面请求的顺序是不同的，然而，他们的聚合兴趣却非常类似，因为每个页面上花费的归一化访问时间是相似的。虽然在第 2 个会话和第 6 个会话中，前 5 个页面访问的顺序是相同的，但是，他们的聚合兴趣却不同。根据本文提出的聚类原则，第 5 个会话和第 6 个会话将会被分到同一个聚类中，而第

2 个 session 将属于另一个不同的聚类。

这个独立性假设使得可以使用 n 个独立的概率分布来对用户session中的每一维进行建模。为了对这个数据建模，假定每一维上的数据是由混合的泊松分布生成。一个随机变量 x 有一个参数为 m 的泊松分布 $(m>0)$[107]：

$$p(X=k)=\frac{m^k e^{-m}}{k!} \qquad k=0,1,... \tag{4.14}$$

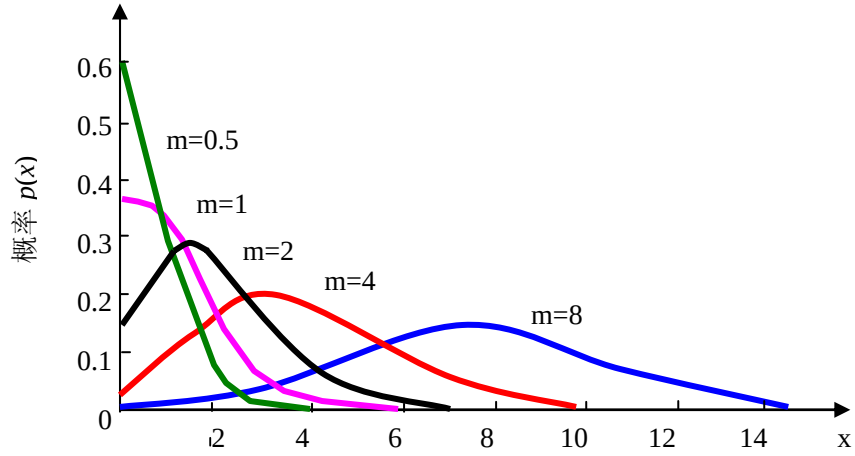

图 4.2 不同参数的泊松分布形状

Fig.4.2 Poisson distribution shapes with different parameters

上图表示了不同参数的泊松分布图的形状。可以看出，随着 m 的增大，泊松分布变得类似于钟形的分布。泊松分布可以被用来对单个事件发生的比率，如，对一个用户 session 在某一特定页面上的归一化时间为 1 这样的事件进行建模。为了验证本章提出的假设，也就是说，每一维上的数据是通过一个泊松分布来生成的，画出了每一维上十个可能值的每一个值出现的频率图，其中的大多数图形证实了本章提出的假设。图 4.3 表示了其中的一个柱状图。可以看出，该柱状图的形状类似于带有一个较小的参数值 m 的泊松分布图。

根据独立性假设，聚类 h 中的一个用户 session x_i 是按照公式 4.15 的泊松模型来生成的：

$$p(x_i|c_h,\Phi_h)=\prod_{j=1}^{n}\frac{(\varphi_{hj})^{x_{ij}}\ e^{-\varphi_{hj}}}{x_{ij}!} \tag{4.15}$$

其中，φ_{hj} 是聚类 h 中的第 j 维上的泊松分布的参数。

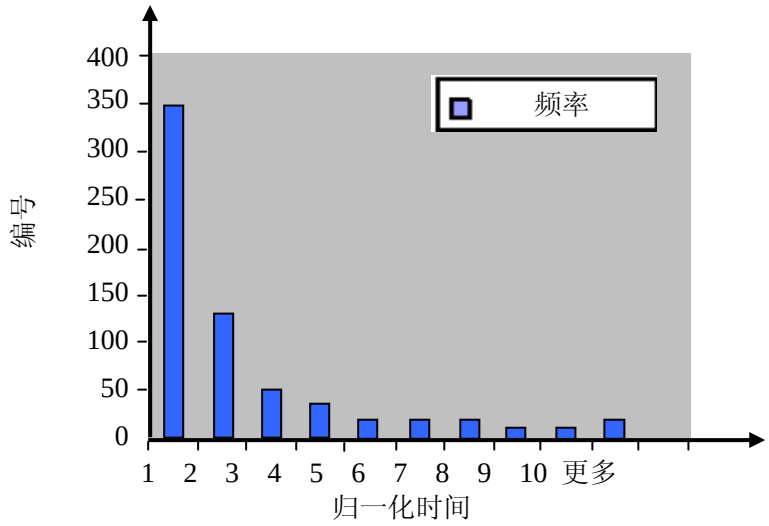

图 4.3 NASA web 服务器的网页柱状图

Fig.4.3 Histogram of NASA Web server

通过结合等式 4.11 和等式 4.14，得到：

$$p\left(x_i\middle|\Phi\right)=\sum_{h=1}^{H}\xi_h\left(\prod_{j=1}^{n}\frac{\left(\varphi_{hj}\right)^{x_{ij}}e^{-\varphi_{hj}}}{x_{ij}!}\right) \tag{4.16}$$

其中，$\varphi_{hj}\left(h\in\left[1,\cdots,H\right],j\in\left[1,\cdots,n\right]\right)$是聚类$c_h$在维度 j 上的泊松分布参数。

例 4.3　对于表 4.2 中的用户 session 的样本集，每一个聚类有 10 个泊松参数，其中，数据集中的唯一性页面的个数为 10。模型参数为：

$$\Phi=\left\{\Phi_1,\cdots,\Phi_H,\xi_1,\cdots,\xi_H\right\},\Phi_h=\left(\varphi_{h1},\cdots,\varphi_{hn}\right),\sum_{h=1}^{H}\xi_h=1 \tag{4.17}$$

4.3.2 训练模型参数

使用 EM 算法，在 EM 算法的最大化步骤中增强条件依赖的假设来训练在前面章节中提出的混合模型的模型参数，为模型中的每一个组件建立一个学习算法。使用 EM 算法主要有下面几个原因：

（1） 使用泊松分布表示一个 session 中的用户行为

（2） 算法的性能和 session 的个数是线性关系

（3） 对噪声数据来说，鲁棒性较好

（4） 接受期望的聚类数作为输入

（5） 为每一个 session 提供成为类成员的概率

（6） 可以处理高维数据

（7） 给定一个好的初始条件，能够快速收敛

为了实现 EM 算法，首先选取要聚类的个数(H)，初始化起点$\left(\Phi^{'}(0)\right)$及为了对模型参数进行最大后验概率估计需要的Φ 的收敛条件和先验概率。为了确定聚类数，使用几个聚类数运行算法。首先评估单个组件模型的泊松参数，然后再随机地打乱部分参数，得到参数集合H，来初始化组件的参数，$\Phi_h\left(h\in[1,\cdots,H]\right)$。如果在训练数据上的两次连续迭代的 log 概率差别值小于 0.001%，认定条件收敛。这样一来，在参数评估的准确度和迭代次数之间就需要一种平衡。迭代的差别值越小，迭代的次数就越多，算法收敛的时间就越长。反之，参数的精确度就低。最后，为了进行 MAP 评估给Φ 分配先验概率，本章使用了泊松分布的先验分布。

例 4.4 为表 4.2 中的数据集确定初始化参数。假定聚类的数量为 3，聚类的先验可能性设定为$\xi_h=1/3$，其中$h\in[1,2,3]$。为每一个聚类的泊松分布参数确定 10 个初始值，总共是 30 个泊松参数。表 4.3 表示了这些参数。然后，对于 EM 算法的第一次迭代，第一个聚类的先验聚类值是$\xi_1=1/3$，该类中的第 2 维的泊松参数为$\varphi_{11}=0.401$。

表 4.3 聚类数为 3 时的泊松参数

Tab.4.3 the Poisson parameters when clustering number is 3

网页号	聚类 1	聚类 2	聚类 3
1	0.432	0.401	0.418
2	0.140	0.155	0.162
3	0.095	0.092	0.111
4	1.002	1.007	0.994
5	0.213	0.238	0.210
6	0.192	0.203	0.205
7	0.220	0.221	0.200
8	0.146	0.138	0.148
9	0.171	0.169	0.173
10	0.486	0.482	0.497

4.3.2.1 模型参数的 ML 估计（最大似然估计）

从数据中训练参数的一个过程就是发现那些最大化数据似然性的参数值：

$$\Phi^{ML}=\arg\max_{\Phi}\left\{p\left(D\middle|\Phi_1,\cdots,\Phi_H,\xi_1,\cdots,\xi_H\right)\right\} \tag{4.18}$$

这些参数经常被用为最大似然性或者 ML 估计。

在对参数进行 ML 评估的 E-步中包含了在给定的当前参数集合$\Phi^{'}$下，对缺失类标号的条件概率的更新问题。

为了计算用来获取 E-step 中 ML 参数所需要的等式，在当前给定的参数集合$\Phi^{'}$下，计算缺失的类标号的条件概率，定义这个概率为聚类后验概率$P_{ih}\left(\Phi^{'}\right)$，其中，会话$x_i$来源于第$h$个聚类。使用 Bayes 规则得到聚类后验概率为：

$$\begin{aligned} P_{ih}\left(\Phi^{'}\right) &= p\left(C=c_h \middle| x_i\right) \\ &= \frac{p\left(C=c_h\right)p\left(x_i \middle| c_h, \Phi^{'}_h\right)}{p\left(x_i\right)} \\ &= \frac{\xi_h p\left(x_i \middle| c_h, \Phi^{'}_h\right)}{\sum_{j=1}^{H}\xi_j p\left(x_i \middle| c_j, \Phi^{'}_j\right)} \end{aligned} \tag{4.19}$$

这样，S 函数就可以表示为：

$$S\left(\Phi,\Phi^{'}\right)=\sum_{i=1}^{M}\sum_{h=1}^{H}P_{ih}\left(\Phi^{'}\right)\left[\ln p\left(X_i \middle| c_h,\Phi_h\right)+\ln\xi_h\right] \tag{4.20}$$

在 M-step 中，为了使训练数据的预期可能性最大，实验中保持聚类后验概率不变，但重新给定一组参数集合$\Phi^{'}\left(n+1\right)$。在聚类的先验概率和为 1 的前提下，最大化S函数。为了实现条件的最大化，实验中使用了 Lagrange 乘法因子。聚类先验概率的评估等式如下：

$$\begin{aligned} &\frac{\partial}{\partial\xi_h}\left[S\left(\Phi,\Phi^{'}\right)-\lambda\sum_{j=1}^{H}\xi_j\right]=0 \\ &\sum_{i=1}^{M}P_{ih}\left(\Phi^{'}\right)\left[\frac{1}{\xi_h}\right]-\lambda=0 \end{aligned} \tag{4.21}$$

其中，

$$\lambda\xi_h=\sum_{i=1}^{M}P_{ih}\left(\Phi^{'}\right) \tag{4.22}$$

对 4.22 中的h求和，得到：

$$\lambda=\sum_{i=1}^{M}\sum_{h=1}^{H}P_{ih}\left(\Phi^{'}\right)=M \tag{4.23}$$

其中，等式 4.23 满足这样一个事实：

$\sum_{h=1}^{H}P_{ih}\left(\Phi^{'}\right)=1$。

将等式 4.22 和 4.23 结合，得到更新聚类概率的等式：

$$\hat{\xi}_h = \frac{1}{M}\sum_{i=1}^{M} P_{ih}\left(\Phi^{'}\right) \tag{4.24}$$

类似地，在独立性假设（4.25）的前提下，对于泊松模型的参数Φ_h，可以最大化等式 4.21 中的S 函数。

$$\begin{aligned}
&\frac{\partial}{\partial\varphi_{hm}}\left[S\left(\Phi,\Phi^{'}\right)\right]=0\\
&\frac{\partial}{\partial\varphi_{hm}}\left[\sum_{i=1}^{M} P_{ih}\left(\Phi^{'}_{h}\right)\left(\ln\prod_{j=1}^{n}\frac{\left(\varphi_{hj}\right)^{x_{ij}}e^{-\varphi_{hj}}}{x_{ij}!}+\ln\xi_h\right)\right]=0\\
&\sum_{i=1}^{M} P_{ih}\left(\Phi^{'}\right)\left[\frac{x_{im}}{\varphi_{hm}}-1\right]=0
\end{aligned} \tag{4.25}$$

从而得到下面的更新泊松模型参数的等式：

$$\hat{\varphi}_{hm}=\frac{\sum_{i=1}^{M}\left(P_{ih}\left(\Phi^{'}\right)x_{im}\right)}{\sum_{i=1}^{M}P_{ih}\left(\Phi^{'}\right)} \tag{4.26}$$

4.3.2.2 模型参数的 MAP 评估（最大后验概率评估）

使用最大似然估计方法的一个困难是零可能性的出现。例如，对于数据集中的一个页面p_i，如果没有对该页面的请求，那么，对该页面的泊松参数的评估值就为 0。也就是说，根据本章提出的模型，对于该页请求的可能性是 0。为了解决这一个问题，事先为Φ 分配一个先验概率，并使用最大后验分布作为参数评估。这样，通过最大化给定数据Φ 的后验概率可以对应于Φ 的最大后验分布的 MAP 参数，也就是：

$$\Phi^{MAP}=\arg\max\Phi=\left\{p\left(D\middle|\Phi_1,\cdots,\Phi_H,\xi_1,\cdots,\xi_H\right)p\left(\Phi\right)\right\} \tag{4.27}$$

其中，第二个等号利用了 Bayes 规则，是模型参数的先验分布。

为了对参数进行MAP评估，首先需要为先验概率$p\left(\Phi\right)$选择一个函数形式。参数集Φ 包括泊松参数集和类权重。常用的泊松分布的先验分布是具有两个参数α和β的Gamma分布[109]。选择类的分布可以被认为是一个多项式分布，而多项式分布的共轭先验分布是具有参数γ的Dirichlet分布。

下面具体的来证明一下。

证明 4.1：如果样本符合泊松分布，则 Gamma 分布族是泊松分布的共轭先验分布。共轭先验分布是指，产生由先验概率产生的后验概率和先验概率在功能形式是一样的。对于数据集合 $D=\{X_1,\cdots,X_n\}$，假定 $X_1,\cdots,X_n$ 是在符合一个 φ 值未知的泊松分布中随机产生的样本，φ 的先验分布是参数为 $\alpha,\beta(\alpha>0,\beta>0)$ 的 Gamma 分布：

$$p(\varphi)=\frac{\beta^{\alpha}\varphi^{\alpha-1}e^{-\beta\varphi}}{\Gamma(\alpha)}\propto\varphi^{\alpha-1}e^{-\beta\varphi} \tag{4.28}$$

其中，对于正整数 α：

$$\Gamma(\alpha)=(\alpha-1)!$$

在这种情况下，对于给定的 $X_i=x_i(i=1,\cdots,n)$，φ 的后验分布就是一个参数为 $\alpha+\sum_{i=1}^{n}x_i$ 和 $\beta+n$ 的 Gamma 分布。

证明：设 $y=\sum_{i=1}^{n}x_i$，则似然函数 $p(D|\varphi)$ 满足关系 $p(D|\varphi)\propto e^{-n\varphi}\varphi^{y}$

如果 X 是一个符合泊松分布的随机变量，，则一个特定数据 x_i 的概率为：

$$p(X=x_i|\varphi)=\frac{\varphi^{x_i}e^{-\varphi}}{x_i!} \tag{4.29}$$

全部数据集 $D=\{X_1,\cdots,X_n\}$ 的概率作为似然性，定义为：

$$\begin{aligned}p(D|\varphi)&=\prod_{i=1}^{n}p(X=x_i|\varphi)\\&=\prod_{i=1}^{n}\frac{\varphi^{x_i}e^{-\varphi}}{x_i!}\propto\varphi^{\sum_i x_i e^{-n\varphi}}\end{aligned} \tag{4.30}$$

参数 φ 的后验分布为：

$$p(\varphi|D)=\frac{p(D|\varphi)p(\varphi)}{p(D)}\propto p(D|\varphi)p(\varphi) \tag{4.31}$$

将公式 4.28 和 4.30 代入 4.31，得到：

$$\begin{aligned}p(\varphi|D)&\propto\varphi^{\sum_i x_i}e^{-n\varphi}\varphi^{\alpha-1}e^{-\beta\varphi}\\&\propto\varphi^{\alpha+y-1}e^{-\varphi(n+\beta)}\end{aligned} \tag{4.32}$$

因此，φ 的后验分布再一次变为参数为 $\alpha+\sum_{i=1}^{n}x_i$ 和 $\beta+n$ 的 Gamma 分布。

证明 4.2：假定随机变量 $D=\{X_1,\cdots,X_k\}$ 来源于一个多项分布，则联合概率函数为：

$$p\left(X_1=x_1,\cdots,X_k=x_k\right)=\frac{N!}{x_1!\cdots x_k!}\xi_1^{x_1}\cdots\xi_k^{x_k}\propto\prod_{i=1}^{k}\xi_i^{x_i} \tag{4.33}$$

其中，

$$x_1+\cdots+x_k=N$$

$$\xi=\{\xi_1,\cdots,\xi_k\},\xi_1+\cdots+\xi_k=1$$

ξ 的共轭先验是参数为 $\gamma=\{\gamma_1,\cdots,\gamma_k\}(\gamma_j>0)$ 的 Dirichlet 分布，如公式 4.34 所示。

$$p(\xi)=\frac{\Gamma\left(\sum_{i=1}^{k}\gamma_i\right)}{\prod_{i=1}^{k}\Gamma(\gamma_i)}\prod_{j=1}^{k}\xi_j^{\gamma_j-1}\propto\prod_{j=1}^{k}\xi_j^{\gamma_j} \tag{4.34}$$

那么，ξ 的后验分布是另外一个其参数为 $\gamma_1+x_1,\cdots,\gamma_k+x_k$ 的 Dirichlet 分布。

证明：对于给定的 ξ 的 Dirichlet 先验分布，设待研究的数据集合 D 来自于多项分布的，状态 i 发生的概率为 $x_i(i=1,\cdots,k)$，那么似然函数满足关系 $p\left(D|\xi\right)\propto\prod_{i=1}^{k}\xi_i^{\gamma_i+x_i}$

观察 D 的概率为：

$$p\left(D|\xi\right)\propto\prod_{i=1}^{k}\xi_i^{x_i} \tag{4.35}$$

用 Bayes 规则将 ξ 的后验分布表示为：

$$p\left(\xi|D\right)=\frac{p\left(D|\xi\right)p(\xi)}{p(D)}\propto p\left(D|\xi\right)p(\xi) \tag{4.36}$$

将公式 4.34 和 4.35 代入公式 4.36，得到：

$$\begin{aligned}p\left(\xi|D\right)&\propto\prod_{i=1}^{k}\xi_i^{x_i}\prod_{j=1}^{k}\xi_j^{x_j}\\&\propto\prod_{i=1}^{k}\xi_i^{x_i+\gamma_i}\end{aligned} \tag{4.37}$$

该结果同样是一个参数为 $\gamma_1+x_1,\cdots,\gamma_k+x_k$ 的 Dirichlet 分布。

一般说来，通过对问题的知识的预先了解，确定共轭先验分布参数的选择。然而，实际中很难获取这些先验知识。如果没有这些先验知识，人们通常使用“无信息”的先验值，典型的是使用均匀分布。本文中，尝试了几种参数的组合。

MAP 的 E-step 中对参数的评估方法与 ML 中的评估方法相同，对给定的当前参数集$\Phi^{'}$，通过等式 4.19 来计算缺失标号的条件概率。EM 算法中，计算 log 后验函数的S函数定义如下：

$$S\left(\Phi,\Phi^{'}\right)=\sum_{i=1}^{M}\sum_{h=1}^{H}P_{ih}\left(\Phi^{'}\right)\left[\ln p\left(x_i\left|c_h,\Phi_h\right.\right)+\ln\xi_h\right]+\ln p\left(\Phi\right) \tag{4.38}$$

其中，参数Φ由所有的混合模型参数组成，如式 4.39 示：

$$\begin{aligned}&\Phi=\left\{\Phi_1,\cdots,\Phi_H,\xi\right\}\\&\Phi_h=\left\{\varphi_{h1},\cdots,\varphi_{hn}\right\}\\&\xi=\left\{\xi_1,\cdots,\xi_H\right\},\sum_{h=1}^{H}\xi_h=1\end{aligned} \tag{4.39}$$

等式 4.37 中的$p\left(\Phi\right)$包括了泊松分布的参数先验和聚类先验，可以被分解为：

$$p\left(\Phi\right)=\prod_{h=1}^{H}\prod_{k=1}^{n}p\left(\varphi_{hk}\left|\alpha_{hk},\beta_{hk}\right.\right)p\left(\xi\left|\gamma\right.\right) \tag{4.40}$$

使用参数为α和β的 Gamma 先验作为泊松参数，使用 Dirichlet 先验作为聚类权重：

$$\begin{aligned}p\left(\varphi_{hk}\left|\alpha_{hk},\beta_{hk}\right.\right)&\propto\varphi_{hk}^{\alpha_{hk}}e^{-\beta_{hk}\varphi_{hk}}\\p\left(\xi\left|\gamma\right.\right)&\propto\prod_{h=1}^{H}\xi_h^{\gamma_h}\end{aligned} \tag{4.40}$$

等式 4.38 中的S函数可以写为：

$$\begin{aligned}S\left(\Phi,\Phi^{'}\right)=&\sum_{i=1}^{M}\sum_{h=1}^{H}P_{ih}\left(\Phi^{'}\right)\left[\ln p\left(x_i\left|c_h,\Phi_h\right.\right)+\ln\xi_h\right]\\&+\sum_{h=1}^{H}\sum_{k=1}^{n}\left(\alpha_{hk}\ln\varphi_{hk}-\beta_{hk}\varphi_{hk}\right)+\sum_{h=1}^{H}\gamma_h\ln\xi_h\end{aligned} \tag{4.42}$$

为了计算最优参数，在聚类先验和为 1 的前提下，最大化等式 4.42 中的S函数：

$$\begin{aligned}&\frac{\partial}{\partial\xi_h}\left[S\left(\Phi,\Phi^{'}\right)-\lambda\sum_{j=1}^{H}\xi_j\right]=0\\&\sum_{i=1}^{M}P_{ih}\left(\Phi^{'}\right)\left[\frac{1}{\xi_h}\right]+\frac{\gamma_h}{\xi_h}-\lambda=0\end{aligned} \tag{4.43}$$

该等式满足条件：

$$\lambda\xi_h = \sum_{i=1}^{M} P_{ih} + \gamma_h \tag{4.44}$$

对等式 4.44 中的 h 求和，得到：

$$\lambda = \sum_{j=1}^{H}\left[\sum_{i=1}^{M} P_{ij} + \gamma_j\right] \tag{4.45}$$

将等式 4.45 代入到 4.44，并应用先验聚类，得到更新聚类先验概率的等式：

$$\hat{\xi}_h = \frac{\sum_{i=1}^{M} P_{ih}\left(\Phi^{'}\right) + \gamma_h}{\sum_{j=1}^{H}\left[\sum_{i=1}^{M} P_{ij}\left(\Phi^{'}\right) + \gamma_j\right]} \tag{4.46}$$

对于泊松参数，优化 S 函数：

$$\begin{aligned} &\frac{\partial}{\partial\varphi_{hm}}\left[S\left(\Phi,\Phi^{'}\right)\right] = 0 \\ &\sum_{i=1}^{M} P_{ih}\left(\Phi^{'}\right)\left[\frac{x_{im}}{\varphi_{hm}} - 1\right] + \frac{\alpha_{hm}}{\varphi_{hm}} - \beta_{gm} = 0 \end{aligned} \tag{4.47}$$

用等式 4.47 求解 φ_{hm}，从而获得更新泊松参数的等式：

$$\hat{\varphi}_{hm} = \frac{\sum_{i=1}^{M} P_{ih}\left(\Phi^{'}\right) x_{im} + \alpha_{hm}}{\sum_{i=1}^{M} P_{ih}\left(\Phi^{'}\right) + \beta_{hm}} \tag{4.48}$$

EM 算法 MAP 评估过程的输出是聚类参数集，每一个聚类都有自己的参数：

$$pc_h = \left\{\xi_h, \left(\varphi_{h1}, \cdots, \varphi_{hn}\right)\right\}$$

例 4.5　对于表 4.2 中的数据集合，在 E 步中使用等式 4.13 计算聚类的后验概率。在 M-step 中，使用公式 4.24 和 4.26 更新模型参数。这样，表 4.3 中的参数和聚类的先验值在每一个 M 步中被更新。重复 E 步和 M 步，直到满足了收敛规则。算法的输出是聚类的参数集合。例如：$\left\{0.3;\left(0.02,1.2,\cdots\right)\right\}$ 表明，聚类的先验可能性为 0.3，第一个网页的泊松参数是 0.02，第二个网页的泊松参数为 1.2，……，依此类推。当使用 MAP 评估时，根据共轭对的上级参数，具体的参数可能会有所不同。表 4.4 中的聚类是通过将每一个 session 分配给具有最高后验概率的聚类中获得的。

表 4.4 使用 EM 算法产生的聚类
Tab.4.4 the cluster result using EM algorithm

聚类号	会话号	网页集	归一化时间
1	3	{p_7, p_6, p_4, p_8, p_1, p_9, p_2}	{10,10,6,0,2,9,0,10,0,10}
	5	{p_7, p_1, p_2, p_9}	{10,0,7,0,3,9,0,0,0,0}
	6	{ p_7, p_9, p_2, p_1, p_6 }	{10,0,7,0,2,10,0,0,0,10}
2	1	{p_3, p_2, p_1, p_{10}, p_6, p_5}	{0,0,0,1,0,0,8,0,7,9}
	7	{p_{10}, p_5, p_2, p_6, p_3}	{0,0,0,1,0,10,9,0,7,9}
3	2	{p_7, p_9, p_2, p_1, p_6, p_5, p_8}	{9,9,2,0,10,5,0,0,10,3}
	4	{ p_5, p_6, p_2, p_9, p_8, p_7}	{9,8,1,0,0,5,0,0,10,3}

4.3.3 生成推荐页面集

为了得到一个用来推荐的页面集合，同时对这些页面进行排序，需要利用每一个类的泊松参数为该类中的每一个页面计算推荐值。这样，每一个类除了拥有前面章节中创建的参数集合，还拥有一系列的推荐值集。修改聚类参数以使每一个类有一个推荐值集，$RS_h=\{rs_{h1},\cdots,rs_{hn}\}$，其中，$rs_{hi},i\in[1,\cdots,n]$是类$c_h$中对于页面 p_i 的推荐值。被更新的聚类的参数形式为：$pc_h=\{\xi_h;(\varphi_{h1},\cdots,\varphi_{hn});(rs_{h1},\cdots,rs_{hn})\}$。这些是系统需要的用来生成推荐页面集的参数。定义存储在内存中的参数个数为模型大小。显然，模型越小，预测速度越快。

使用五种方法来为每一个页面计算其推荐值，并对推荐值进行归一化处理，使得其最大值为 1。这五种方法如下。

方法 1：仅仅使用泊松参数作为推荐值，也就是说：

$$rs_{hi}=\varphi_{hi} \tag{4.49}$$

对于后面的计算，将训练集合中的每一个 session 分配给一个具有最大后验可能性的类。然后，计算每一个聚类中的每一个页面的请求数量。定义该数量为流行度，(f_{hi})，其中$i\in[1,\cdots,n],h\in[1,\cdots,H]$。例如，如果$R_{p_i}$为聚类$c_h$中对页面 p_i 的全部请求数量，而 R_p 是对所有页面的请求数量，则页面 p_i 在该类中的流行度为：$f_{hi}=R_{p_i}/R_p$。

方法 2：在本方法中，仅使用流行度这一信息进行网页推荐。直观的想法就是，推荐最可能被访问的页面。这样，在聚类c_h中，p_i的推荐值为：

$$rs_{hi} = f_{hi} \tag{4.50}$$

方法 3：在这种方法中，将流行度和泊松参数的乘积作为推荐值：

$$rs_{hi} = f_{hi} \times \varphi_{hi} \tag{4.51}$$

方法 4：采用平均信息量的熵值作为推荐值。根据本章提出的聚类条件，在一个聚类内的会话中对某一特定网页的归一化的访问时间值相差不会很大。因此，利用决策论中的熵值作为推荐值。使用归一化时间的十个可能的访问时间值的相对频率来计算每一个页面的熵。熵值小说明对该页访问的时间比较固定，而熵值大则说明不同的会话对该网页的访问时间的差别较大。推荐值计算方法如下：

$$rs_{hi} = f_{hi} \times \frac{1}{(entropy)_{hi}} \times \varphi_{hi} \tag{4.52}$$

方法 5：使用流行度的对数值来减少流行度在推荐值中的影响。

$$rs_{hi} = \left[\log f_{hi}\right] \times \frac{1}{(entropy)_{hi}} \times \varphi_{hi} \tag{4.53}$$

4.4 实验结果与分析

为了评估结合相对停留时间的基于泊松分布的用户兴趣建模方法，实验中设定了不同的初始参数值运行 EM 算法，并且为最大后验概率估计设定了不同的先验分布参数。

对于最大后验概率估计问题，对 Gamma 和 Dirichlet 先验进行了从 1 到 5 不同参数的实验。对于泊松分布，使用简单的先验概率，具体说来，使 Gamma 分布中的所有的α_i和β_i的值相等。由于数据集稀疏，最初以为利用最大后验概率估计会得到更好的结果。然而，实验却得到了令人惊讶的结果，最大后验概率估计的结果和最大似然估计的结果竟然相差不到 0.001%。由于最大似然估计在模型训练阶段相对不复杂，这样，在接下来的实验中，使用最大似然估计，分别用不同的聚类数进行实验。在实验中，试验了从 0.1 到 0.9 不同的推荐阈值。如果设定的推荐阈值高，那么推荐结果集中包含很少的内容。如果设定的推荐阈值低，那么有着很低推荐值的非相关网页也将会被推荐。实验表明，将阈值设定为 0.5，可以产生很少的且高质量的相关推荐。同时使用从 5 到 30 不同的聚类数进行实验，实验表明归一化时间在 1 和 2 之间时，可以提高预测的准确性。为了进行比较，这里仅仅给出了运行归一化时间值在 1-2 之

间和 1-10 之间的实验结果。图 4.4 和图 4.5 是在 NASA 数据集上进行的实验结果。从图中可以看出，归一化时间在 1-2 之间可以增加预测的准确度。本文，定义有最高点击率的聚类数为最佳聚类数。如果归一化时间范围发生变化，那么最佳聚类数也会随之发生变化。可以看出，当归一化时间在 1-2 之间时，测试数据集中的最佳聚类数为 30。图 4.6 是归一化时间在 1-2 之间，拥有最佳聚类数时，基于泊松分布的模型在 NASA 上的实验结果。

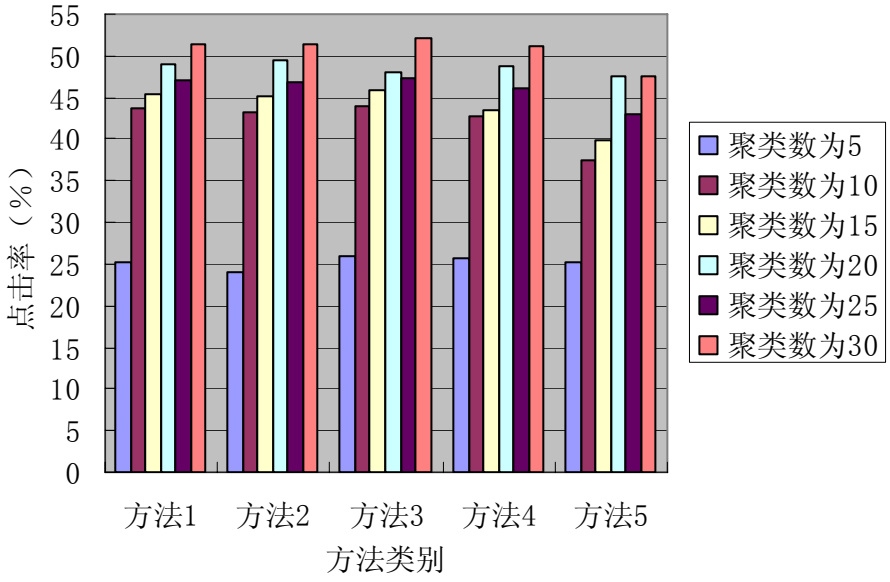

图 4.4 归一化访问时间为 1-2 时，基于泊松分布模型在 NASA 数据集上的实验结果

Fig.4.4 Results in the NASA data when visiting time is normalized between 1 and 2

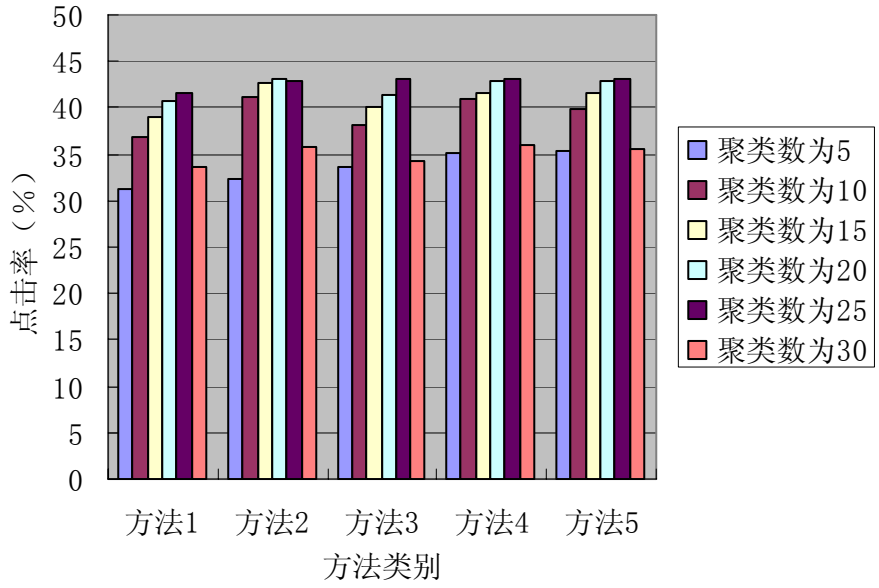

图 4.5 归一化访问时间为 1-10 时，基于泊松分布模型在 NASA 数据集上的实验结果

Fig.4.5 Results in the NASA data when visiting time is normalized between 1 and 10

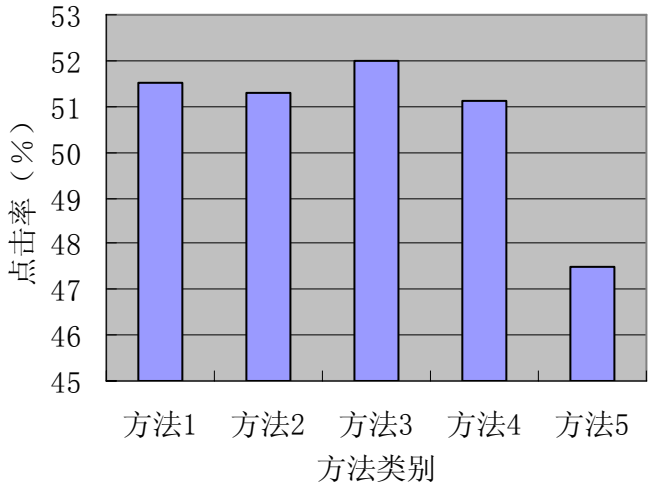

图 4.6　归一化访问时间为 1-2，聚类数为 30 时，基于泊松分布模型在 NASA 数据集上的实验结果

Fig.4.6 Results in the NASA data when clustering number is 30 and visiting time is normalized between 1 and 2

为了进一步评估构建的泊松模型，同时也进行了二项分布的实验。在二项分布中，没有考虑归一化的网页访问时间，仅仅使用网页的二值权重对用户会话进行建模，二值权重指的是在用户会话中网页出现或者网页不出现。使用 EM 算法学习二项分布。实验仅仅使用前三种方法计算推荐值，实验结果如图 4.7 所示。

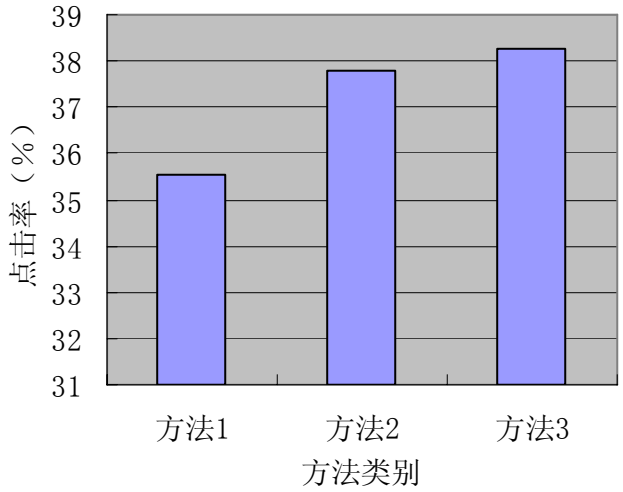

图 4.7　聚类数为 30 时，基于二项分布模型在 NASA 数据集上的实验结果

Fig.4.7 Results in the NASA data when using Binomial distribution and the clustering number is 30

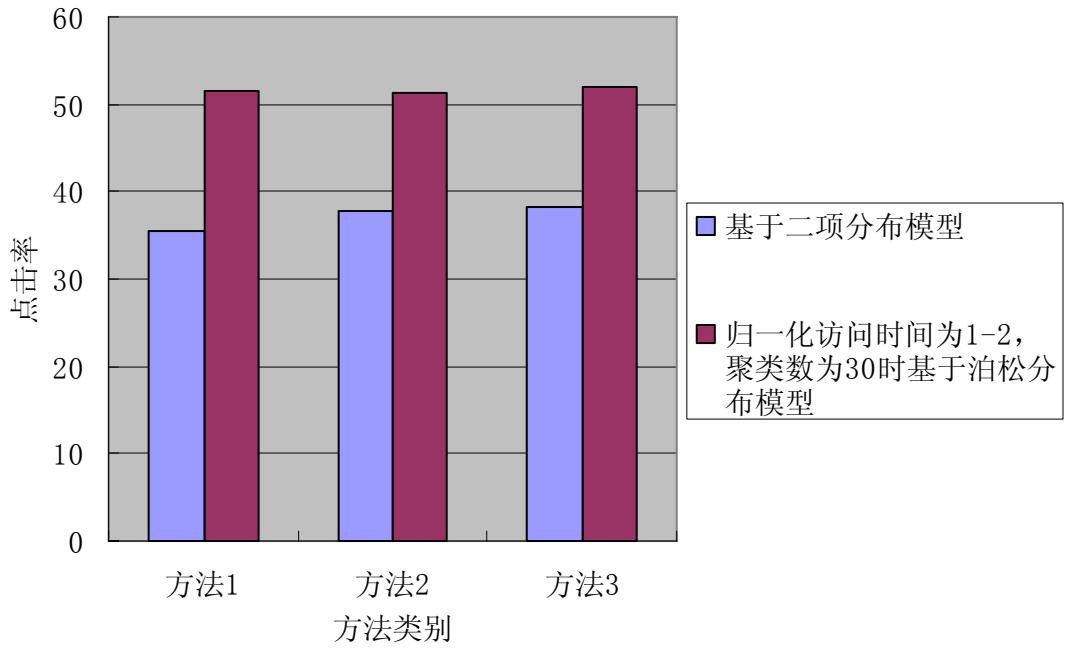

图 4.8 基于二项分布模型和归一化访问时间为 1-2，聚类数为 30 时基于泊松分布模型的网页推荐方法的点击率对比

Fig.4.8 the hit-ratio comparision between the web page recommendation methods based on Binomial distribution model and the Poisson model when clustering number is 30 and visiting time is normalized from 1 and 2

从图 4.8 的对比实验结果中可以看出，基于泊松模型的网页推荐算法比基于二项分布模型的点击率高，推荐效果要好得多。

所有的这些结果都表明了泊松模型可以用来对用户会话进行建模，使用归一化的访问时间可以提高准确率。除了本章提及的实验和柱状图，也研究了每一个可能的归一化网页访问时间的出现频率。根据本文提出的聚类规则，类间的归一化访问时间不会有很大变化。在对用户会话进行聚类之后，为了观察聚类的结构，文中绘制了柱状图。这些柱状图证明了在同一个聚类中的用户会话中的网页访问时间没有太大变化。这也是使用泊松分布对用户会话进行建模的另一个证据。

4.5 小结

用户模型的构建是进行个性化服务的基础，是进行网页推荐的基础。面对日益增长的 Web 信息，为了满足不同背景、不同目的和不同时期的查询请求，本章提出了基

于泊松分布的用户兴趣建模方法，给出了基于泊松分布的用户兴趣模型的构建过程，并基于该模型对网页进行推荐，论证了利用泊松分布对用户兴趣建模的正确性。通过 EM 算法训练模型参数，用最大似然估计和最大后验概率估计模型参数，一系列的实验结果表明，用基于泊松分布的混合模型对用户事务进行建模，当网页的归一化时间在 1-2 之间时，对用户可能访问的网页进行预测，提供一系列推荐网页供用户选择，能够产生很好的点击率和可以接受的计算复杂性。

第五章 基于改进的汉宁窗函数的信息检索算法

Web 信息是由庞大的、分布在世界各地的资源集合而成，这些 Web 信息资源通常以页面形式出现，每一个页面又可以包含指向世界上任何地方的其他相关 Web 信息资源的链接，人们可以跟随一个链接到其所指向的其他 Web 页面或 Web 站点，并且这一过程可无限次重复，从而带给人们不尽的 Web 信息。

Internet 上的信息资源爆炸性增长除了给人们带来丰富的信息之外，同时也使人们难以准确地获得所需的特定信息，在面对如此庞大的网上信息资源时人们会普遍感到不知所措，从而给信息资源的利用带来极大不便。美国《时代周刊》曾评论："Internet 与其说把人们带入了信息世界，不如说把他们领进了茫茫无际的大海。"因此，如何行之有效地对 Internet 上的各类信息资源进行合理的组织和利用，让人们可以轻松地驾驭信息之舟，已经成为人们广为关注的热点问题之一。

信息检索(Information Retrieval，IR)是指利用一定的检索算法，借助于特定的检索工具，针对用户的检索需求，从结构化或非结构化的文档集中找出与用户需求相关的信息，获取有用知识的过程。它主要处理非结构化的数据，如：文本数据(新闻、科技论文等)，网页(HTML，XML)，多媒体数据(图像、图形、视频、音频)等。典型的IR任务是对于给定的自然语言的文档集或者用户的查询(Query)，查找出和Query相关的经过排序的文档子集[110]。信息检索应用广泛，可应用于数字图书馆(Digital Library)、电子政务、电子学习(E-learning)、电子商务和内容安全等方面。

信息检索的过程可以简单的描述为：用户提交查询条件，信息检索系统根据用户提出的查询条件检索 WWW 中与其相关的文档子集，按照这些文档子集与用户查询条件的相关度排序这些文档，最后按照相关度的降序序列将检索结果返回用户。可以将信息检索的整个过程分为三个方面：信息的存储与组织、信息的检索、信息的展示。图 5.1 给出了信息检索原理示意图。

信息检索的目标是用快速、有效的方式搜索到用户想要获得的文档，就是从纷繁复杂的大量信息中筛选出符合用户需求的信息[111]。它首先要解决的问题就是如何表示要检索的文档和查询需求，以及如何计算文档与查询间的相关程度。人们已经提出了大量的检索模型及能够准确反映查询和文档之间相似度的计算模型，其中主要的三个基本检索模型是布尔模型、向量模型和概率模型[112]。

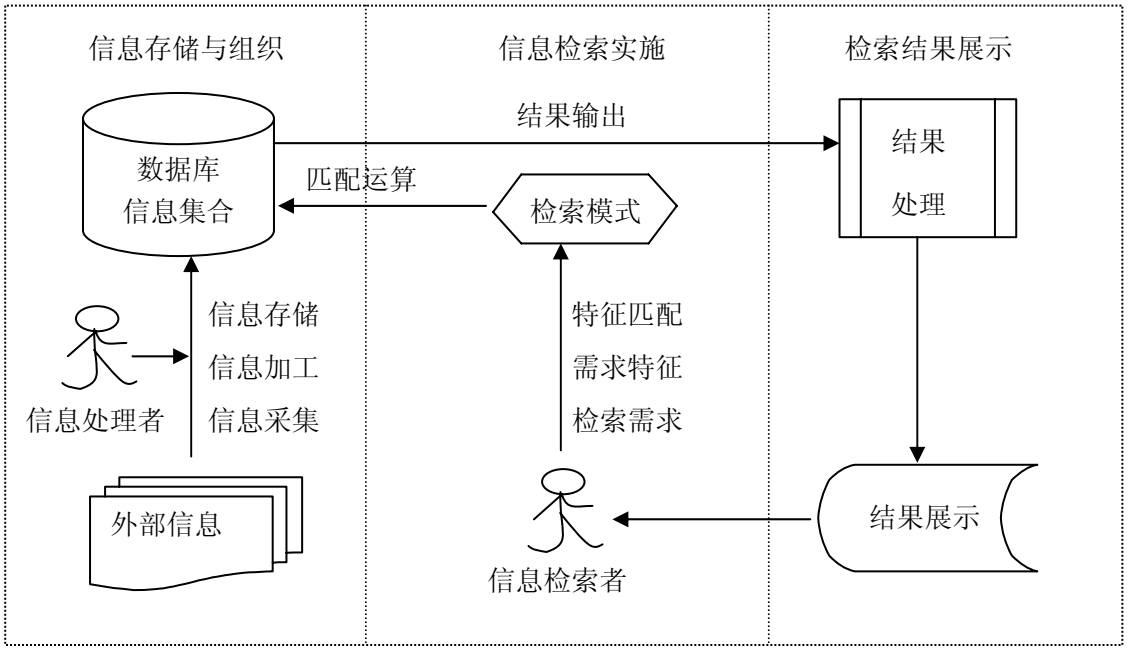

图 5.1　信息检索原理示意图

Fig.5.1 The chart of information retrieval principle

5.1 几种检索模型

5.1.1 布尔模型

布尔模型是最早提出的一个信息检索模型。1957 年，巴·希列尔（Y. Bar-Hille）就对布尔模型应用于计算机信息检索的可能性进行了探讨；20 世纪 60 年代末期，布尔检索模型正式被大型文献检索系统所采用；70 年代时逐渐成为各种商业性联机检索服务系统的标准检索模式。

布尔模型在解释信息检索处理过程时，主要遵循以下两条基本规则：

（1）每一个索引词在一篇文档中只有两种状态：出现或者不出现。相应的，每个索引词的权值 $w_{ij} \in \{0,1\}$

（2）检索条件 q 由三种布尔逻辑运算符“and”、“or”、“not”连接构成。

对于任一篇文档 d_j，定义 d_j 与用户查询 q 的匹配函数为：

$$sim(d_i,q)=\begin{cases}1 & \text{如果文档向量与查询条件匹配}\\ 0 & \text{其他}\end{cases} \tag{5.1}$$

布尔模型是基于集合论和布尔代数的一个简单的检索模型，它具有简单、容易理解和简单的形式化等突出优点。但是，它的查询条件与文档间的相似性只分为相关和

非相关两种情况，它精确匹配文档和查询条件，没有考虑他们之间的部分匹配，可能导致因查询词在大量文档中出现而返回过多的匹配文档，或者由于找不到与查询词匹配的文档而返回太少的检索结果[113]；布尔模型没有提供评价函数，返回的检索结果中的文档地位平等，不能反映出返回的文档与用户查询条件的相关程度。

5.1.2 向量空间模型

20 世纪 60 年代末期，信息处理专家、美国著名学者萨尔顿（G.Salton）基于“部分匹配”策略的信息检索思想在其开发的实验性检索系统 SMART 中最早提出并采用线性代数的理论和方法构建出一种新型的检索模型——向量空间模型（Vector Space Model， 简称 VSM）。向量模型的信息检索方法提出了部分匹配的框架结构，首先用向量表示用户查询和文档，然后采用 tf 和 idf 的概念计算相似度。

向量空间模型在文本信息处理领域中一直占据着非常重要的地位。简单的形式化表示、有效的匹配算法设计以及已取得的较为满意的处理结果，已使得基于向量空间模型的研究思路大为流行，并近乎成为文本处理领域的经典方法。典型的基于向量空间模型的文本信息处理主要包括以下几个分支领域：文本信息检索（Text Retrieval）、文本分类（Text Categorization/Classification）、文本过滤（Text Filtering）、文本聚类（Text Clustering）、文本浏览与可视化（Text Browsing and Visualization）等。

向量空间模型中，文档d_j与用户查询q的相似度值可以通过式（5.2）获得：

$$sim(d_i,q)=(d_j\cdot q)\Big/(|d_j|\times|q|)=\frac{\sum_{i=1}^{t}w_{ij}\times w_{iq}}{\sqrt{\sum_{i=1}^{t}w_{ij}^{\ 2}}\times\sqrt{\sum_{i=1}^{t}w_{iq}^{\ 2}}} \qquad (5.2)$$

向量空间模型理论也存在着明显缺陷，其中，最为后来研究人员不能认同的是对从文档中抽取出的各索引词之间的关系做了相互独立的基本假定，也就是两两正交假设。这一正交假设在实际的文本信息处理环境中一般是难以满足的。向量空间模型理论在应用过程中也存在着一些困难，对处理结果的可解释性较差。所有的这些问题，仍有待于进一步的研究与探索。

5.1.3 概率模型

经典的概率模型是一种实现简单、效果较好的信息检索模型，最早于 1976 年由英国城市大学的罗伯逊（S.E.Robertson）和斯帕克·琼斯（K.Sparck-Jones）提出。

概率模型建立在概率论框架之上，采用概率论原理来解决信息检索问题，通过估

计文档与用户查询条件的相关概率对文档集进行排序。整个检索是对结果不断采用判断和反馈机制，对检索策略不断优化和改进，使检索结果逐渐趋近理想中的结果集合。它需要假定初始的相关和不相关文档集合，没有考虑文档内部索引检索词的频率信息，假定索引检索词是相互独立的。

在经典概率模型中，文档和用户检索提问用索引词向量来表示，每一个索引词的权值是二值的，非 0 即 1，即 $w_{ij}\in\{0,1\}, w_{iq}\in\{0,1\}$，其中 w_{ij} 代表第 i 个索引词 k_i 在文档 d_j 中的权重。给定一个用户检索提问 q，则存在一个相关文档集合 R。令 R_C 为 R 的补集，即非相关文档集合，同时令 $P\left(R\middle|d_j\right)$ 表示文档 d_j 与提问 q 相关的概率，$P\left(R_C\middle|d_j\right)$ 表示文档 d_j 与提问 q 不相关的概率，则 d_j 和 q 之间的相似度 d_j 与提问 $sim(d_j,q)$ 可以定义为：$sim\left(d_j,q\right)=P\left(R\middle|d_j\right)\Big/P\left(R_C\middle|d_j\right)$

利用贝叶斯公式，则有

$$sim\left(d_j,q\right)=\left(P\left(d_j\middle|R\right)\times P(R)\right)\Big/\left(P\left(d_j\middle|R_C\right)\times P\left(R_C\right)\right) \tag{5.3}$$

式（5.3）中，$P\left(d_j\middle|R\right)$ 表示从相关文档集合 R 中随机选择文档 d_j 的概率，或者说文档 d_j 属于相关文档集合 R 的概率；$P\left(d_j\middle|R_C\right)$ 表示从非相关文档集合 R_C 中随机选择文档 d_j 的概率，也即文档 d_j 属于非相关文档集合 R_C 的概率；$P\left(R\right)$ 和 $P\left(R_C\right)$ 则分别表示在整个文档集合中随机选择一篇文档是相关和不相关时的先验概率。由于 $P\left(R\right)$ 和 $P\left(R_C\right)$ 的值对于所有文档来说都是一样的，因此

$sim\left(d_j,q\right)\backsim P\left(d_j\middle|R\right)\Big/P\left(d_j\middle|R_C\right)$ 和 $P\left(R_C\right)$

假定索引词之间时相互独立的，又有：

$$sim\left(d_j,q\right)\backsim\frac{\left(\prod_{w_{ij}=1}P\left(k_i\middle|R\right)\right)\times\left(\prod_{w_{ij}=0}P\left(Nonk_i\middle|R\right)\right)}{\left(\prod_{w_{ij}=1}P\left(k_i\middle|R_C\right)\right)\times\left(\prod_{w_{ij}=0}P\left(Nonk_i\middle|R_C\right)\right)} \tag{5.4}$$

式（5.4）中，$P\left(k_i\middle|R\right)$ 和 $P\left(Nonk_i\middle|R\right)$ 分别表示从文档集合 R 中随机选择一篇文档，其中含有索引词 k_i 和不含有索引词 k_i 时的概率，$P\left(k_i\middle|R_C\right)$ 和 $P\left(Nonk_i\middle|R_C\right)$ 分别表示从非相关文档集合 R_C 中选择一篇文档，其中含有索引词 k_i 和不含有索引词 k_i 时的概率。对式（5.4）取对数，又有 $P\left(k_i\middle|R\right)+P\left(Nonk_i\middle|R\right)=1$ 和 $P\left(k_i\middle|R_C\right)+P\left(Nonk_i\middle|R_C\right)=1$，则有：

$$sim\left(d_j,q\right)\backsim\sum_{i=1}^{t}w_{iq}\times w_{ij}\times\log\frac{P\left(k_i\middle|R\right)\times\left(1-P\left(k_i\middle|R_C\right)\right)}{P\left(k_i\middle|R_C\right)\times\left(1-P\left(k_i\middle|R\right)\right)}$$

进一步地，可以简记为：

$$sim\left(d_j,q\right)\backsim\sum\log\frac{P\left(k_i\left|R\right.\right)\times\left(1-P\left(k_i\left|R_C\right.\right)\right)}{P\left(k_i\left|R_C\right.\right)\times\left(1-P\left(k_i\left|R\right.\right)\right)} \tag{5.5}$$

式（5.5）就是概念模型得到检索结果并对检索结果实施排序输出的主要计算依据。

概率模型不同于布尔模型和向量空间模型，它具有一种内在的相关反馈机制，它把检索处理过程看作是一个不断逼进并最终确认命中文档集合特征的过程，并通过运用某种归纳式学习方法实现系统对检索结果的优化与完善。但是，概念模型仍然存在一定的局限性。各种参数估计难度较大，索引词权值的计算方法为 0/1 式，没有考虑到词频等加权因素，沿用了索引词之间相互独立的假定等。

用户进行检索时，不会关注返回结果的多少，而是更加关注返回结果是否和自己的需求相吻合[115]。用户在返回的结果中一遍一遍筛选满足自己需求的信息，增加了用户的检索负担。而且，在现实生活中，社会成员的信息需求千差万别。但是，现有的检索模型检索时间长，检索结果质量差，无法适应用户群体的多样性，无法根据用户访问Web的兴趣、爱好和使用目的，最大程度地满足用户的个性化需求。如何提高信息检索结果的精度和检索有效性，进一步改善信息检索质量和效果是人们面对的一大难题[116]。本章中，通过分析用户模型，以自然语言作为检索语言进行语义检索，考虑了检索词含义、顺序及词密度，提出了基于改进的汉宁窗函数的信息检索模型。

5.2 基于改进的汉宁窗函数的信息检索模型

5.2.1 基于《知网》的检索关键词概念扩展

如果从检索思想的本质入手分析检索方式，基于串匹配的检索都属于“关键词检索”的范畴。关键词检索的弊端显而易见，一篇以“计算机”为主题的文档通篇没有出现“电脑”这个词，仅仅基于关键词字面匹配的方法，当用户输入“电脑”这个检索词时，该文档根本无法命中，但实际上，“计算机”和“电脑”在很多情况下都表达相同的意思，也就是说概念相同。所以，有必要对用户输入的查询关键词进行概念扩展，以得到更好的检索结果。

知网（英文名称为HowNet）是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库，由多个数据文件组成[96]。由于用户群体的多样性，如：知识水平、文化习惯的差异，可能导致他们使用不同的词进行检索。为了更好的表达用户的查询需求，对用户输入

的查询关键词进行了基于“知网”的关键词概念扩展。概念是对事物本质特征的概括和抽象[98]，它不受词汇语种、多义性和歧义性的影响，对用户输入的查询关键词进行概念扩展，可以更好地表达用户的查询条件，以使返回结果更能满足用户的检索需求。

用户提交一系列的查询关键词$Q=\{q_1,q_2,\cdots,q_n\}$后，在知网中查找每个关键词的同义词，得到Q的同义词集合$Syn_Q=\{Syn_{q_1},Syn_{q_2},\cdots,Syn_{q_n}\}$，将他们也作为查询关键词加入到查询条件中，得到扩展后的查询关键词$Q'=\{q_1,q_2,\cdots,q_n,Syn_{q_1},Syn_{q_2},\cdots,Syn_{q_n}\}$。

5.2.2 改进的汉宁窗函数

事实上，查询词在文档出现的密度可以表达文档与查询词之间的相符程度。查询词在文档中出现的密度越大，表明该文档越满足用户提出的检索要求。本文定义了一个评估函数——改进的汉宁窗函数，当用户提出的查询词在某一汉宁窗口中出现的密度很大的时候，就给该窗口分配一个很高的权重值[117]。也就是说，在一个汉宁窗口中用户的查询词密度越大，那个汉宁窗口越重要。如果一个文档中，出现权值很高的汉宁窗口的数目很多，那么就认为这篇文档满足用户的检索需求。改进的汉宁窗评估函数就是基于这个很简单的想法提出来的。词密度的例子如图 5.2 所示。

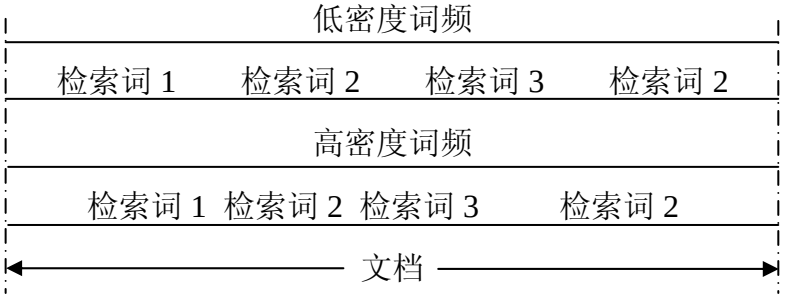

图 5.2 词密度的例子
Fig.5.2 Examples of words density

假定W是汉宁窗的大小（本文就是一篇文档中的一个段落），l是窗口的中心位置，i是文档中的位置。改进的汉宁窗函数$f_H(i,l)$定义如下：

$$f_H(i,l) \overset{def}{=} \begin{cases} \frac{1}{2}\left(1+\cos 2\pi \frac{i-l}{W}\right) & \left(|i-l| \le W/2\right) \\ 0 & \left(|i-l| > W/2\right) \end{cases} \quad (5.6)$$

定义窗口中心l的值$S(l)$为：

$$S(l) \overset{def}{=} \sum_{i=l-W/2}^{l+W/2} f_H(i,l) \cdot a(i) \quad (5.7)$$

其中，$a(i) \overset{def}{=} \begin{cases} idf(t_i) & \text{如果}t_i\text{出现在i位置之后} \\ 0 & \text{否则} \end{cases}$，

$idf(t_i) \overset{def}{=} \log\left(\frac{D}{df(t_i)}\right)$，$df(t_i)$是包含关键词$t_i$的文档出现的频率，$D$是搜索引擎的 web 文档总数。

改进的汉宁窗函数如图 5.3 所示。

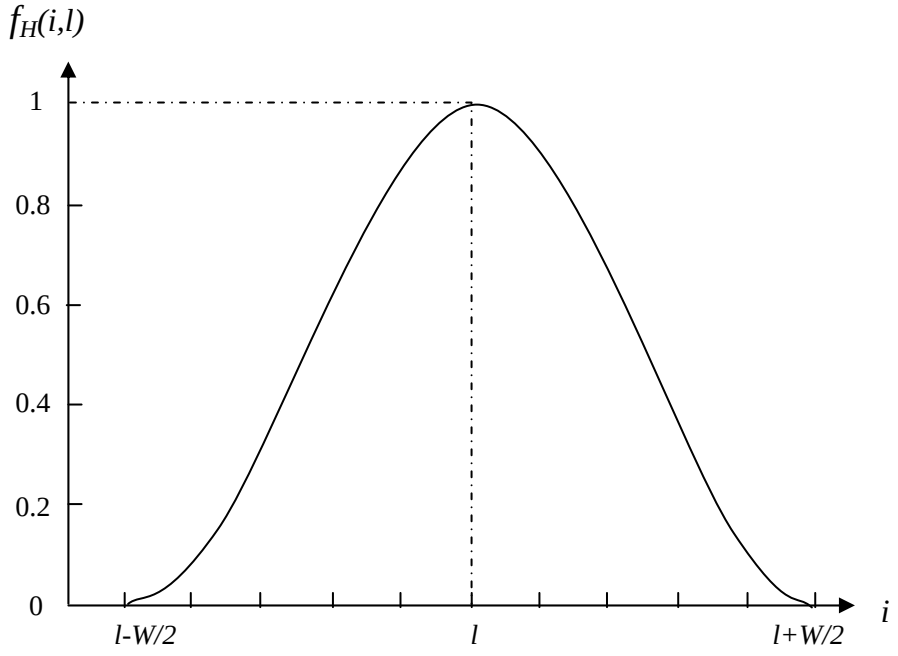

图 5.3　改进的汉宁窗函数

Fig.5.3　Improved Hanning Window Function

5.2.3 基于改进的汉宁窗函数的信息检索算法

输入：扩展后的查询关键词集合Q'，搜索引擎返回的前 m 项结果集合R(文中

令 $m=100$)。

输出：排序后的查询文档结果集

基于改进的汉宁窗函数的信息检索算法步骤：

初始化：$i=1$， $j=1$；

（1） 把结果集 R 中的一篇文档划分为自然段 $p_1, p_2, \cdots, p_n$，令 $P=\{p_1, p_2, \cdots, p_n\}$；

（2） 选取 P 中的一个自然段 p_i；

（3） 从 p_i 段的起始开始移动汉宁窗，每次移动一个字符位置，直到段结束，按照公式 (6.2) 依次计算 $S(l)$ ($1 \le l \le |p_i|$，$|p_i|$ 为 p_i 所含的字符数)；

（4） 令 $S_{p_i}=\max S(l)(1 \le l \le |p_i|)$；

（5） 如果i<n，那么i++,重复上述步骤(2)～(4)；

（6） 令 $S_{d_j}=\sum_{i=1}^{n} S_{p_i}$ 作为当前文档满足用户检索需求的度量函数；

（7） 如果j<m，那么j++，重复上述步骤(1)～(6)；

（8） 按 S_{d_j} ($1 \le j \le m$)的递减顺序输出 R 中文档对应的链接。

5.3 实验与结果分析

任何对信息检索技术的评价在一定程度上都是基于用户对查询结果的相关性判断,而这是一个非常主观的概念,因此判断一个网络搜索工具的性能是一件十分困难的事情。本文采用评价检索系统的两个主要指标查准率(precision)和召回率（recall）[118]来评价基于改进的汉宁窗函数的信息检索模型。查准率和召回率的定义如下：

$$\text{查准率}=\frac{\text{检索结果中与检索相关的文档数}}{\text{检索结果中的文档总数}}$$

$$\text{召回率}=\frac{\text{检索结果中与检索相关的文档数}}{\text{文档库中所有和检索相关的文档总数}}$$

5.3.1 扩展检索关键词对查准率及召回率的影响

让6名学生使用各自熟悉领域中的60组关键词以及使用同样的关键词及扩展后的

查询关键词集在 Google 上进行检索。通过他们自己分析两种检索方式的检索结果，得到了如图 5.4a 所示的查准率及如图 5.4b 所示的召回率。从实验结果中可以看出：使用检索关键词概念扩展后，检索的查准率和召回率有所提高,但是提高的幅度不大。这主要是因为使用“知网”进行检索关键词的概念扩展后，虽然扩大了检索的范围，可以获得由用户提供的检索关键词所不能找到的用户需要的网页，但也同时导致了检索结果中包含了一些用户根本不需要的网页。

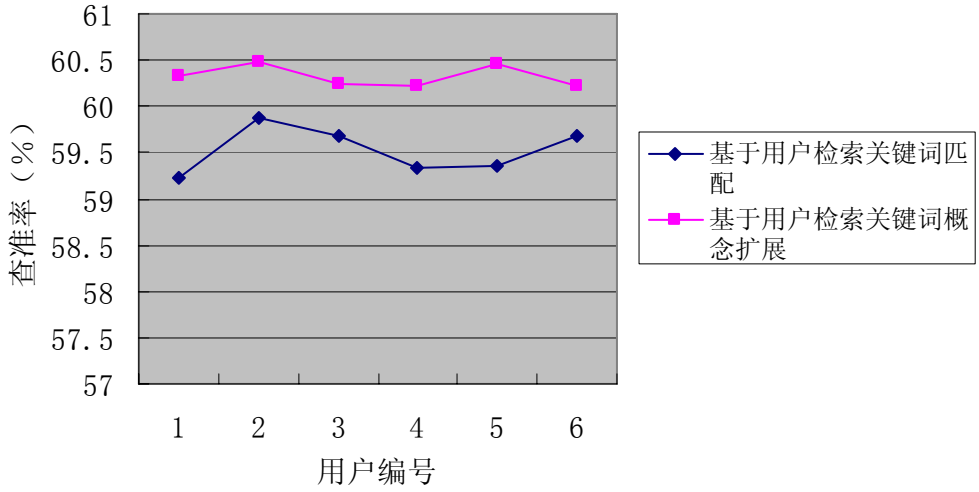

图 5.4a 扩展检索关键词概念对查准率的影响

Fig.5.4a the impact of the precision after extending the retrieval keywords’ concepets

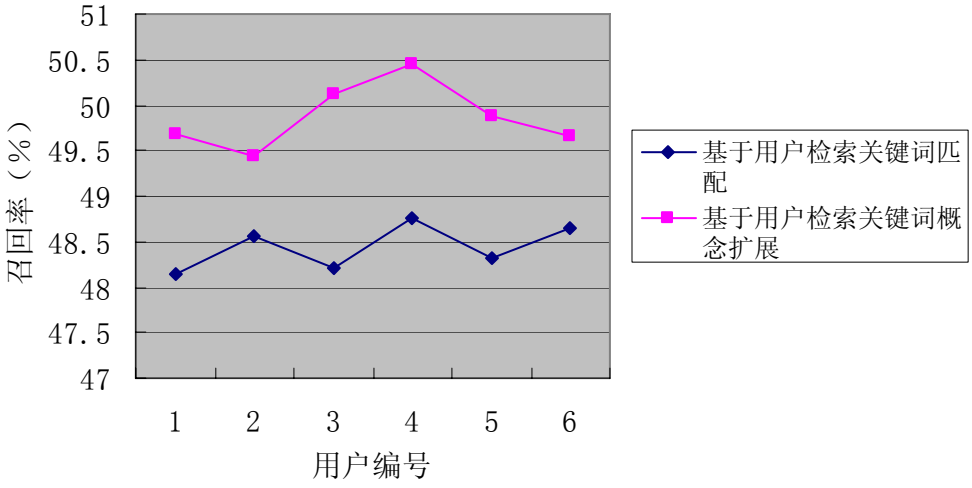

图 5.4b 扩展检索关键词概念对召回率的影响

Fig.5.4b the impact of the recall after extending the retrieval keywords’ concepets

5.3.2 基于改进的汉宁窗函数的信息检索算法的查准率及召回率

继续让这6名学生使用上述实验中的检索关键词在Google上进行基于改进的汉宁窗函数的信息检索算法的检索。通过他们自己分析，检索结果的查准率及召回率如图5.5 所示。从实验结果不难看出：使用检索关键词概念扩展后，再利用基于改进的汉宁窗函数的信息检索算法进行检索，使得检索的查准率和召回率有大幅提高。

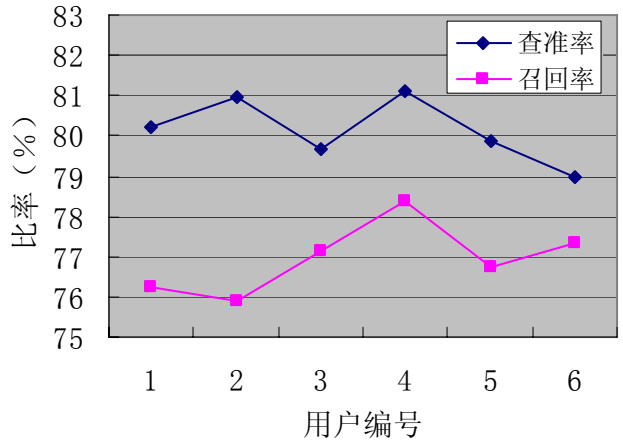

图 5.5 基于改进的汉宁窗函数的信息检索算法的查准率及召回率

Fig.5.5 The precision and the recall of the information retrieval algorithm based on improved Hanning Window

5.3.3 基于汉宁窗函数的信息检索算法与基于改进的汉宁窗函数的信息检索算法的查准率及召回率比较

仍让这6名学生使用上述实验中的检索关键词在Google上进行基于汉宁窗的信息检索算法的检索。通过他们自己分析，对比基于改进汉宁窗的信息检索算法的检索结果，查准率及召回率比较如图 5.6a 及 5.6b 所示。从实验结果不难看出：使用基于改进的汉宁窗函数的信息检索算法比使用基于汉宁窗的信息检索算法的查准率和召回率均有提高。

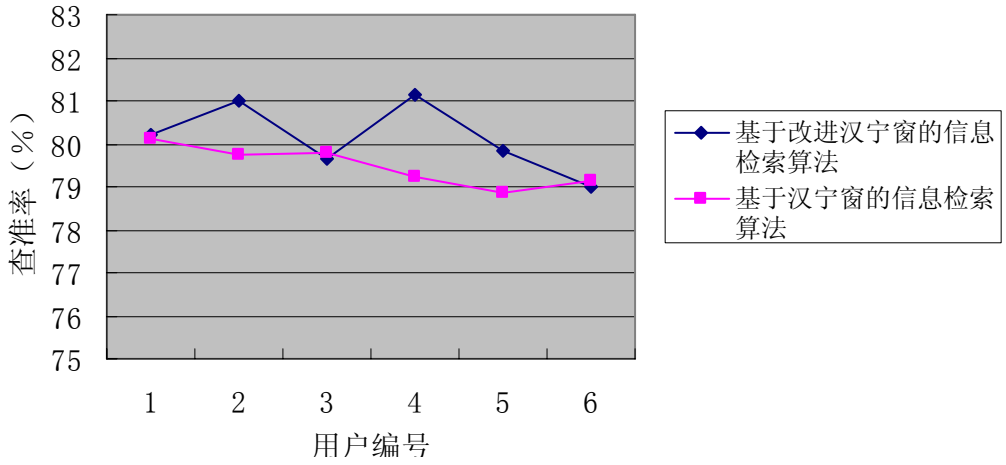

图 5.6a 基于改进的汉宁窗与基于汉宁窗的信息检索算法查准率比较

Fig.5.6a the precision comparison of the information retrieval algorithms based on improved Hanning window and the Hanning window

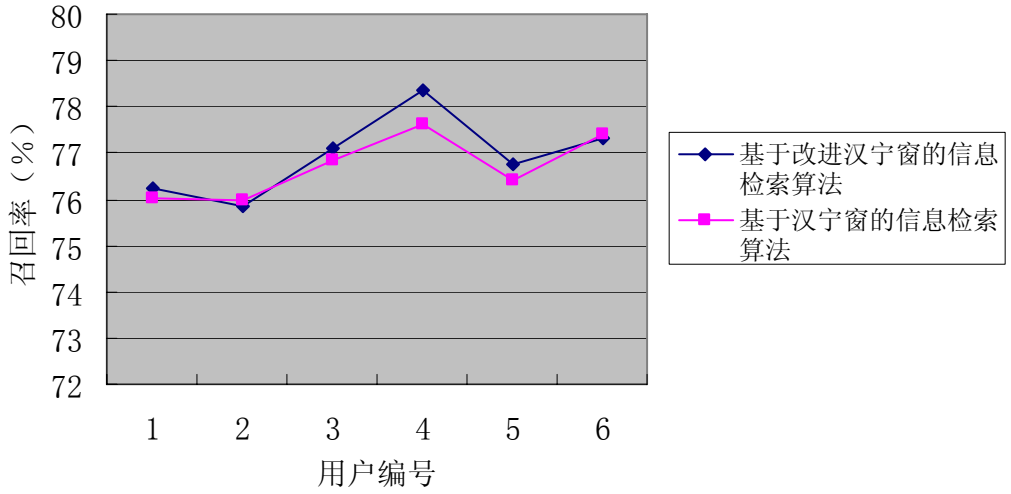

图 5.6b 基于改进的汉宁窗与基于汉宁窗的信息检索算法召回率比较

Fig.5.6b the recall comparison of the information retrieval algorithms based on improved Hanning window and the Hanning window

5.4 小结

面对人类社会不断发展而积累起来的海量知识，如何高效、准确地查找所需要的信息，是每一个人在学习、研究和生活中都无法回避的问题。信息检索技术满足了人们的一定需要，Web 信息检索将 IR 技术应用于 WWW 上的 HTML 网页，致力于捕捉到与用户需求精确匹配的内容。但它仍不能满足不同背景、不同目的和不同时期用户的查询请求。本章通过分析布尔模型、向量空间模型及概率模型的缺陷，考虑检索词的含义、顺序及词密度，将汉宁窗函数进行改进，将改进的汉宁窗函数引入信息检索

模型中，提出了基于改进的汉宁窗函数的信息检索算法，首先对检索关键词进行概念扩展的概念检索，它突破了机械式字面匹配局限于表面形式的缺陷，从词所表达的概念意义层次上来认识和处理用户的检索请求。实验结果表明，该算法可以使得检索的查准率和召回率有大幅提高，很好的改善了检索的性能，提高了信息检索结果的精度和检索有效性，为用户提供更好的个性化检索服务奠定了基础。

第六章 基于 RSS 的个性化信息推送

信息检索系统容易实现，检索速度快，但是它存在的问题也显而易见[119]：

（1） 信息检索没有考虑用户个人的兴趣爱好，不同的用户使用同样的检索关键词检索出来的信息是相同的，可以说信息检索是一种通用性技术。

（2） 用户提交一个查询，返回数以千计的结果，有些是相关的，但大多数是不相关的，用户需要花费大量的时间和精力去作选择。

（3） 信息检索系统提供的是被动式的信息服务，只有用户提交请求，才会得到信息资源，但网上的信息瞬息万变，当信息发生变化时，用户往往无法及时得知最新的资源。用户为了跟踪这些变化，需要进行多次反复查询，浪费大量时间。

个性化服务[120]是一种能够满足用户个体需求的服务。它根据用户提出的明确要求提供准确的信息服务；或通过对用户专业特征、使用偏好的分析而主动地向用户推荐其可能需要的信息。无论采取何种形式描述用户需求，采用何种推荐技术计算结果集，最终它都以一定的推送方式体现出来。推送方式的好坏，将直接影响用户对个性化服务的取舍，也将影响个性化推荐系统的性能。通过研究大量提供个性化服务的网站，发现主要有 3 种信息推送服务方式：频道推送、邮件式推送和用户专用信息网页[121]。这 3 种推送方式存在如下不足：

（1） 用户必须访问相关网站才能获得定制的信息；

（2） 在获取信息的同时，无法屏蔽用户没有订阅的内容和随时弹出的广告；

（3） 通过 E-mail 方式通知用户进行更新信息，病毒和垃圾邮件将给用户带来网络安全上的隐患。

6.1 信息推送技术

所谓信息推送技术，是指依据一定的技术标准和协议，主动从服务器上选择用户所需要的信息，并通过一定的方式有规律地将用户所需要的信息传送给广大订阅用户的技术，也有人称其为广播技术。信息推送技术以其提供检索的主动性、返回信息的新颖性、及时性以及它在个性化信息检索方面的优势，越来越获得人们的青睐。

主动将信息从信息源推送给用户是一种已经存在的信息获取方式，如电视和广播

就是采用这种方式。信息推送技术实际上是传统的广播技术与现代计算机技术（网络技术）相结合的产物。从用户的角度来看，信息推送技术实际上是一种信息获取技术，其基本思想就是用户可以根据服务器所提供的用户界面或服务器的提示信息定制用户所需的信息，服务器根据用户所提供的信息将用户所需的信息传送到用户计算机上，用户可以在想查看的时候再查看它，甚至可以离线浏览。这种技术使用户不必每次都访问网站就可以自动获得由网站主动发送的最新资源。从信息发送方来看，信息推送技术实际上是一种信息发布技术，是网站运营商通过一定的协议，从服务器上的信息源或信息制作商那里获取信息，再通过固定的频道向用户发布信息的技术。应用信息推送技术建立的网络信息广播系统，通过智能化的服务器从网站中不断取回用户所需信息，再将信息进行类聚，同时在服务器上设立固定的“信息频道”、“信息树”等，供用户预定和选择网站的信息。用户联网后，通过客户机随时可获得经过更新的各类信息。

推送技术是一种信息发布技术，意指信息服务机构依据一定的技术标准或协议，主动从网上的信息源或信息制造商选择并获取信息，并以一定的方式（如电子邮件等）有规律地将信息传递给用户的一种技术。Push 技术最早于 1996 年由美国 Point Cast 公司提出，它也因而成为第一个在 Internet 上使用 Push 技术发布信息的公司。Push 技术与传统的使用浏览器查找信息的 Pull 技术不同之处在于：后者由客户机发送服务请求，服务器根据请求进行处理并返回用户所需的结果，在这里用户是数据传输操作的发起者，它从服务器中把信息拉出来。这样，网络上传输的只是用户的请求和服务器针对该请求所作的内应，服务器所提供的服务是被动的。

在 Push 技术系统中，要求用户事先选择所需信息频道，服务器根据用户预先设定好的触发事件和发送内容（不是用户的即时要求），在条件满足时将用户感兴趣的信息推送给客户机系统；虽然数据传输的方向仍然是从服务器流向用户，但操作的发起者却成了服务器，而不是用户。这样服务器所提供的服务是主动的，即在客户端没有请求的情况下主动把数据发送给客户端，因而它又称 webcasting （网播）。

目前常用的信息推送技术主要有以下几种方式[122]：

（1） 频道式推送——是目前网站中普遍采用的一种信息推送方式，它是将某些网页定义为浏览器中的频道，用户可以像选择电视频道那样去选择收看感兴趣的(通过网络播送的信息。关于“网络频道”，目前还没有统一的定义，Microsoft、Nescape 等都有各自的频道定义格式。例如，Microsoft 公司提

出的频道定义格式是为站点信息内容建立的“目标索引文件”，以便于个人化定制信息的推送；而 Nescape 提出的则是基于“元内容”的网播方式。

（2） 邮件式推送——用 E-mail 将有关信息主动传送给用户。

（3） 网页式推送——在特定网页（如某企业、某机构或某个人的网页）上将信息提供给对此信息感兴趣的用户。

（4） 专用式推送——通过机密的点对点通信方式，将指定的信息发送给专门的用户。

6.2 个性化信息推送服务流程

6.2.1 RSS 技术

RSS是 20 世纪 90 年代末由网景公司（Netscape）为发送新闻标题而开发的，当时称为“推”技术，后来Dave Winner对其进行了扩展和完善。由于RSS技术有不同的源头，所以不同的技术团体对其作出了不同的解释，比如，它可以是“Rich Site Summary”（丰富站点摘要），“RDF Site Summary”（RDF站点摘要），或是“Really Simple Syndication”（真正简单聚合）的缩写[123]。简单说来，RSS是基于XML技术的互联网内容发布和集成技术，是一种描述新闻或其他Web内容的方式，通过“Feed”（提要）将信息传递到网络用户面前，网络用户可以在客户端借助于支持RSS的新闻聚合工具软件，在不打开网站内容页面的情况下阅读支持RSS输出的网站内容。

RSS的原理十分简单，只要内容提供者根据RSS规范对各种信息用RSS格式打包，即创建RSS Feed文件，然后采用“推”技术将其发布到网络中，这个RSS Feed中包含的信息就能直接被其他站点调用。RSS具有即时性、私密性、易用性、免除垃圾邮件等优势。RSS阅读器自动更新用户定制的内容，保持新闻的及时性，在用户不知道有新闻发生时将新闻送到用户面前。RSS技术在信息推送方式上具有很好的动态性、时效性、可操作性，但缺乏交互性，没有考虑用户的个性化需求[124]。而个性化服务在满足用户需求方面，已经形成了一套自己的理论和方法，包括用户模型和推荐技术，但是在个性化推送方式上存有不足。本章将个性化技术与RSS技术相结合，提出了基于RSS技术的个性化信息推送方法。

6.2.2 信息推送系统的设计

一般来说，设计一个信息推送系统包括如下步骤：

（1） 建立用户个性化的兴趣模型；

（2） 根据用户兴趣模型对新信息进行过滤，把用户可能感兴趣的信息自动推送给用户；

（3） 根据用户的反馈，对兴趣模型进行修正。

由上可见，建立一个真正能体现用户兴趣的模型是至关重要的，它对推送系统的质量起着非常重要的作用。

基于 RSS 技术的个性化信息推送服务将个性化技术与 RSS 技术很好结合，按照个性化推荐系统构建原则，信息推送系统的建立分为 4 个阶段：

（1） 采用已有的个性化技术方法建立用户需求模型；

（2） 根据构建的用户模型组织数据，按照某种推荐技术优化结果集；

（3） 利用 RSS 格式包装用户所需信息；

（4） 将信息推送推送给用户。

基于 RSS 技术的个性化信息推送系统的服务流程如图 6.1 所示。

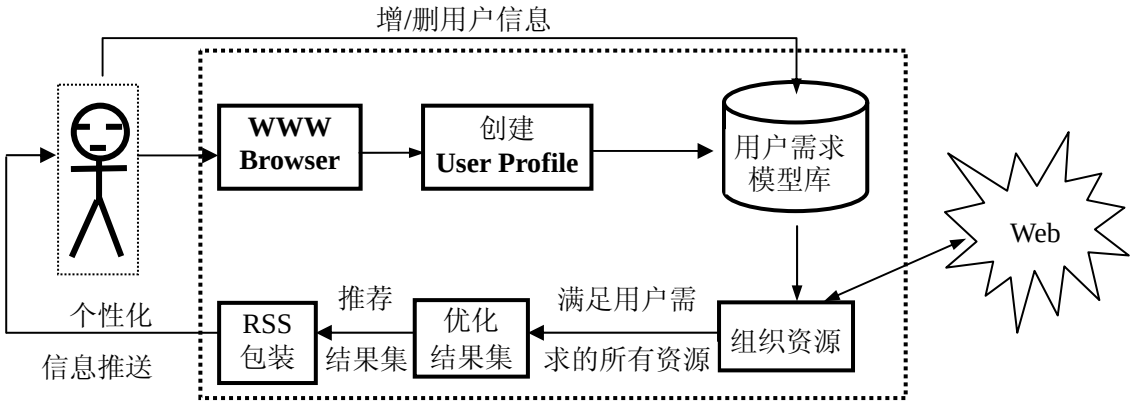

图 6.1 基于 RSS 技术的个性化信息推送系统服务流程图

Fig.6.1 the flowchart of the personalized information pushing system based on RSS technology

6.3 基于 RSS 的信息推送系统的功能组成

为了验证提出方法的可行性，设计开发了一个原型系统，系统由 4 个功能模块组成：构建用户需求模型、资源组织、资源推送和资源更新。

6.3.1 用户需求模型构建

首先采用显示方式全面收集用户信息，构建用户需求模型。包括：用户基本信息和用户需求信息。利用收集到的用户信息，为用户构建用户描述文件（User Profile），每一个用户都对应一个 Profile 文件。用 SQL Server 数据库来存储用户信息（见表 6.1）。

表 6.1 用户需求信息表结构

属性名称	数据类型	是否为空	注释
UserID	Char(10)	N	用户标识
InfoClass	Char(6)	N	感兴趣的资源类别
KeyWord	VChar(50)	N	关键词 需求项的描述信息
Time	DateTime	N	操作时间
needTag	Char(1)	N	需求是否可用

6.3.2 资源组织

系统根据用户信息中感兴趣的资源类别和关键词，自动产生标准的查询请求格式，并作为用户的需求模型加以保存。表 6.2 为存储用户信息需求模型的表结构。资源组织模块定期检查用户的信息需求模型，对每个需求模型利用搜索引擎，用个性化推荐技术中的方法，对可能含有“噪声”的检索结果集进行优化，得到满足用户需求的经过重新组织的信息资源，并交给资源推送模块。

表 6.2 用户信息需求模型表结构

属性名词	数据类型	是否为空	注释
UserID	Char(10)	N	用户标识
Query_SQL	Vchar(50)	N	用户需求模型
Query_Time	DateTime	N	查询操作时间

6.3.3 资源推送

根据 RSS 的语法描述，设计了一个能够根据资源组织模块得到用户需求信息结果记录的 title，identifier-url，description 和 date 字段，自动形成 RSS feed 中 Item 描述的对应项，即<title>、<link>、<description>、< pubDate >的 creator_rss Java 类。当用户

接到 RSS 的 URL 通知，将 RSS 文件的 URL 加入到客户端 RSS 阅读器后，满足用户需求的信息就被“推送”到用户的桌面，并且随着网站上信息的更新，RSS 阅读器中的信息同步将自动更新。这样，一个满足用户需求的个性化信息推送系统就建立起来了。

6.3.4 资源更新

资源更新包括更新已有的信息资源和组建新的信息资源。

（1） 更新已有的信息资源

为了保证用户需求信息的新鲜度，系统按照用户的需求模型定时“增量式”的检索，用检索得到的更新结果结合定义的新鲜度函数修改相应的 RSS 文件。即便 RSS 文件内容有改动，系统也无需通知用户，RSS 阅读器能够自动捕获到相应的信息，提醒用户查阅。

（2） 组建新的信息资源

“拉”[125]机制是，当用户的兴趣改变时，及时向系统提交用户的兴趣描述。在信息推送系统中，如果用户的某个需求改变了，或系统提供的信息不能满足用户要求时，用户可以在RSS阅读器中直接删除系统提供的信息资源。如果用户有新的需求，则在信息推送系统的用户管理界面对用户需求信息表进行添加，同时删除过时的需求信息。

“推”[125]机制是，系统定期地检查每个用户是否有新的需求模型提出，根据用户的需求描述，应用系统的搜索引擎，将满足条件的信息优化整理，以RSS格式打包，然后“推”给用户。

将“推”和“拉”的技术结合，这样，当用户需求发生变化的时候，可以随时删除已建立的信息资源，并能根据新的需求，组建满足用户需求的新的信息资源。

6.4 实验

（1） 用户需求：将标题为”model”的信息组织起来。

（2） 资源组织：查找满足用户需求的所有数据，并进行优化处理。下面是查询到的部分记录（按照 XML 格式显示）。

<title> A Parallel Execution Model for Logic Programming </ title> //结果记录的标题中包含“model”

< subject> All Records </subject>

<description> The Sync Model, a parallel execution method for logic programming, is proposed. The Sync Model is a multiple-solution data driven model that realizes AND-parallelism and OR-parallelism in a logic program assuming a message-passing multiprocessor system.

</ description>

……

< title> Conceptual Development and Empirical Testing of an Outdoor Recreation. Experience Model: The Recreation Experience Matrix (REM)</ title> //结果记录的标题中包含“Model”

<subject> Forestry </subject>

<description> This dissertation examines four issues, including: ...…

</ description >

……

可以看出，在返回的检索结果中，每条记录的 title 中都包含查询词“model”，说明检索结果符合用户的要求。

（3） RSS 包装

根据上面的检索结果，调用 creator_rss 类自动形成 RSS 文件。下面是形成的 RSS 文件的片断，包括频道的描述和一个资源的描述。

<?xml version="1.0" encoding="iso-8859-1" ?>

<rss version="2.0">

<channel> //频道描述

<title>model from DL </title> //频道标题

<description>BIT TestBed</description> //频道简介

<item> //资源描述

<title> A model-based frequency constraint for mining associations from transaction data </title> //资源标题

<link> http://epub.wu-wien.ac.at/dyn/dl/wp/epub-wu-01_7a9 </link>

//资源原文链接

<description> In this paper we develop an alternative to minimum support which utilizes knowledge of the process which generates transaction data and allows for highly skewed frequency distributions. We apply a simple stochastic model (the NB model), which is known for its usefulness to describe item occurrences in transaction data, to develop a frequency constraint. This model-based frequency constraint is used together with a precision threshold to find individual support thresholds for groups of associations.... </description> //资源简介

<pubDate>2004/04/12</pubDate> //资源出版时间

</item>

……

</channel>

</rss>

（4） 利用 RSS 阅读器（RSS Reader） 接收检索结果集

将生成的 RSS 文件对应的网址（URL）加入到 RSS Reader，就为用户构建了一个满足用户需求的信息推送系统。用户通过 RSS Reader 可以得到推送的信息资源内容，如图 6.2 所示。

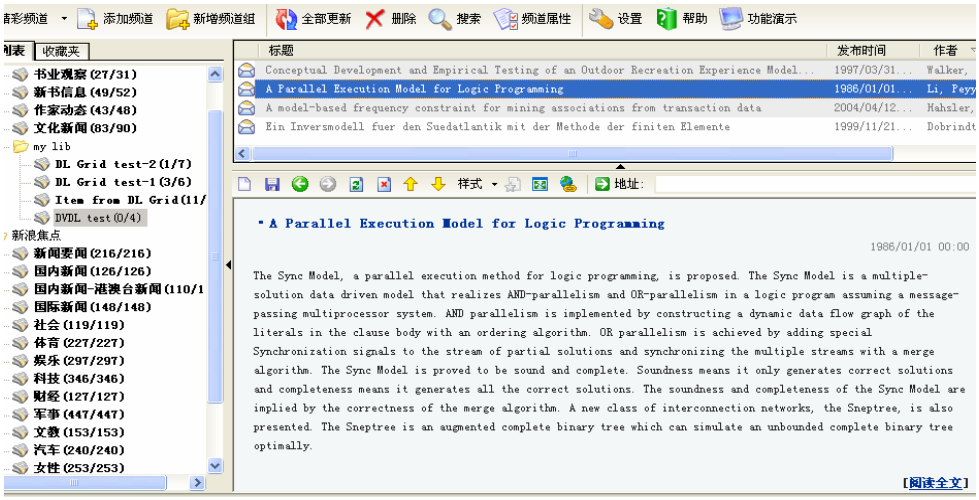

图 6.2 满足用户需求的信息资源内容

Fig.6.2 the information resource satisfied the user requirement

6.5 小结

在 WWW 信息检索服务中，最常用的信息获取方式是“拉”模式，即信息提供者把所提供的信息直接发布在网页上，需要该信息的用户去访问此网站，然后在内容中查找需要的信息。这样将导致用户需要花费大量的时间寻找信息，甚至，有的用户虽然花费了大量的时间在网页之间跳转，但仍没有找到需要的内容。而在“推”模式下，信息提供者直接将最新信息的标题和摘要发布给订阅的用户，然后用户根据自己的需要点击链接访问网站进行阅读。本章分析了“推”技术和“拉”技术及 RSS 技术的优缺点，设计了将“推”技术和“拉”技术有效结合的基于 RSS 技术的个性化信息推送系统，给出了系统的服务流程图，详细说明了系统的模块组成。实例表明，将 RSS 技术和个性化技术相结合的基于 RSS 技术的个性化信息推送方法是可行、有效的，避免了用户网上漫无边际的查找与长时间的等待，提高了信息检索效率，可以为用户提供方便的个性化服务。

结束语

论文工作总结

Internet 上的“信息过载”和“资源迷向”问题的出现，导致人们迫切需要能够提供个性化服务的系统。个性化服务的形式多种多样，但无论采用何种形式，都首先需要建立个性化的用户兴趣模型。如何构建一个能够反映用户真实兴趣，如何了解用户不断变化的兴趣是构建用户兴趣模型的难点。

本文一是围绕着如何建立用户个性化兴趣模型，分析了解用户兴趣，以便进行个性化服务进行了研究。首先研究了建立用户个性化兴趣模型的关键技术，然后研究了用户个性化兴趣模型的更新。二是围绕着用户个性化兴趣模型的应用进行了研究，包括个性化信息检索、个性化信息推荐、个性化信息推送。本文针对 web 数据挖掘用户兴趣建模技术进行了研究，内容创新点有如下几个方面。

（1） 提出了基于标记窗的网页正文信息提取方法

提出了标记窗的概念和基于标记窗的网页正文信息提取方法。该方法能够解决网页正文存放在多个 td 中的情况，能够解决正文文字短的网页的正文提取问题，尤其重要的是，它能够处理非 table 结构的网页正文提取问题。本方法无须将网页表示成一棵树，只需利用正则表达式，就可以直接提取出网页中标记对之间的正文，大大降低了算法的复杂度。

（2） 提出了基于概念关系的用户兴趣模型的构建方法及基于用户兴趣森林模型的网页预取方法

用户访问一个页面后，一般会随着页面的链接来访问其它页面，可以对用户即将访问的链接进行预测，预先下载用户即将访问的页面，从而加快用户的浏览速度。根据用户的访问历史，利用“知网”建立了基于概念关系的用户兴趣森林模型，在此基础上，提出了基于用户兴趣森林模型的预取方法。通过计算超链接描述文字的平均带权语义距离，得到每个超链接描述文字的评价函数，进行网页预取。实验结果表明，该方法的预取命中率在 61%左右，具有较高的系统性能。

（3） 提出了结合用户相对停留时间的用户兴趣模型的构建方法及基于泊松分布模型的网页推荐方法

用户兴趣模型的构建是进行个性化服务的基础，是进行网页推荐的基础。Web 推

荐是基于数据挖掘和机器学习的方法，对用户可能感兴趣的网页进行推荐，它根据用户的爱好兴趣对用户可能访问的网页进行预测，并提供给用户进行选择。利用泊松模型对用户进行建模，当网页的归一化时间在 1-2 之间时，对用户可能访问的网页进行预测，提供一系列推荐网页供用户选择，实验结果点击率在 52％左右。

（4） 提出了基于改进的汉宁窗函数的信息检索模型

传统的基于关键词匹配的检索方法检索时间长，检索结果质量差，无法适应用户群体的多样性。基于改进的汉宁窗函数的信息检索模型从词密度角度对用户的查询进行分析建模，对检索关键词进行概念扩展后进行概念检索，突破了机械式字面匹配局限于表面形式的缺陷，从词所表达的概念意义层次上来认识和处理用户的检索请求，根据用户访问 Web 的兴趣、爱好和使用目的，最大限度的满足用户的个性化检索需求。实验表明，该算法可以使得检索的查准率和召回率有大幅提高，很好的改善了检索的性能，提高了信息检索的有效性。

（5） 提出了基于 RSS 的个性化信息推送方法

将 RSS 技术和个性化技术相结合的基于 RSS 技术的个性化信息推送方法将“推”技术和“拉”技术有效结合，打破了传统的信息获取方式，减少了用户上网搜索的工作量，将个性化的信息直接送给用户，提高了用户获取信息的效率。实例表明此方法是可行、有效的，可以为用户提供方便的个性化服务。

进一步工作

本文对 web 个性化服务中的几个关键技术进行了研究，取得了一些成果。但是，作为一门新兴的技术，对个性化服务中的预取和推荐的研究仍然有大量的工作要做，准备在下面的一些方面进行进一步的研究：

（1） 结合更多的语义信息对用户兴趣模型进行构建。面向用户群体的建模研究。本文构建的用户模型是单用户模型，但是在某些情况下，针对一组用户而不是单个用户兴趣建模更有意义。拓展和完善群体个性化领域研究的新技术和新方法有待进一步提出和实现。

（2） 用户长期兴趣和短期兴趣的集成技术研究。用户长期兴趣反映了用户持续的偏好，短期兴趣则是用户的暂时性需求。将两者有效结合，进一步提高推荐准确性和有效性是今后的努力方向。

（3） 将自然语言作为检索语言，以更符合人们说话语言的方式进行检索，更好

的实现个性化信息检索

（4） 将更多的信息结合到网页预取和推荐方法中，为用户提供更好的个性化服务。

参考文献

[1] Arasu A , Cho J et al. Searching the Web[J]. ACM Transactions on Internet Technology, 2001,1(1):2-43.

[2] 范亚芹，刘颖，李兴男. Web数据挖掘原理及实现[J]. 吉林大学学报（信息科学版）. 2003，21（4）：370-373.

Fan Ya-qin, Liu Ying, Li Xing-nan. Web Data Mining Principle and Implementation[J]. Journal of Jin Lin University(Information Science Edition). 2003,21(4):370-373.(in Chinese)

[3] Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Raghavan P., and Rajagopalan, S. Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text. [C]// Proceedings of 7th World Wide Web Conference, 1998, 30(1-7): 65-74.

[4] S. Chakrabarti et al. Mining the Web's Link Structure[J]. IEEE Computer, 1999,32(8):60-67.

[5] 杨炳儒，李岩，陈新中，王霞. Web结构挖掘[J]. 计算机工程，2003，29（20）：28-30.

Yang Bing-ru, Li Yan, Chen Xin-zhong, Wang Xia. Web Structure Mining[J]. Computer Engineering, 2003,29(20):28-30.(in Chinese)

[6] Davison B, Gerasoulis A, Kleisouris K, Lu Y, Seo H, Wang W, Wu B. DiscoWeb: Applying link analysis to web search (extended abstract). [C]// Proceedings of the 8th ACM-WWW International Conference. Toronto: ACM Press, 1999. 148-149.

[7] H. Kao, S. Lin, J. Ho, and M. Chen. Entropy-based link analysis for mining web informative structures. [C]// Proceedings of the 11th ACM CIKM, 2002:574-581.

[8] Cooley R, Mobasher B, Srivastava J. Web mining: Information and pattern discovery on the World Wide Web. [C]// Proceedings of the 9th International Conference on Tools with Artificial Intelligence. Newport Beach: IEEE Computer Society, 1997. 558-567.

[9] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon. Pagerank, hits and a unified framework for link analysis.[C]// Proceedings of the 25th ACM SIGIR,

2002:353-354.

[10] Liu Jun. Web Usage Mining[D]**.** Master Paper, Nan Jing University Computer Department, 2000.

[11] J. Srivastava, R Cooley et al. Web usage mining: discovery and application of usage patterns from Web data[J]. SIGKDD Explorations, 2002,1(2):1-12.

[12] Yan T.W., Jacobsen M., Garcia-Molina H. &Umeshwar D. From User Access Patterns to Dynamic Hypertext Linking. [C]// Proceedings of 5th international WWW Conference,1996,28(7-11):1007-1014.

[13] Armstrong, R., Freitag, D., Joachims, T., Mitchell, T. WebWatcher: A learning Apprentice for the World Wide Web. [C]// In Working Notes of AAAI Spring Symposium: Information Gathering form Heterogenous, Distributed Environments, Standford University, AAAI Press, 1995:6-12.

[14] Chakrabarti et al. Mining the Web's Link Structure [J]. IEEE Computer, 1999, 32(8):60-67.

[15] Chen M.S, Park J.S, Yu P.S. Data Mining for Path Traversal Patterns in a Web Environment. [C]// Proceedings of the 16th International Conference on Distributed Computing Systems,1996:385-392.

[16] K.-L. Wu, P.S. Yu, and A. Ballman. SpeedTracer: A Web Usage Mining and Analysis Tool[J]. IBM Systems J., 1998, 37(1): 89-105.

[17] Seon-Mi Woo, et al. User-centered Filtering and Document Ranking. [C]// TENCON 99. Proceedings of the IEEE Region 10 Conference, 1999:1059-1062.

[18] 阳小华，刘振宇. 基于浏览过程的LFU：一个新的WWW缓冲清理算法[J]. 计算机工程与应用，2000，36(6):133-134.

Yang Xiao-hua, Liu Zhen-yu. LFU Based on Navigation Processes:A Removal Policy of WWW Caches[J]. Computer Engineering and Applications. 2000,36(6):133-134.(in Chinese)

[19] J. Srivastava, R Cooley et al. Web usage mining: discovery and application of usage patterns from Web data[J]. SIGKDD Explorations, 2002,1(2):1-12.

[20] Page, L. PageRank: Bringing Order to the Web[R]. Stanford Digital Libraries Working Paper,1997-0072.

[21] M. Chau, D. Zeng, and H. Chen. Personalized spiders for Web search and analysis.[C]// Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2001), 2001:79-87.

[22] Filippo Menczer. Combining Link and Content Analysis to Estimate Semantic Similarity[J]. [C]// Proceedings of the 13th International World Wide Web. Conference, 2004:452-453.

[23] Brian D.Davison. Toward a Unification of Text and Link Analysis. [C]// Proceedings of 26th annual international ACM SIGIR conference on Research and development in information retrieval, Toronto, Canada. SIGIR ’03:367-368.

[24] Y. Wang and M. Kitsuregawa. Evaluating contents-link coupled Web page clustering for Web search results. [C]// Proceedings of 11th International Conference on Information and Knowledge Management, 2002:499-506.

[25] Peter Dolog, Nicola Henze, Wolfgang Nejdl, Michael Sintek. Personalization in Distributed e-Learning Environments. [C]// Proceedings of the International World Wide Web Conference, Education Track, New York, 2004:170-179.

[26] Glen. Jeh, Jennifer Widdom. Scaling Personalized Web Search[EB/OL]. Http://www2003.org/cdrom/papers/referred/ html/p185

[27] J. Dean and M. Henzinger. Finding related page in the World Wide Web. [C]// Proceedings of the 8th International World Wide Web Conference, Toronto, 1999:389-401.

[28] D.Hawking, N.Craswell, P.Bailey, et al. Measuring search engine quality[J]. Information Retrieval, 2001, 4(1):33-59.

[29] Miller, C. S., & Remington, R. W. Effects of structure and label ambiguity on information navigation[J]. Conference Extended Abstracts on Human Factors in Computer Systems. New York: ACM Press. 2002:630-631.

[30] Chen Zheng, Liu Wenyin, Peng Xiao. User modeling for building personalized web navigation assistants. [C]// 11th International World Wide Web Conference. Hawaii,USA:ACMPress, 2002.

[31] P.Merialdo , P.Atzeni , G.Mecca. Semistructured and structured data in the web: Going back and forth. [C]// Proceedings of the Workshop on the Management of

Semistrctured Data (in conjunction with ACM SIGMOD), 1997,26(4):16-23.

[32] C. C. Aggarwal. Collaborative Crawling: Mining User Experiences for Topical Resource Discovery[R]. IBM Research Report, 2002.

[33] C. C. Aggarwal, F. Al-Garawi, P. Yu. Intelligent Crawling on the World Wide Web with Arbitrary Predicates. [C]// Proceedings of the 10th international conference on World Wide Web, HongKong, 2001(01-05):96-105.

[34] .IBM. Web site personalization[EB/OL]. http://www-106.ibm/developworks/library/personalization/

[35] Richard Dean. Personalizing Your Web Site[EB/OL]. http://www.builder.com/Business/Personal/index.html.

[36] 丁浩，林云. Internet上的个性化信息服务[J]. 计算机系统应用. 2000，5：36-39.
Ding Hao, Lin Yun. Personalized Information Service in Internent[J]. Applications of the Computer Systems. 2000,5:36-39. (in Chinese)

[37] Alexander Pretshner, Susan Gauch. Personalization on the Web[R]. Technical Report ITTC-FY2000_TR-13591-01. The University of Kansas,1999.

[38] M.Eirinaki, M.Vazirglannis. Web Mining for Web Personalization[J]. ACM Transactions on Internet Technology. 2003,3(1):1-27.

[39] Sun-Mi Woo, et al. User-Centered Document Ranking Technique Using Statistical Analysis. [C]// Proceedings of the 1998 ICCCS , Computer & Communication Center Taegu University, Korea , 1998:159-164.

[40] Collaborative Proposal Submitted to the NSF 99-2. Information and Data Management/IIS/CISE and Computation and Social Systems/IIS/CISE[EB/OL]. Http://scils.rutgers.edu/etc/mongrel/proposal.htm.

[41] 曾春，刑春晓，周立柱. 个性化服务技术综述[J]. 软件学报，2002，13(10):1952-1961.
Zeng Chun, Xing Chun-xiao, Zhou Li-zhu. A Survey of Personalization Technology[J]. Journal of Software. 2002,13(10):1952-1961. (in Chinese)

[42] T. Kurki, S. Jokela, R. Sulonen et al. Agents in delivering personalized content based on semantic metadata.[C]// Proceedings 1999 AAAI Spring Symposium

Workshop on Intelligent Agents in Cyberspace, Stanford, USA,1999:84-93.

[43] H. Libermann. Letizia: An agent that assists Web browsing. [C]// Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-95), 1995: 475-480.

[44] H. Liebermann. Antonomous interface agents. [C]// Proceedings ACM Conference on Computers and Human Interaction, CHI97, Atlanta, USA,1997.

[45] Sliverstri F., Baraglia R., Palmerini P., &Serrano M. On-line Generation of Suggestions for web Users. [C]// Proceedings of IEEE International Conference on Information Technology: Coding and Computing, 2004,(1):392-397.

[46] 郝沁汾，祝明发，郝继升. 一种新的代理缓存替代策略[J]. 计算机研究与发展，2002，39（10）：1178-1185.

Hao Qin-fen, Zhu Ming-fa, Hao Ji-sheng. A New Proxy Cache Replacement Policy[J]. Journal of Computer Research and Development, 2002, 39(10): 1178-1185. (in Chinese)

[47] Zukerman I., Albrecht W., Nicholson A. Predicting user's request on the WWW. [C]// Proceedings of the 7th International Conference on User Modeling. Banff, Canada, 1999.

[48] T. Lau, E. Horvitz. Patterns of Search: Analyzing and Modeling Web Query Refinement. [C]// Proceedings of. the 7th International Conference on User Modeling, 1999:119-128.

[49] Zhong Su, Qiang Yang, Ye Lu, Hongjiang Zhang. WhatNext: A Prediction System for Web Requests using N-gram Sequence Models. [C]// Proceedings of the First International Conference on Web Information Systems Engineering, 2000:200-207.

[50] Bestravros A. Using speculation to reduce server load and service time on the WWW. [C]// Proceedings of the CIKM'95, Baltimore, 1995:403-410.

[51] Schechter S, Krishnan M, Michael DS. Using path profiles to predict http requests.[C]// Proceedings of the seventh International World Wide Web Conference. Brisbane, 1998: 457-467.

[52] 徐宝文，张卫丰. 数据挖掘技术在Web预取中的应用研究[J]. 计算机学报，

2001，24（4）：431-436.

Xu Bao-wen, Zhang Wei-feng. Applying Data Mining to Web Pre-Fetching[J]. Chinese Journal of Computers. 2001, 24(4):431-436. (in Chinese)

[53] 许欢庆，王永成，孙强. 基于隐马尔可夫模型的web网页预取[J]. 上海交通大学学报， 2003，37（3）：404-407.

Xu Huan-qing, Wang Yong-cheng, Sun Qiang. Intelligent Web Pre-Fetching Based on Hidden Markov Model[J]. Journal of Shanghai Jiaotong University. 2003, 37(3):404-407. (in Chinese)

[54] Ken ichi Chinen and Suguru Yamaguchi. An interactive prefetching proxy server for improvement of www latency. [C]// Proceedings of the 7th Annual Conference of the Internet Society, Kuala Lumpur, July 1997.

[55] Dan Duchamp. Prefetching hyperlinks. [C]//Proceeding Of the Second USENIX Symposium on Internet Technologies and Systems, Boulder, CO, USA, 1999:127-138.

[56] E.Marktos and C. Chronaki. A top-10 approach to prefetching on the web. [C]// Proceeding of the INET 98 Conference, Geneva, Switzerland, July, 1998:276-290.

[57] Lieberrnan H, Letizia. An Agent that Assists Web Browsing. [C]// Proceedings of the International Joint Conference on Artificial Intelligence, Montreal, August, 1995:924-929.

[58] Pazzani M, Muramatsu J. and Billsus U. Syskill&Webert. Identifing Interesting WebSites.[C]//Proceedings of the 13th National Conference on Artificial Intelligence, MenloPark. California, 1996: 54-61.

[59] 何军，周明天. 信息网络中的信息过滤技术[J]. 系统工程与电子技术，2001，23（11）：76-79.

He Jun, Zhou Ming-tian. Information Filtering Technology in Information Network[J]. Systems Engineering and Electronics, 2001,23(11):76-79. (in Chinese)

[60] 欧洁，林守勋. 个性化智能信息提取中的用户兴趣发现[J]. 计算机科学，2001，28（3）：112－115.

Ou jie, Lin Shou-xun. Discovering User’s Interests in Personalized Intelligent

Information Retrieval System[J]. Computer Science, 2000, 28(3):112-115. (in Chinese)

[61] Lieberman H., Dyke N. V. and Vivacqua A. Let's Browse: a Collaborative Browsing Agent[J]. Knowledge-Based Systems, 1999, 12: 427-431.

[62] Terveen L., Hill W., Amento B., et al. A System for Sharing Recommendations[J]. Communications of the ACM, March 1997, 40(3): 59-62.

[63] Kautz H., Selman B. and Shah M. Referral Web: Combining Social Nerivorks and Collaborative Filtering[J]. Communications of the ACM, 1997, 40(3):63-65.

[64] Rucker J. and Polanco M.J. Siteseer: Personalized Navigation for the Web[J]. Communications of the ACM. March 1997, 40(3): 73-75.

[65] Konstan J. A., Miller B. N., Maltz D, et al. GroupLens: Applying Collaborative Filtering to Lsenet News[J]. Communications of the ACM, 1997, 40(3):77-87.

[66] 路海明，卢增祥，李衍达. 基于多Agent混合智能实现个性化网络信息推荐[J]. 计算机科学，2000，27（7）：32-34.

Lu Hai-ming, Lu Zeng-xiang, Li Yan-da. Personal Informational Recommendation Service: Based on Multi-Agent Hybrid Intelligence[J]. Computer Science, 2000, 27(7):32-34. (in Chinese)

[67] 赵亮，胡乃静，张守志. 个性化推荐算法设计[J]. 计算机研究与发展，2002，39（8）：986－991.

Zhao Liang, Hu Nai-jing, Zhang Shou-zhi. Algorithm Design for Personalization Recommendation Sysetems[J]. Journal of Computer Research and Development, 2002,39(8):986-991. (in Chinese)

[68] 林鸿飞，王剑峰. 基于合作模式的文本过滤模型[J]. 小型微型计算机系统，2001，22（11）：1372－1374.

Lin Hong-fei, Wang Jian-feng. Model for Text Collaborative Filtering[J]. Mini-micro Systems, 2001, 22(11):1372-1374. (in Chinese)

[69] Wooju Kim, Yong U. Song, June S.Hong. Web enabled expert systems using hyperlinks-based inference[J]. Expert Systems with Applications, 2005(28):79-91.

[70] 冯翱，刘斌，卢增祥等. Open Bookmark——基于Agent的信息过滤系统[J]. 清华大学学报（自然科学版），2001，41（3）：85－88.

Feng Ao, Liu Bin, Lu Zeng-xiang, et al. Open Bookmark——an Agent-based information filtering system[J]. Journal of TsingHua University (Science and Technology) 2001, 41(3):85-88. (in Chinese)

[71] Glover E. J., Lawrence S., Gordon M. D., et al. Web Search——Your Way[J]. Communications of the ACM, 2001,44(12): 97-102.

[72] 潘金贵，胡学联，李俊等. 一个个性化的信息搜集Agent的设计与实现[J]. 软件学报，2001，12（7）：1074－1079.

Pan Jin-gui, Hu Xue-lian, Li Jun, et al. Design and Implementation of a Personalized Information Retrieval Agent[J]. Journal of Software, 2001,12(7):1074-1079. (in Chinese)

[73] 徐振宁，张维明，陈文伟. 基于Ontology的智能信息检索[J]. 计算机科学，2001，28（6）：21－26.

Xu Zhen-ning, Zhang Wei-ming, Chen Wen-wei. Intelligent Information Retrieval Using Ontology[J]. Computer Science, 2001, 28(6):21-26. (in Chinese)

[74] Peter Dolog, Nicola Henze. Personalization Services for Adaptive Educational Hypermedia.[C]// Proceedings of International Workshop on Adaptivity and User Modelling in Interactive Systems (ABIS'2003).

[75] 卢增祥，关宏超，李衍达. 利用bookmark服务进行网络信息过滤[J]. 软件学报， 2000，11（4）：545-550.

Lu Zeng-xiang, Guan Hong-chao, Lu Yan-da. Network Information Filtering Using Bookmark Service[J]. Journal of Software.2001,11(4):545-550.(in Chinese)

[76] Salton G. Developments in automatic text retrieval[J]. Science, 1991,253(5023): 974-979.

[77] Stefani A., Strappavara C. Personalizing access to web sites:The SiteIF project[EB/OL].

http://wwwis.win.tue.nl/ah98/Stefani/Stefani.html.1998-06-24/2004-05-12.

[78] 卢增祥，路海明，李衍达. 网络信息过滤中的固定文章集表达方法[J]. 清华大学学报（自然科学版），1999，39（9）：118-121.

Lu Zeng-xiang, Lu Hai-ming, Lu Yan-da. FDS expressive method in information filtering[J]. Journal of TsingHua University (Science and Technology),1999,

39(9):118-121. (in Chinese)

[79] Lieberman, H. Letizia. An agent that assists web browsing. [C]// Proceedings of the International Joint Conference on Aritificial Intelligence. Morgan Kaufmann Publishers.1995.

[80] Joachims, T., Freitag, D. & Mitchell, T. WebWatcher: A Tour Guide for the World Wide Web. [C]//Proceedings of the 15th International Joint Conference on Artificial Intelligence. IJCAI-97.1997:770-775.

[81] Pazzani M.J. and Billsus D. Learning and Revising User Profiles: The identification of Interesting WebSites. Machine Learning 1997, 27(6):313-331.

[82] D. Mladenic. Personal WebWatcher: design and implementation[R]. Technical Report IJS-DP-7472, J.Stefan Institute, Department for Intelligent Systems, Ljubljana,1998.

[83] Liu, L.， Pu, C. et al. XWRAP: An XML-enable Wrapper Construction System for the Web Information Source. [C]// Proceedings of the 16th IEEE International Conference on Data Engineering, 2000:611-620.

[84] 王琦，唐世渭，杨冬清，王腾蛟. 基于DOM的网页主题信息自动提取[J]. 计算机研究与发展，2004，41（10）：1786-1792.

WANG Qi , TANG Shi-wei , YANG Dong-qing , and WAN G Teng-jiao. DOM-Based Automatic Extraction of Topical Information from Web Pages[J]. Journal of Computer Research and Development, 2004, 41(10):1786-1792.(in Chinese)

[85] 梁邦勇，李涓子，王克宏. 基于语义web的网页推荐模型[J]. 清华大学学报（自然科学版），2004，44（9）：1272-1281.

Lang Bangyong, Li Juan-zi, Wang Ke-hong. Web Page Recommendation Model for the Semantic Web[J]. Journal of TsingHua University (Science and Technology), 2004,44(9):1272-1281.(in Chinese)

[86] Kristina Lerman, Craig Knoblock, Steven Minton. Automatic Data Extraction from Lists and Tables in Web Sources. [C]// Proceedings of the Automatic Text Extraction and Mining workshop(ATEM-01), IJCAI-01, Seattle, WA, USA, August 2001.

[87] 崔继馨，张鹏，杨文柱. 基于DOM的web信息抽取. 河北农业大学学报，2005，28(3):90-93.

CUI Ji-xin, ZHANG Peng, YANG Wen-zhu. DOM based Web information extraction. Journal of Agricul Tural University of Hebei, 2005,28(3):90-93. (in Chinese)

[88] 孙承杰，关毅. 基于统计的网页正文信息抽取方法的研究. 中文信息学报，2004，18（5）：17-22.

SUN Cheng-jie, GUAN Yi. A Statistical Approach for Content Extraction from Web Page. Journal of Chinese Information Processing, 2004, 18(5):17-22.(in Chinese)

[89] A Finn, A Kushmerick, B. Smyth. Fact or fiction: Content classification for digital libraries. [C]//Proceedings of the 2nd DELOS Network of Excellence Workshop on Personalization and Recommender Systems in Digital Libraries, Dublin, Ireland, 2001.

[90] E. Kassinen, M. Aaltonen, J. Kolari, et al. Two approaches to bringing Internet services to WAP devices.[C]// Proceedings of the 9th International World Wide Web Conference on Computer Networks. Amsterdam: North-Holland Publishing Corporation, 2000:231-246.

[91] O. Buyukkokten, H. Garcia-Molina, A. Paepacke.Accordion Summarization for end-game browsing on PDAs and cellular phones.[C]//Proceedings of ACM Conference on Human Factors in Computering Systems(CHI 2001), New York: ACM Press,2001:213-220.

[92] O. Buyukkokten, H. Garcia-Molina, A. Paepcke. Seeing the whole in parts: Text summarization for Web browsing on handheld devices. [C]// Proceedings of the 10th International Conference on World Wide Web. New York: ACM Press, 2001:652-662.

[93] S. Gupta, G. Kaiser, D. Neistadt, et al. DOM-based content extraction of HTML documents.[C]//Proceedings of the 12th International World Wide Web Conference, New York: ACM Press, 2003:207-214.

[94] 孙承杰，关毅. 基于统计的网页正文信息抽取方法的研究[J]. 中文信息学报，

2004，18（5）：17-22.

SUN Cheng-jie, GUAN Yi. A Statistical Approach for Content Extraction from Web Page[J]. Journal of Chinese Information Processing, 2004, 18(5):17-22.(in Chinese)

[95] Levenshtein V I. Binary Codes Capable of Correcting Deletions[J]. Insertions and Reversals.Sov.Phys.Dokl., 1966:705-710.

[96] 董振东. 语义关系的表达和知识系统的建造[J]. 语言文字应用. 1998，27（3）：76-82.

Dong Zhen-dong. The Expression of the Semantic Relation and the Construction of the Knowledge System[J]. Applied Linguistics. 1998, 27(3): 76-82. (in Chinese)

[97] 刘群，李素建. 基于《知网》的词汇语义相似度计算[J]. Computational Linguistics and Chinese Language Processing，2002, 7(2):59-76.

Liu Qun, Li Su-jian. Word Similarity Computing Based on How-net[J]. Computational Linguistics and Chinese Language Processing，2002, 7(2):59-76. (in Chinese)

[98] 邓珞华. 概念空间——定义、意义和局限[J]. 情报学报. 2003，22(4)：393~397.

Deng Luo-hua. Concept Space－Its Definition, Significance and Limitation[J]. Journal of The China Society For Scientific and Technical Information. 2003, 22(4):393-397. (in Chinese)

[99] 孙强，李建华，李生红，许欢庆. 基于概念联想网络的网页预取模型[J]. 上海交通大学学报，2004，38(5):779-782.

Sun Qiang, Li Jian-hua, Li Sheng-hong, Xu Huan-qing. Web Pre-fetching Model Based on Concept Association Network[J]. Journal of ShangHai JiaoTong University. 2004,38(5):779-782.(in Chinese)

[100] Catledge, L.D. and Pitkow, J.E. Characterizing browsing strategies in the World-Wide Web[J]. Computer Networks and ISDN Systems, 1995:1065-1073.

[101] Nick Golovin, Erhard Rahm. Reinforcemen learning architecture for web recommendations. [C]// Proceedings of the International Conference on Information Technology: Coding and Computing.2004:398-407.

[102] R.Burke. Hybrid Recommender System: Survey and Experiments[J]. User

Modeling and User-Adapted Interaction, 2002(12):331-370.

[103] M.Nakagawa, B. Mobasher. A Hybrid Web Personalization Model Based on Site Connectivity. [C]// Proceedings of the 5th WEBKDD workshop, Washington, DC, USA, 2003:59-70.

[104] Hsiangchu Lai, Tzyy-Ching Yang. A System architecture for intelligent browsing on the web[J]. Decision Support Systems, 2000,28(3):219-239.

[105] Dempster, A.P., Laird, N. M. and Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of Royal Statistical Society, 1977(39):1-38.

[106] Hand, D., Mannila, H. and Smyth, P. Principles of Data Mining[M]. The MIT Press, USA,2001.

[107] Bartoszynski, R., and Niewiadomska-Bugaj, M. Probability and Statistical Inference[M]. John Wiley&Sons, Incorporation, 1996.

[108] Mobasher B., Dai H., Luo T. and Nakagawa M. Discovery of aggregate usage profiles for Web personalization.[C]// Proceedings of the International WEBKDD Workshop, Web Mining for E-Commerce: Challenges and Opportunities, Boston, MA, USA, 2000, August 20.

[109] DeGroot M.H., Schervich M.J. Probability and Statistics[M]. Addison-Wesley Publishing, 2002.

[110] R.Baeza-Yates, B. Ribeiro-Neto. Modern Information Retrieval[M]. Addison-Wesley, 1999.

[111] Zou Tao, Wang Ji-cheng, Zhang Fu-yan, et al. The survey of text information retrieval[J]. Computer Science, 1999, 26 (9):72-75.

[112] Joon Ho Lee. On the evaluation of Boolean operators in the extended Boolean retrieval framework.[C]// Proceedings of the 17th annual international ACM-SIGER conference on Research and development in information retrieval, 1994:182-190.

[113] Jeffrey Dean and Monika R. Henzinger. Finding Related Web Pages in the World Wide Web.[C]// Proceeding of the 8th International World Wide Web Conference. Toronto, Canada.1999:389- 401.

[114] Tsutomu Hirao. A Study on Generic and User-Focused Automatic Summarization[D]. Doctor's Thesis.2002.

[115] Stephen S. Yau, Huan Liu, Dazhi Huang and Yisheng Yao. Situation-aware Personalized Information Retrieval for mobile Internet.[C]// Proceedings of the 27th Annual International Computer Software and Applications Conference(COMPSAC 03), 2003:639-644.

[116] Fang Liu, Clement Yu, WeiyiMeng. Personalized Web Search for Improving Retrieval Effectiveness[J]. IEEE Transactions on Knowledge and Data Engineering., 2004,16(1):28-40.

[117] Salton, G. and Buckley, C. Term-Weighting Approaches in Automatic Text Retrieval[J]. Information Processing and Management,1988:24(5), 513-523.

[118] Han Li-xin, Yang Xue-lin, Xie Li and Chen Dao-xu. A New Method to Improve WEB-IR Precision[J]. Journal of the China Society for Scientific and Technical Information. 2002, 21(5):524-531.

[119] 王继成，萧嵘，孙正兴，张福炎. Web信息检索研究进展[J]. 计算机研究与发展，2001，38（2）：187-193.

Wang Ji-cheng, Xiao Rong, Sun Zheng-xin, Zhang Fu-yan. State of the Art of Information Retrieval on the Web[J]. Journal of Computer Research and Development, 2001,38(2):187-193.(in Chinese)

[120] 白首晏．大学图书馆个性化信息服务研究[J] ．沈阳建筑大学学报：社会科学版，2005，7（1）：68-70．

Bai Shou-yan. The study of individuation information service in university library[J]. Journal of ShenYang Jian Zhu Uninversity (Social Science), 2005,7(1):68-70.(in Chinese)

[121] Jamie C, Alan S. Personalisation and recommender systems in digital libraries[EB/OL]. Joint NSF-EU DELOS Working Group Report, [2003].http://dlib.cs.odu.edu/ htm.

[122] 陆广能. 数字图书馆个性化信息检索中信息推送技术的应用研究[J]. 电脑知识与技术. 2005，7：9-12.

Lu Guang-neng. Research on the Application of the information pushing

technology in digital library personalized information retrieval[J]. Computer Knowledge and Technology, 2005,7:9-12.(in Chinese)

[123] Danny S. Making an RSSFeed [EB/OL].[2004] . http://searchenginewatch.com/ sereport/ article.php/2175271.htm.

[124] RSS 2.0 Specification [EB/OL].[2005]. http://blogs.law.harvard.edu/tech/rss.htm.

[125] Leonardo C, Donatella C, Pasquale P. A service for supporting virtual views of large heterogeneous digital libraries.[C]// Proceedings of 7th European Digital Library Conference. Norway, 2003: 362 -373.

攻读博士期间发表的论文

1. 赵欣欣，索红光，刘玉树，张利萍.基于带权语义距离的网页预取方法.北京理工大学学报（EI 刊源）.2006,26(4).
2. 赵欣欣，郑志蕴，刘玉树.基于 RSS 的个性化推送服务.沈阳建筑工业大学学报（EI 刊源）.2006，22（2）：334－337.
3. Zhao Xin-xin, Suo Hong-guang, Liu Yu-shu. A Web page Prefetching Method based on User Interest Forest Model. Journal of Computational Information Systems（EI 刊源）.（已录用）
4. Zhao Xin-xin, Suo Hong-guang, Liu Yu-shu. A Web Page Content Information Extraction Method based on Tag Window. International Confernece Machine Learning and Cybernetics 2006.（EI 检索源）（已录用）
5. 赵欣欣，朱铁丹，刘玉树.不同粒度下的文档分类.计算机工程（已录用）
6. 赵欣欣，刘玉树，高影繁，王丽.web 缓存、预取、推荐.计算机科学.2005,32(9A):157-159.
7. 赵欣欣，刘玉树，高影繁.一种结合用户兴趣的推荐算法.计算机工程与应用.2005,41:160-162.
8. 赵欣欣，索红光，刘玉树.基于改进汉宁窗的信息检索模型.广西师范大学学报.（已录用）
9. 赵欣欣，索红光，刘玉树.基于标记窗的网页正文信息提取方法.计算机应用研究（已录用）
10. 赵欣欣，索红光，刘玉树.基于用户兴趣森林模型的网页预取方法.中文信息学报.(已投稿)
11. Tie-dan Zhu, Xin-xin Zhao, Yu-shu Liu. A New Text Classification Model Based on the Sentence Space. International Confernece Machine Learning and Cybernetics 2005:1774-1777.（EI 收录）
12. Su Fei, Ci Lin-lin, Zhu Li-ping, Zhao Xin-xin. Semantic Data caching for XQuery. Journal of Beijing Institute of Technology (English Edition). (Ei 刊源)(已录用)

致　谢

首先，真诚地感谢我的导师刘玉树教授。在相处的几年里，刘老师严谨的治学态度，渊博的学识，对问题敏锐的洞察力，真诚和蔼的为人，平易近人的学者风范，使我受益匪浅，为我今后的研究工作树立了榜样。更重要的是，在刘老师身上，流露出一种可贵的对待工作的敬业精神，这是让我终身受益的财富。回想几年来走过的路，虽无大的波折，但也经历了无数的困惑和迷茫。每次同刘老师的交谈都能增加我的信心，刘老师的每次指导都能使我茅塞顿开。我就是在这样的一次次交流中逐渐成长和进步。从论文的开题到研究过程中方向的把握，刘老师都提出了非常宝贵的意见和建议。刘老师不仅在学术上严格负责，在生活上更是无微不至的关怀学生。在此，谨向刘老师表达我由衷的感谢和诚挚的敬意。

感谢李侃副教授在论文完成过程中给予的中肯建议，在此表示我由衷的感谢。

感谢师兄索红光副教授在课题研究过程中、在论文构思上给予的帮助和启迪。

感谢我的大学老师邵永运副教授在我遇到数学问题时给予的耐心细致指导，感谢他在我完成论文的过程中给予的关心、支持和帮助，在此对他说声谢谢。

感谢苏斐、付慧、杜剑侠、周艺华、朱铁丹同学在生活和学习上给予的帮助，感谢 2002 级秋季博士生班级的全体同学，和你们一起学习和相处的日子充满了欢乐和温馨，为我整个的学习过程增添了无限的乐趣。

感谢 904 实验室的所有老师，所有师兄师姐、师弟师妹，感谢你们与我一起共享欢乐、分担忧愁。

感谢我的爸爸和妈妈，在我的学海生涯中，他们一直在生活上无私的关心我、爱护我，在学业上支持我、鼓励我，让我在求学的道路上充满信心，勇往直前。在课题研究过程中，他们的鼓励和支持是让我走完这艰辛的求学过程的重要精神支柱，尤其是在课题遇到难点时，爸爸妈妈总能耐心倾听，不断激励我前进。他们的爱，我将永远铭记在心。

最后，感谢所有关心和帮助过我的老师和朋友们，感谢参与本文评审的各位老师在百忙之中对我论文的悉心指正。