



2016 湖南省研究生数学建模竞赛参赛承诺书

本文仔细阅读了湖南省研究生数学建模竞赛的竞赛规则。

本文完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与队外的任何人（包括指导教师）研究、讨论与赛题有关的问题。

本文知道，抄袭别人的成果是违反竞赛规则的，如果引用别人的成果或其他公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

本文郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，本文将受到严肃处理。

本文授权湖南省研究生数学建模竞赛组委会，可将本文的论文以任何形式进行公开展示（包括进行网上公示，在书籍、期刊和其他媒体进行正式或非正式发表等）。

本文参赛选择的题号是：B 题

本文的参赛报名号为：201518001012

所属学校：国防科学技术大学

参赛队员（打印并签名）：1.杨勇

2.张胜

3.李济廷

指导教师或指导教师组负责人(打印并签名)：

日期：2016 年 4 月 19 日

评阅编号（由组委会评阅前进行编号）：

2016 湖南省研究生数学建模竞赛

编号专用页

评阅编号（由组委会评阅前进行编号）：

评阅记录（可供评阅时使用）：

[illegible]

湖南省第二届研究生数学建模竞赛

题 目 B 题 法律咨询网站用户行为分析及新服务模式探索

摘 要

随着互联网的日益发展，新的服务模式在各行各业应运而生。“互联网+”法律也是近些年来迅速发展的一个新型服务模式。人们通过网络进行法律咨询、法律知识查找、律师查找、法律文书咨询等。为了更好地使用户知法、懂法、守法，提升法律服务网站的用户体验，本文针对法律快车网站的用户日志文件，分析并研究以下问题。

在问题一中，针对用户日志文件，本文进行了单维度和双维度的数据分析。单维度分析统计了用户 IP、用户所在地区、用户使用系统、用户浏览的页面信息、用户浏览的主题信息等。双维度分析比较了不同地区用户对于不同主题的偏好信息，结果显示地区之间用户对主题的关注方向是有差别的。

为了计算用户查找信息或寻找律师的成功率和效率，本文分析了关键词匹配算法的缺陷，分析了用户查询信息的流程，定义了四种页面类型：有效页面、无效页面、终止页面和过渡页面。并通过这四个概念计算出特定类别问题（离婚类别问题）的成功率和效率分别为 12.6%和 56.8%；而其他类型问题的成功率 22.4%和效率 93.9%。

在问题二中，为了提升该网站的成功率和效率，本文从三个方面提出了优化和改进现有服务的建议和措施。

其一，构建智能推荐系统。本文比较了基于协同过滤的推荐算法，但考虑到网页之间存在主题的内在联系，最后选择基于隐语义模型（LFM）的方法做为推荐算法。该模型将用户和网页的关系分解成用户与隐含主题和网页与隐含主题的关系。针对该模型本文进行了离线测试，发现模型稳定性较好。特别地，当隐含主题个数接近真实主题个数时效果最好。结合问题一，验证了引入智能推荐系统确实使用户找到相关页面或专业律师的成功率和效率有所提高；

其二，优化页面结构。本文分析了用户的不同类型，针对新用户设计了推荐系统冷启动的方式。通过挖掘新老用户的特征以及相似特征用户的行为偏好，定义重要度的概念。并且根据重要度来进行页面结构优化设计，重要度越高的信息，越优先显示。但同时兼顾热度低的领域，满足长尾用户的需求；

其三，构建智能提示系统。本文使用 Levenshtein 距离和动态规划的算法，基于已知关键词库对用户输入进行预测和提示，以提升用户的输入效率和用户体验。为探究智能系统的提升效率，本文进行离线测试。结果显示智能提示系统能够大幅度提升用户查找效率，且提示备选列表长度为 5~10 时，提升效率和用户体验较好。

关键词：推荐系统；隐语义模型；页面结构优化；智能提示

1 问题重述与分析

随着中国国家实力的不断增强，人们的法律意识也随之提高，对于法律服务的需求也不断提升，而目前中国法律服务行业整体水平偏低，主要体现在分工不细致、律师事务所规模偏小，律师占总人口的比例偏低等。目前全国执业律师约 27 万，北京万人律师比最高，约为 12.6，多个省份万人律师比不足 1。限制了法律服务的范围和质量。随着“互联网+”这一概念的提出，法律服务也进入了新时代，法律咨询网站应运而生。利用互联网自身的特点，普及法律常识，增强公众法律意识，提升“普法”的广度与深度，降低法律咨询的难度和成本，缩小地区差异，帮助用户，特别是经济不发达地区的用户，通过法律途径伸张公平与正义。互联网正逐步成为人们获取法律服务的重要途径。然而，由于法律事务的复杂性、多样性、专业性，目前的法律咨询网站还有很多的不足之处，用户很多时候还是难以通过网站高效准确地找到有针对性的法律知识、相关案例或专业律师，影响用户体验，限制网站发挥作用。为了优化法律咨询网站的服务模式和用户体验，在本文中本文主要对以下问题进行讨论：

问题一：根据附件数据，分析用户关注领域、访问路径、用户体验等方面的特征或规律。尝试评估针对特定问题找到相关页面或专业律师的成功率和效率。

分析：首先本文需要分析题目所附带的日志数据，对用户的行为进行观察和统计，用图表的形式来描述表示用户的行为特征和规律。

而后，针对特定的问题，本文需要结合特定问题的数据，对成功率和效率两个概念给出精确的定义，进而求解，在得到成功率和效率的结果后还要做进一步的结果分析，分析结果背后可能存在的原因。

问题二：尝试通过优化页面组织结构，增加新功能（如，自动推荐），建立新型服务模式（如，向导式助手）等方式，改进服务效率，并对改进效果进行对比、评估。改进方案原则上应为自动服务，如果确实需要人工干预，则应明确人工干预的具体方式，并估计工作量。

分析：在问题二中，本文从三个角度来改进网站服务质量，提升用户体验。三个角度分别为构造智能推荐系统、优化页面结构以及构建智能提示系统。在构造智能推荐系统这个角度，需要了解有关推荐系统的知识，借鉴亚马逊等购物网站成功的推荐方法；在优化页面结构方面，需要定义与法律相关的各个知识域的重要程度，从而对页面的显示做出优化；在构建智能提示系统这个角度，需要了解和析主流搜索引擎所使用的智能输入提醒功能，从而找出解决方法。

2 假设说明

1. 当网页系统记录的停留时间为 0 时，则假定页面对用户是无效信息；
2. 假设用户感兴趣的信息主要集中在以“HTML”关键词结束的网页中，其他的网页可以理解为目录页面或者路径页面（非信息类页面）；
3. 假设日志数据均是由真实用户产生，不考虑爬虫等其他因素影响；

3 符号说明

符号	说明
$P_{xi} = \{P_{x1}, P_{x2}, \dots, P_{xN}\}$	有效页面集合
$T_{xi} = \{T_{x1}, T_{x2}, \dots, T_{xN}\}$	每个有效页面停留的时间集合
$P_{yi} = \{P_{y1}, P_{y2}, \dots, P_{yN}\}$	无效页面集合
$P_{zi} = \{P_{z1}, P_{z2}, \dots, P_{zN}\}$	过渡页面集合
$T_{zi} = \{T_{z1}, T_{z2}, \dots, T_{zN}\}$	每个过渡页面停留的时间集合
φ	成功率
ω	效率
k	隐含主题的个数
J	损失函数
α	学习速率
Precision	精度
Recall	召回率
Coverage	覆盖率
Important	重要度
R_u	为目标用户提供的推荐列表
T_u	目标用户的预测行为列表
$\text{lev}_{a,b}(i, j)$	编辑距离
IMP	输入提升效率

4 模型的建立、求解与应用

4.1 问题一

在问题一中，首先要求分析用户关注领域、访问路径、用户体验等方面的特征或规律，而后尝试评估针对特定问题找到相关页面或专业律师的成功率和效率。在分析用户数据方面，首先采取单维度的数据分析，即对单个有价值的属性进行分析总结；而后进行多维度的分析，即从众多属性中找出可能会有联系有价值的组合进行分析。而在成功率和效率的定义和评价方面，本节也给出了详细的过程，并对所得结果进行了分析。

4.1.1 基于日志数据的用户行为分析

本小节中对题目中提供的数据进行了处理和分析，处理的原始数据为 log.csv 文件，数据文件包括 837450 条数据，描述了与法律网站和用户行为相关的 21 条属性，包括真实用户 IP，地区编号，浏览器代理，用户浏览器类型，页面路径，页面类型等。

首先，本小节将先做单维度的属性分析，从 21 个属性中选取以下几个有价值的属性进行描述和分析：

■ 数据的时间范围

原始数据为 2015 年 2 月 1 日 00:00:06 到 2015 年 4 月 29 日 17:11:03 产生的用户浏览网站的数据。

■ 真实用户 IP

IP 地址即互联网地址或 Internet 地址。是用来唯一标识互联网上计算机的逻辑地址。每台连网计算机都依靠 IP 地址来标识，全世界的 IP 地址都是唯一的，这也是标识用户最有效的方法。通过对数据的分析，本文发现原始数据中不重复的真实 IP 有 230149 条，表明数据中包括 230149 为用户的浏览记录。对所有用户浏览网站的次数以及频数进行分析，得到下图：

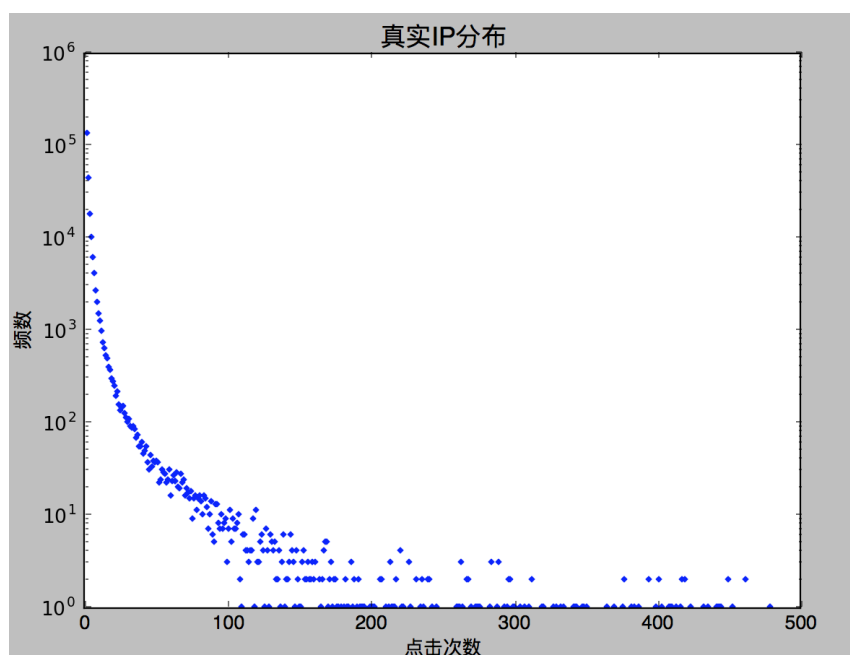


图 1 真实用户 IP 长尾效应图

通过上图可以发现，在该网站的用户行为方面服从幂律分布，当点击次数较少时，其频数较高，而随着次数的增多，频数迅速下降。

长尾效应是用来描述诸如亚马逊、Netflix、Rhapsody 之类网站的商业和经济模式，是指那些数量巨大种类繁多的产品或者服务，其中很大一部分得不到足够重视，但是零零散散的冷门产品或服务，总收益却超越了主流商品。因此在本问题中对于部分用户的零散的需求以及针对这些需求的个性化服务设计是值得注意的一个方面。

■ 地区

在用户所处地区数据方面，做了如下的统计：

表 1 各个地区记录数统计表

地区编号	记录数
140100	760164
140107	21436
140103	18814
140108	7479
140106	7384
140104	6724
140105	6425
140102	4241
140109	1794
140110	1507
140101	743
140111	618
140112	121

从上表中可以看出，各个区域对于该网站的关注程度是有所区别的，则在各个区域的网站运营中应该注意选取不同的运营方式。

■ 操作系统

用户浏览网页所使用的操作系统统计如下图所示：

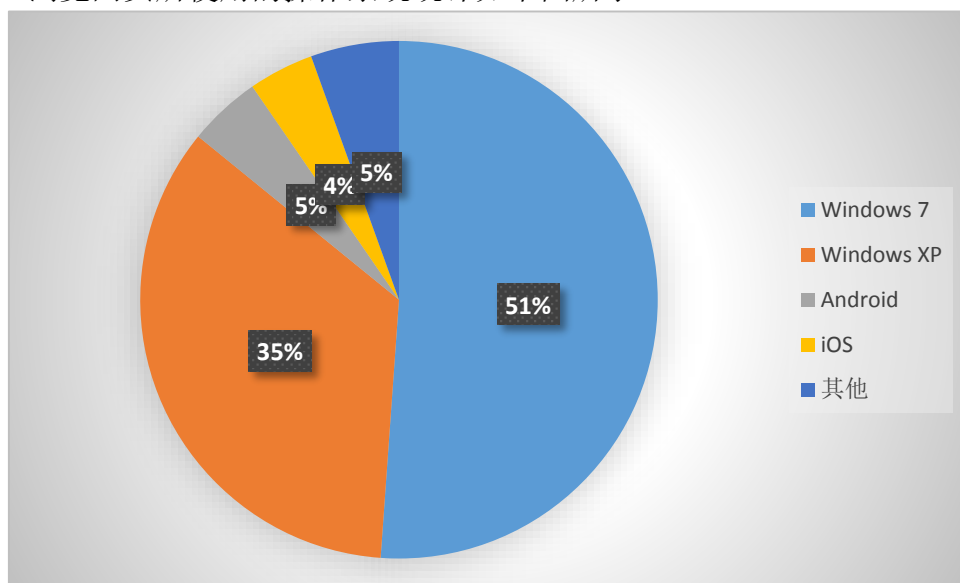


图 2 操作系统统计饼图

通过统计可以发现，用户在浏览网页时使用 Windows 系列系统的高达 86%，同时也发现有些用户会在移动端浏览网页。所以针对移动互联网的发展，网站需要注意针对移动端进行特别的网页设计

■ 页面分类信息

在原始数据 log.csv 中包含一列属性为页面类型编号，结合附录 3 中的信息，可以得到页面类型编号与具体页面分类信息的一一对应。经过统计，可得到下图：

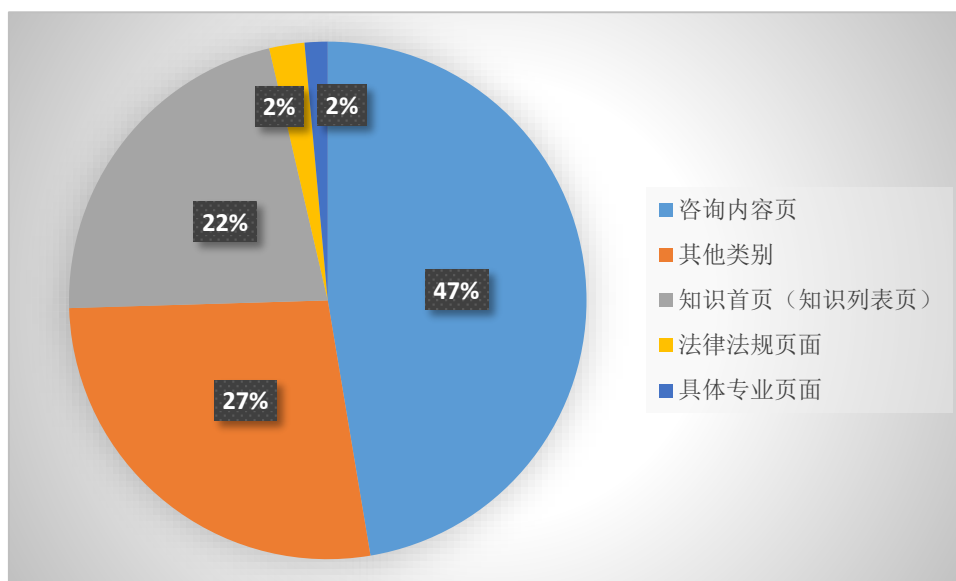


图3 页面分类信息统计饼图

在上图中可以发现，用户最经常浏览的页面类型为咨询内容页和知识首页。

■ 访问入口

访问入口为用户通过何种方式到达本题中的目标网站，通过统计可以得到下图：

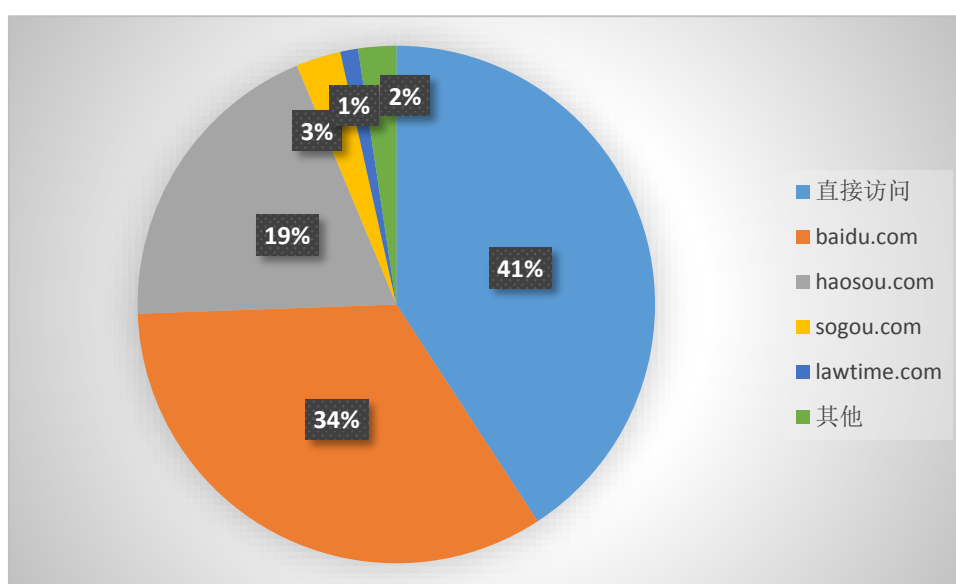


图4 访问入口信息统计饼图

在上图中可以发现用户浏览目标网站最常用的是输入网址直接到达，使用搜索引擎到达目标网站的也占了很大一部分比例，其中使用最多的搜索引擎是百度和好搜。

■ 标题关键词

标题关键词是用户在网站询问法律问题时在标题中所提炼出的关键部分，表现了用户在咨询这个问题时的主要内容，对所有网站用户在该网站咨询的问题情况进行了一般性描述。原始数据中包含 2115 个关键词，其中搜索次数大于 10000 条的关键词是本文主要关注的对象。通过统计这部分的信息可以了解该网站用户最感兴趣的话题是什么，最需要咨询的问题是什么，对标题关键词的统计如下图所示：

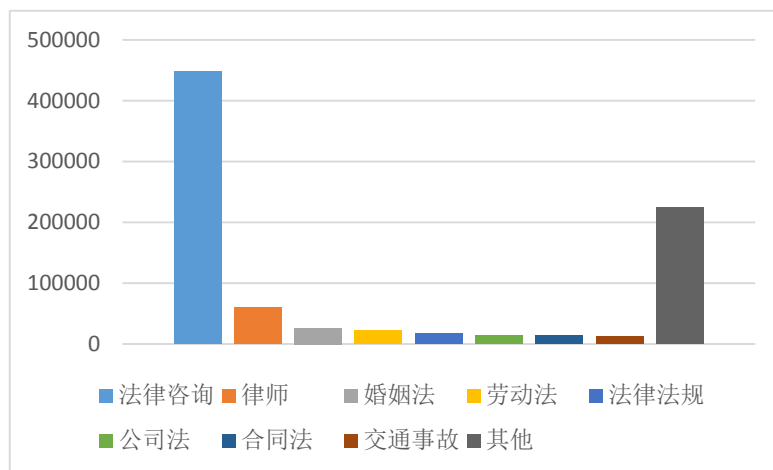


图 5 标题关键词信息统计柱形图

通过上图可以发现，用户最常搜索的关键词为法律咨询、律师、婚姻法等。

■ 标题类型

标题类型指的是用户所提出的咨询问题所属的领域，对于这一属性的统计分析可以得知用户最为关注的问题领域有哪些，这可以为之后的页面结构优化提供有效依据，原始数据中包含的标题类型有 72 个，对这些标题类型的统计结果如下图所示：

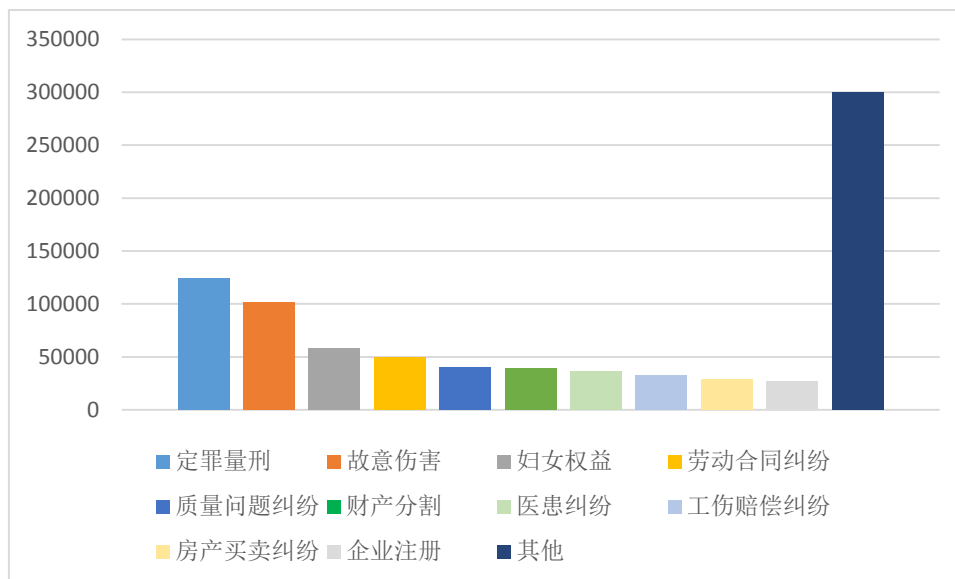


图 6 标题关键词信息统计柱形图

上图展示的用户最关注的前十个与法律相关的领域，在其他中，股权纠纷、劳资纠纷、侵权纠纷、五险一金、立案侦查等也是用户关注很多的法律领域。

接下来本小节将做双维度的属性分析，本文在单维度分析中发现在不同的地区，用户关注的问题会有所不同，所以选取了地区编号和标题类型进行分析：

■ 地区和标题类型

在不同的地区中，用户关注的领域也不同，本文在数据中所涵盖的 13 个区域中选取 3 个有代表性的区域，分别是 140103 地区、140105 地区和 140106 地区进行分析，并得到统计柱形图如下图所示。在图中，不同的颜色表示不同的标题类型，在不同的图中，同一种标题类型的颜色是一致的，本文着重关注每个地区排名前五名的标题类型：

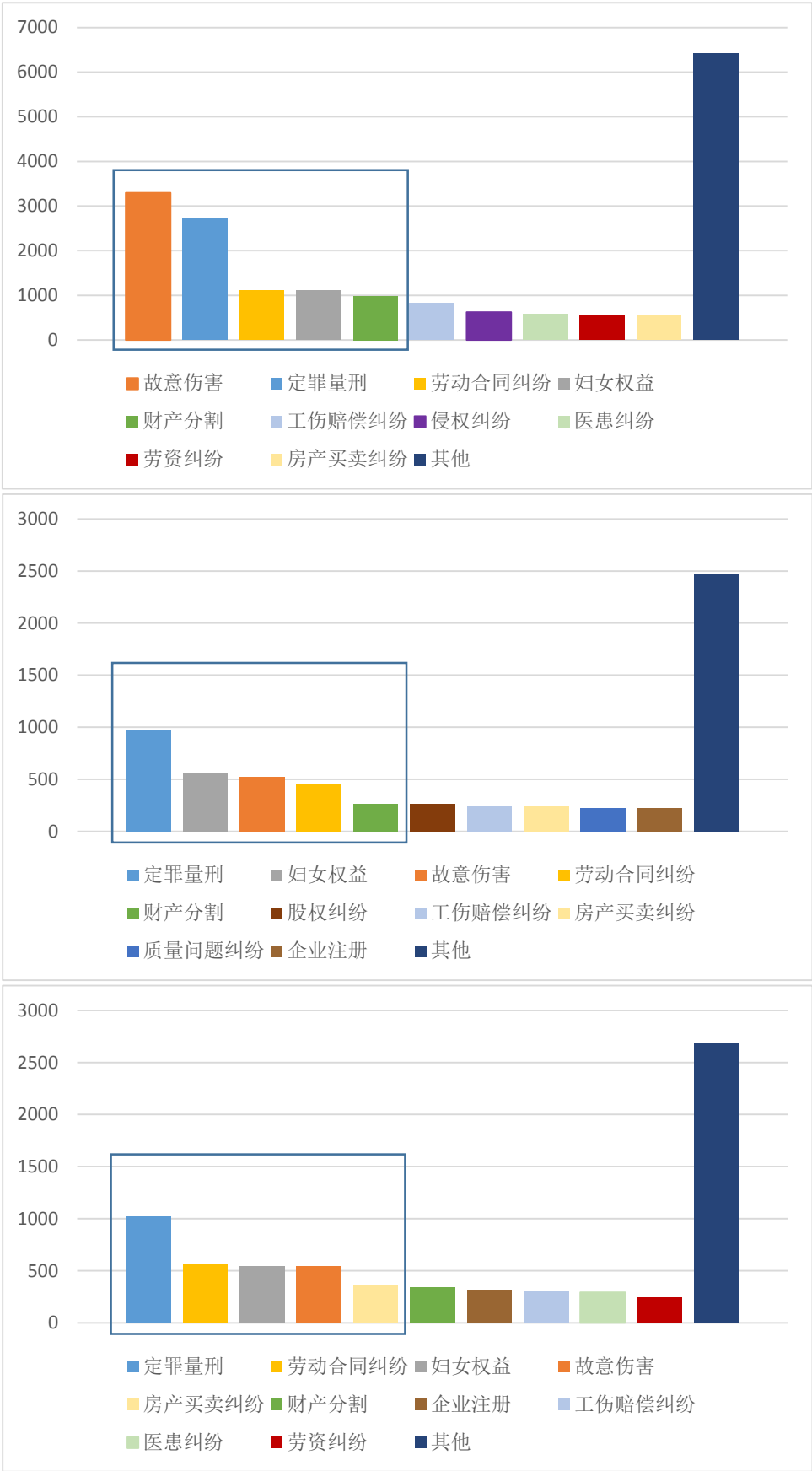


图 7 编号 140103/140105/140106 地区标题关键词信息统计柱形图

从图中可以发现，不同区域的用户关注的法律问题是有一些区别的。总的来说，用户们最关注的几个领域是定罪量刑、妇女权益、故意伤害、劳动合同纠纷等。但每个地区内用户们对这几个问题的关注程度是有细微差别的，如在 140103 地区，用户最关心的是故意伤害方面的法律知识，在 140105 和 140106 区域，除了定罪量刑这个方面，用户们分别对妇女权益和劳动合同纠纷方面也有很高的关注度。

4.1.2 模型的建立

问题一中所提到的要求是评估针对特定问题找到相关界面或者专业律师的成功率和效率。因此，本节选取热门搜索内容——离婚类别作为建模分析对象。

本文认为，一次准确的浏览行为即指用户在发生浏览行为后能更好的得到与用户自身意图匹配最好的结果（相关知识页面或专业律师），因此，如何定义用户的搜索意图和定义搜索结果是定义准确率的关键。

通过分析数据 log.csv，发现用户在搜索时使用的关键词能很好的反映用户的实际意图，而最终的停留页面即是最终的搜索结果。因此，一种衡量成功率的方法是对用户搜索的关键词（组）和浏览网页的标题进行匹配。如果得到的匹配度越高，说明用户找到了自己需要的网页和信息。而匹配度越低，则说明用户没有找到自己需要的信息。

在进行匹配之前，需要对用户 i 的关键词（组）合网页的标题进行分词。假设关键词组的分词结果为：

$$K_i = \{k_1, k_2, \dots, k_m\}$$

网页标题 j 的分词结果为：

$$T_j = \{t_1, t_2, \dots, t_n\}$$

则可以用 Jaccard 相似度定义两者的匹配度，其公式如下：

$$J(K_i, T_j) = \frac{|K \cap T|}{|K \cup T|}$$

其中 $|\dots|$ 表示集合中所含元素的个数。

那么，对于整个网站系统的准确率 $J(K, T)$ ，即为上述相似度的平均值，即：

$$J(K, T) = \text{avg} \left(J(K_i, T_j) \right) \quad i = 1, 2, \dots, m; j = 1, 2, \dots, n$$

通过分析数据本文发现，用户的搜索关键词记录主要来自于百度、搜狗等搜索引擎，而好搜等搜索引擎并未给出用户搜索的关键词。因此，计算这些搜索引擎的关键词搜索匹配度比较困难。另外，还有一些直接访问到 lawtime.com 的日志数据，用户通过直接点击网页模块，缺乏搜索关键词的信息，因此也无法通过上述公式计算准确率。因此，该方法不适用于题目所给出的数据。

此外，本文发现百度等搜索引擎所带来的流量中，用户搜索的关键词和文章的标题匹配度很高。但本文认为这很可能是因为百度等搜索引擎本身良好的性能会将匹配度高的网页排在搜索结果的前面，而用户往往只点击结果列表最前面若干条的链接，因此，所得到的匹配度结果普遍偏高。所以，使用用户搜索关键词和网页标题的匹配程度来衡量搜索的成功率不具有参考性。

为了更好的定义用户的意图，本文提出另外一种定义成功率与效率的方法，首先明确用户浏览网站找到自己所需要的相关界面或者专业律师的整体流程，流程如下图所示：

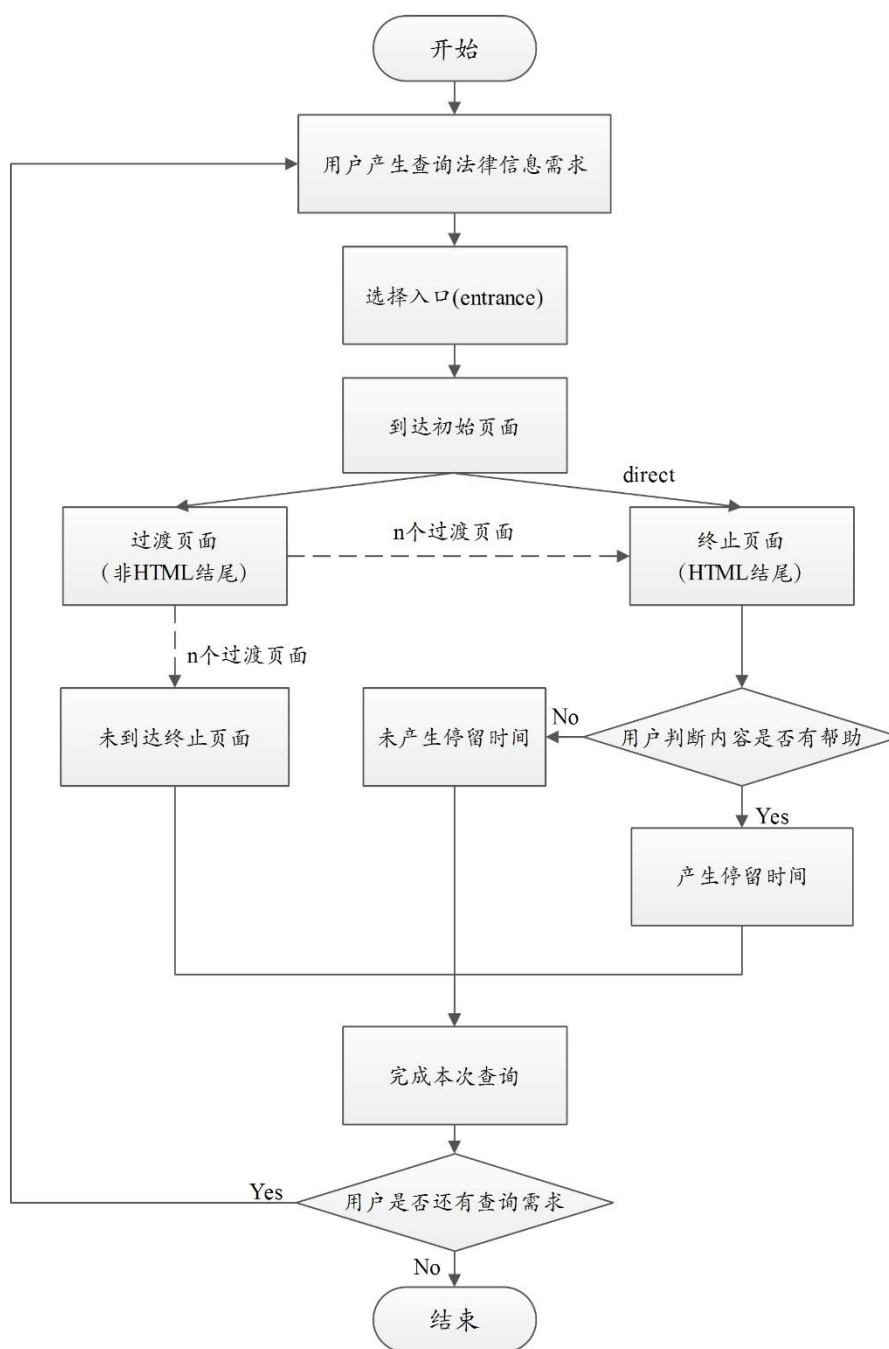


图 8 用户查询信息流程图

而后定义四个概念如下：

- **有效页面：**即在数据中用户浏览过的以“HTML”关键词结束的，且用户在该页面有停留时间的网页。该网页中集中了用户感兴趣的信息。下表是一个有效界面的实例：

表 2 有效页面实例

网页停留时间	URL	标题
262468	http://www.lawtime.cn/ask/question_5721510.html	营业执照下来多久外地小孩子才能在北京上学

定义 x_N 个有效页面集合为 $P_{xi} = \{P_{x1}, P_{x2}, \dots, P_{xN}\}$ 。在每个有效页面停留的时间集合为 $T_{xi} = \{T_{x1}, T_{x2}, \dots, T_{xN}\}$ 。

- **无效页面：**即在数据中用户浏览过的以“HTML”关键词结束的，但是用户在

该界面的停留时间为 0 的网页。这里模型假定用户在发现页面中的内容是对自己无效的信息或者自己之前的提问没有得到回答时立刻关掉网页，故网页系统记录的停留时间为 0。下表是一个无效页面的实例：

表 3 无效页面实例

网页停留时间	URL	标题
0	http://www.lawtime.cn/ask/question_6462095.html	外地结婚证能在北京办离婚吗

定义 y_N 个无效页面集合为 $P_{yi} = \{P_{y1}, P_{y2}, \dots, P_{yN}\}$ 。在每个无效页面停留的时间为 0。

值得注意的是，有效页面和无效页面统称为**终止页面**。

- **过渡页面**：即在数据中用户浏览过的所有非“HTML”关键词结束的网页，在这些网页中，有的网页有停留时间，有的网页没有停留时间，它们都是用户由初始页面到达终止页面过程中浏览的非信息类页面。下表是一个过渡页面的实例：

表 4 过渡页面实例

网页停留时间	URL	标题
0	http://www.lawtime.cn/info/hunyin/ccfglaw/	离婚财产分割 - 法律快车婚姻法

定义 z_N 个过渡页面集合为 $P_{zi} = \{P_{z1}, P_{z2}, \dots, P_{zN}\}$ 。在每个过渡页面停留的时间集合为 $T_{zi} = \{T_{z1}, T_{z2}, \dots, T_{zN}\}$ 。

定义这四个概念后，本文对用户找到相关界面或专业律师的成功率定义为：

$$\varphi = \frac{x_N}{x_N + y_N}$$

上式中的 x_N 代表有效页面的个数， y_N 代表无效页面的个数，对上式的解释为：用户经过若干个过渡界面，最终到达以 HTML 为结尾的终止页面，这些终止页面，正如上文所定义，有一些是有效页面，有一些是无效页面，用所有有效页面计数与终止页面总数的比值即为本文定义的成功率。

对用户针对离婚问题找到相关界面的效率定义为：

$$\omega = \frac{x_N + y_N}{x_N + y_N + z_N}$$

上式中的 x_N 代表有效页面的个数， y_N 代表无效页面的个数， z_N 代表过渡页面的个数，下图用结点和边表示了用户浏览网页的简易过程，其中蓝色结点为开始结点，绿色结点为有效页面结点，红色结点为无效页面结点，黑色结点为过渡页面结点，其中 $a_i(i = 1,2,3)$, $b_i(i = 1,2)$, $c_i(i = 1,2)$, d 边为有效边，代表用户由开始页面直接或者经过若干个过渡页面到达了终止页面；而图中的 e 边为无效边，代表用户到达若干个过渡页面后放弃查询。本文对于效率的定义即为终止页面的总数与所有页面总数的比值。

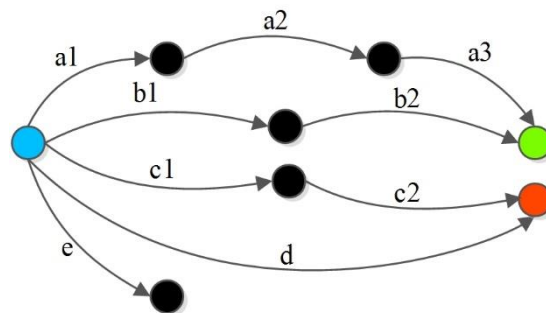


图 9 用户浏览网页简易过程图

4.1.3 模型结果及分析

在模型建立后，本节对用户查询离婚类别问题找到相关界面或者专业律师的成功率和效率进行评价，所得结果如下：

表 5 离婚类别问题数据统计

离婚类别问题数据统计				
页面总数	终止页面	有效页面	成功率	效率
293999	167029	21045	0.125996	0.568128

由结果可以看出，在离婚类别问题的数据统计中，用户有较高的访问效率，但准确率却较低。为不失一般性，本文对非离婚类别问题的数据进行了一定量的抽样分析，所得结果如下：

表 6 非离婚类别问题数据统计

非离婚类别问题数据统计				
总页面	终止页面	有效页面	成功率	效率
189938	178297	39914	0.223862	0.938712

可以看出，在非离婚类别问题的数据中用户的成功率和效率均有明显提升。

本文认为，因为离婚类别问题的数据相较其他类别问题的数据而言，数据量更大；且由于系统并未给用户信息提示或推荐，用户在面对更多的数据信息时依然近似进行了随机选择，这样在最终的页面中用户获取到的无关信息将会增多，用户访问的无信息页面也会增多，最终导致了准确率和效率的降低现象。

因此，根据用户行为数据的分析进而给用户合理的信息和推荐，对于用户在面对大量信息时提高访问的准确率和效率是非常重要的。

4.2 问题二

问题二要求通过优化页面的结构，建立新功能，建立服务模式等方式改进和提高服务的准确率和效率。提高准确率的目的是使得用户尽可能的找到有效的网页和信息，与此同时，通过使用户尽可能的快速找到所需要的网页来提高服务的效率。那么通过对原来问题的分析，本文发现原来的网络系统具有以下几个方面可以进行提升：

- 对于直接访问网站的用户来说，服务的效率较低，用户找到需要的信息耗时较长；
- 通过搜索引擎带来的流量的成功率较低，用户在进入网站后没有发现有用信息时会快速离开；

- 用户在搜索框进行搜索时，可能会因为关键词输入不当导致查询错误；
- 不同地区的用户在进行查询搜索时所看到的法律网站页面是完全一样的，不具有区域区分性等。

针对上述问题，本文提出了构建智能推荐系统、优化页面结构、构建智能提示系统三种策略来尝试解决上述问题，以此提高网站服务的成功率和效率。

4.2.1 模型的建立

4.2.1.1 构造智能推荐系统

推荐系统是根据用户的历史记录和偏好程度，为用户筛选出与之行为或兴趣相关的信息，并按照一定的优先顺序呈献给用户。推荐系统近些年来在互联网中的应用非常普遍。包括网上购物、多媒体推荐、好友推荐、内容推荐等。

本文中，个性化推荐是根据用户的兴趣特点和实际访问行为，向用户推荐感兴趣的页面或专业律师。随着网站的不断更新发展，网站所包含信息量越来越庞大，用户需要耗费越来越多的时间精力用于获取自己所需要的信息。而对于网站开发者而言，信息过载问题日益凸显造成了大量的用户流失，为了解决这些问题，提高用户体验，个性化推荐应运而生。

通过对法律快车网站整体设计的分析发现，当前该网站采用的是传统的非个性化推荐系统，即每个用户被推荐的物品或信息均相同。可以看出，网站开发者人工挑选了推荐条目或是基于条目的流行度将浏览热度靠前的条目进行优先推荐。

常用的推荐系统所使用的算法为协同过滤算法，协同过滤算法主要分为两种，分别是基于用户的协同过滤推荐算法和基于物品的协同过滤推荐算法。新闻类的网站经常根据基于用户的协同过滤推荐算法向用户进行新闻推荐；而亚马逊等电子商务网站目前普遍采用基于物品的协同过滤推荐算法。

■ 基于用户的协同过滤推荐算法

算法主要包含两个步骤：

- a) 找到与目标用户有相同兴趣的用户集合；
- b) 找到集合中其他用户感兴趣的而目标用户未发生行为的对象推荐给目标用户。

■ 基于物品的协同过滤推荐算法

算法主要包含两个步骤：

- a) 计算物品间的相似度；
- b) 根据物品的相似度和目标用户的历史行为对其进行个性化推荐。

协同过滤算法是通过用户之间的行为模式来进行推荐的，该算法具有很好的可解释性。但是，该算法也存在稀疏性和可扩展性差的缺点。

由于网页和网页之间存在某种主题与主题之间的联系，因此，挖掘出用户对这些主题之间的喜好程度，以及网页和主题之间的相关程度，对于更好地进行网页推荐有促进作用。基于隐语义模型的推荐技术在近些年来受到了广泛的研究和关注。

隐语义模型 LFM (Latent Factor Model) 的思想是：通过找到一些隐藏的主题，并分析用户和主题的关系，主题和页面的关系，来构造和预测用户对网页之间的偏好程度。

假设用户和网页之间的关系构成了一个 $m*n$ 的矩阵 R ，其中 m 表示用户个

数， n 表示网页个数。矩阵中 1 表示用户点击了该网页（或者称为用户对网页进行了正反馈），0 表示生成的负样本，这一部分会在下文进一步解释，其他为空。对于用户和网页之间的关系刻画如下图所示：

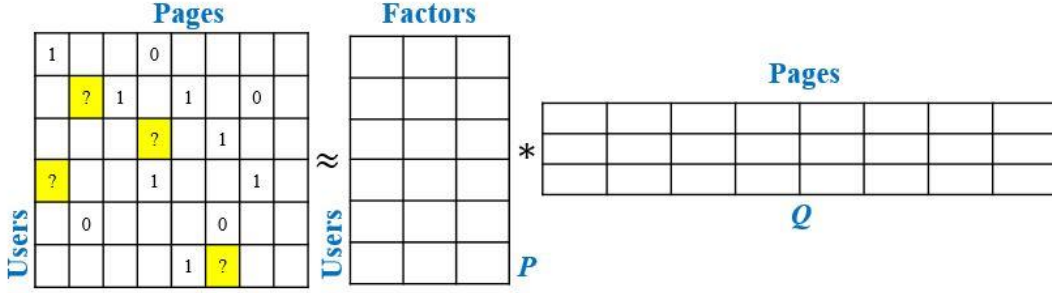


图 10 用户与网页之间关系图

如上图所示，假设矩阵 R 由矩阵 P 和矩阵 Q 相乘得到，其中 P 为 $m \times k$ 的矩阵， Q 为 $k \times n$ 的矩阵， k 表示隐含主题的个数。那么矩阵 P 表示的是用户对隐含主题的偏好程度， Q 表示主题和网页的相关程度。

LFM 的目的是找到 P 、 Q 两个矩阵，使得 P 、 Q 的乘积尽可能的与 R 相等，即：

$$R \approx P \cdot Q$$

或者

$$r_{ij} \approx \sum_k p_{ik} \cdot q_{kj}$$

则可以定义损失函数：

$$J = \frac{1}{|R|} \sum_{(i,j) \in R} (r_{ij} - p_{ik} \cdot q_{kj})^2$$

式中的 r_{ij} 表示的是矩阵 R 中已知的值。

由于计算的过程中可能出现过拟合的问题，本文在损失函数后加上正则化项：

$$J = \frac{1}{|R|} \sum_{(i,j) \in R} (r_{ij} - p_{ik} \cdot q_{kj})^2 + \lambda(|p|^2 + |q|^2)$$

算法的目的是使得该损失函数的值最小。可以通过学习的算法来对该函数进行迭代和优化，常用的算法有随机梯度下降算法。

首先求得损失函数 J 对 P 和 Q 的偏导数 $\frac{\partial J}{\partial p_{ij}}$ 和 $\frac{\partial J}{\partial q_{ij}}$ 。然后利用梯度下降的算法对梯度进行更新。得到以下公式：

$$p_{ij}' = p_{ij} + \alpha \cdot \frac{\partial J}{\partial p_{ij}} \quad q_{ij}' = q_{ij} + \alpha \cdot \frac{\partial J}{\partial q_{ij}}$$

其中 α 为学习速率。由于是学习模型，所以需要对模型进行初始化，初始时使用接近 0 随机数对 P 、 Q 矩阵进行初始化。

在构造数据集时，需要保留部分数据进行交叉验证，防止数据过拟合，并且增加模型的泛化能力。下图为数据集划分的交叉验证：

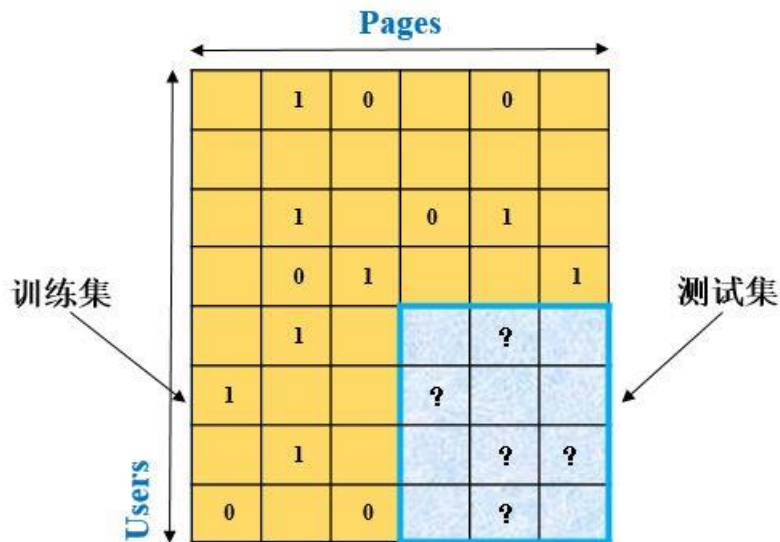


图 11 数据集划分的交叉验证

值得注意的是，LFM 算法在显性反馈数据（如评分数据，既有正反馈，也有负反馈）上能够取得很好的效果，但是在隐形反馈数据（只有正反馈，没有负反馈）上效果不佳。因此，要对数据集进行操作，需要随机采样生成负样本。生成负样本的方法很多，其中 Rong Pan 等人^[11]在通过对比了多种不同的负样本采用方法后，认为对于每一个用户，从他没有过行为的网页中采集出一些网页做为负样本，且采样时保证每个用户的正负样本数目相当，这样的方法得到的效果最好。因此，本文中使用这种方法进行负样本的采样。

综上，可以看出，LFM 算法具有以下优点：

- 基于用户行为数据统计自动聚类生成分类的角度，避免了由网页编辑不合理的自定义；
- 通过设置 LFM 最终分类数可以自动控制分类的粒度，分类数越高，粒度越细，通过模型仿真可以确定较合理的分类粒度；
- 对于每一个网页，采取软分类的方式进行分类，即认为目标网页按一定的概率属于某一个主题，而不是严格的属于某一主题；
- 对于任意一个用户， P 矩阵能表示其对所有主题的兴趣度，值越大代表用户兴趣度越高；
- 对于每一个隐含类， Q 矩阵表示网页在各主题中的权重，权重越大，越能代表隐含主题。

4.2.1.2 优化页面结构

在互联网高速发展的今天，网页能否给用户更好的浏览体验，一定程度上可以决定这个网站未来的发展，所以对于法律咨询网站来说，优化网页页面结构是尤为重要的。好的网页设计一般都有信息突出、页面简洁美观、与用户交互性好的特点，而当前的法律咨询网站的页面存在以下几个问题：

- 信息之间没有明显的区别，用户除了信息的先后性难以区分信息的重要程度；
- 网页没有相应的评分，数据均为隐形反馈；

- 虽然有设置不同的区域，但是只有对应的律师信息发生变化，其余信息没有区别，如下图所示，网站页面对于长沙和杭州两个地区而言除了律师信息发生变化，其余信息均未发生变化；



图 12 网站不同地区网页对比图

因此，本文对网页的结构做如下优化：

针对用户的类型，可将用户分为无行为记录的新用户和有行为记录的老用户。针对新用户，由于没有相应的历史数据，系统处于冷启动的状态，对于冷启动问题，系统通常情况下会根据网页的搜索热度进行推荐或者直接由网页开发者进行内容的排序，很显然，这样会导致马太效应，即随着时间的推移，热门的内容会更加热门，导致很多信息被淹没，系统的覆盖率很低，造成资源的浪费，同时，用户体验度不高，对于有“小众”需求但同时法律知识匮乏的用户，很难检索到有用信息。

通过第一问对数据的分析，本文发现，对于不同的地区、使用不同操作系统的用户所关注的内容有比较明显的不同，可见，用户基本信息与最终的结果是有

先相关性的。并且当前系统有用户注册的功能，但并未记录用户的详细信息，因此，可以建议网站采取以下措施：

- 引导用户完善自身信息，例如用户感兴趣的类目，添加用户属性标签，例如职业、家庭状况等；
- 利用现有的开放数据平台，例如新浪微博、QQ 空间等平台，可以根据用户的授权获取用户的一些基本信息，例如朋友圈、关注的生活信息等。

本文设计算法如下：计算每种特征（如年龄、性别、职业等）的用户喜欢的网页。即对于每种特征 f ，计算具有这种特征的用户对各个商品的喜好程度 $p(f,i)$ 。

$$p_0(f,i) = |N(i) \cap U(f)|$$

其中 $N(i)$ 是喜欢物品 i 的用户集合， $U(f)$ 是具有特征 f 的用户集合。

以上公式可能会出现有一个网页只被一个用户喜欢过，而这个用户刚好就有特征 f 的情况，那么令 $p(f,i)=1$ 。这种情况没有统计意义，因此可以为分母加上一个比较大的数 a ，避免这样的页面产生比较大的权重：

$$p_0(f,i) = \frac{|N(i) \cap U(f)|}{(|N(i)| + a)}$$

通过以上公式，可以计算出每种特征的网页集合。根据已有的用户信息对用户进行一定的“粗粒度”个性化推荐，这样可以避免所有的长尾信息都被覆盖，对于系统的准确率、效率、覆盖率都有一定的提升。

对于老用户而言，用户拥有一定量的浏览记录，根据浏览记录可得到目标用户对于相关网页的兴趣程度 p_1 。此外，对网页运营者来说，网页展示的信息可为其带来经济效益的能力为 C ，运营者应该在用户体验度和经济效益中取得平衡。

综合以上，本文认为系统应该对需要重点显示的内容进行突出，而不仅仅是内容的先后性，对此，本文定义推荐内容的重要度如下，重要度的取值介于 0 和 1 之间，另 $\alpha + \beta + \gamma = 1$ ：

$$Important = \alpha \cdot p_0 + \beta \cdot p_1 + \gamma C$$

根据重要度，本文对网页的版面做如下示意图：



图 13 网页版面设计基本示意图

其中,版面面积的大小与重要度成正相关关系, S 表示网页展示区域的面积, I 表示推荐内容对于目标用户的重要度,则可得到:

$$Square = S * I$$

对于该示意图做如下解释:示意图主要表现的是网页版面的扁平化设计以及明显的分区,同时用模块面积的大小来突出重点,给用户以更直接的指导。该图仅为示意图,如需日后对网站的页面进行改良设计,还需要更多的细节。

4.2.1.3 构建智能提示系统

通过对法律咨询网站——法律快车网的用户交互方面的分析,本文发现,在该网站的搜索功能方面还有一些可以改进的地方。在问题一中,本文发现,用户在到达法律快车网的入口方面,除了直接在地址栏中输入网站的网址到达网站,还有很多用户是通过搜索引擎到达网站的,用户在不知道这个网站的前提下也可以搜索到这个网站,这多归功于搜索引擎的智能提示功能。现代的搜索引擎,如谷歌和百度,都能够在用户输入的同时对用户的输入进行预测和补全,这样的智能提示功能不仅能够提高用户搜索的效率,还能够提升用户的体验以及对网站的认同。

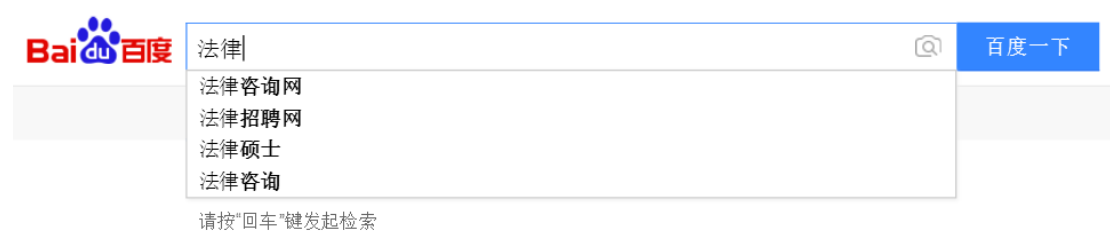


图 14 百度网站智能提示功能

综上,本文想要在用户搜索这个方面做进一步的优化和提升。

常用的自动补全的算法有构建字典树和计算编辑距离,一般来说,构建字典树是将单词拆解成单个字母,然后按照字母在单词中的顺序构建单词树。这种树一般适用于已知共同前缀的单词的补全。而它的缺点是对于不是共同前缀的单词无法进行提示。

编辑距离,又称 Levenshtein 距离,是指两个字符串之间,从一个字符串通过添加、删除、更改单个字母而得到的另一个字符串所要的次数。如 kitten 到 sitting 的编辑距离为 3。

求 Levenshtein 距离,一般使用动态规划算法,编辑距离的公式计算如下。

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0 \\ \min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1, a_i \neq a_j \\ lev_{a,b}(i-1,j-1), a_i = a_j \end{cases} & \text{other} \end{cases}$$

比如,“律师”和“专业律师答疑”之间的编辑距离为 4,其状态转移矩阵如下表所示:

表 7 状态转移矩阵表示

	0	专	业	律	师	答	疑
0	0	1	2	3	4	5	6
律	1	1	2	2	3	4	5
师	2	2	2	3	2	3	4

值得一提的是，上述公式当中将增、删、改的编辑距离都设为相同的值 1，但在实际操作中，可以根据增、删、改对两个字符串匹配的影响定义不同的权重，从而优化智能提示系统的效果。

在构建智能提示系统时，可以将一些常见的词表或者常见的关键词列表进行存储，一旦用户在输入框输入，则计算其输入内容和这些关键词的 Levenshtein 距离，并将所得到的结果由小到大排序，若有多个备选单词并列，则优先排序与输入单词重合字数最多的关键词，最后给出前几项最有可能的关键词备选。

4.2.2 模型的求解

4.2.2.1 构造智能推荐系统

对于推荐系统的实际性能，本文将用系统（算法）预测的准确度和覆盖率来刻画。

■ 精度和召回率

预测准确度即推荐系统或推荐算法预测用户行为的能力。定义精度如下：

$$Precision = \frac{\sum_{u \in U} |R_u \cap T_u|}{\sum_{u \in U} |R_u|}$$

其中， R_u 是根据目标用户在训练集上的行为为其提供的推荐列表， T_u 是目标用户在测试集上的预测行为列表。

同时，本文定义召回率为：

$$Recall = \frac{\sum_{u \in U} |R_u \cap T_u|}{\sum_{u \in U} |T_u|}$$

■ 覆盖率

在问题一中本文发现，在该网站的用户日志文件中存在明显的长尾效应，对于网页开发者而言，这部分被用户忽略的数据会造成信息的浪费和效益的低下。为此，本文定义了覆盖率来描述一个推荐系统对长尾数据的发掘能力：

$$Coverage = \frac{|\bigcup_{u \in U} R_u|}{|I|}$$

其中， $|I|$ 表示总的网页数量。

在使用 LFM 算法，经过对训练集的训练以及通过对测试集的测试后，得到对推荐系统性能检验的结果如下表所示：

表 8 推荐系统性能检验结果

算法	Precision	Recall	Coverage
随机算法	50%	50%	56.8%
LFM k=10	84.3%	53.6%	36.2%
LFM k=20	85.1%	53.7%	36.3%
LFM k=30	85.2%	53.7%	36.3%
LFM k=50	85.3%	53.7%	36.4%
LFM k=60	85.4%	53.8%	36.4%
LFM k=70	85.5%	53.7%	36.4%
LFM k=80	85.5%	53.7%	36.4%
LFM k=100	85.5%	53.7%	36.4%

对以上结果进行解释和分析：其中，随机算法作为空白对照组，随机算法是按照等概率随机选择推荐的网页，因此，该算法推荐的精度和召回率都为 50%。在精度这个指标上，本文所提出的推荐算法体现了很好的性能。在上表中，当 k 值为 60~80 时，LFM 算法的性能是最好的，而在本问题中，数据给出了 72 个隐含主题，包括定罪量刑、故意伤害、妇女权益、劳动合同纠纷等，正落于这个区间，因此证明 LFM 在解决本问题中的实际推荐系统问题时有很好的效果，比随机算法的性能要好得多。同时本文也发现当隐含主题 k 的个数只有 10、20 个的时候，精度也可以分别达到 84.3%、85.1%，说明了该算法较高的稳定性，同时也说明了该算法可以用更小的空间来存储分解的矩阵，大大降低了存储空间。

通过系统的推荐，用户将有更高的几率浏览自己感兴趣的页面，问题一中定义的有效页面数会增加，对应的，无效页面数则会减少，并且有效页面和无效页面的总数保持不变，同时用户也会减少相应的中间环节即过渡界面，更快的访问到终止界面。将以上关系表示如下，其中， $a, b, c \in (0,1)$ ：

$$\begin{aligned} ax'_N &= x_N \\ y'_N &= by_N \\ z'_N &= cz_N \\ x'_N + y'_N &= x_N + y_N \end{aligned}$$

则改进后页面的准确率和效率为：

$$\begin{aligned} \varphi' &= \frac{x'_N}{x'_N + y'_N} = \frac{x_N/a}{x_N + y_N} > \frac{x_N}{x_N + y_N} = \varphi \\ \omega' &= \frac{x'_N + y'_N}{x'_N + y'_N + z'_N} = \frac{x_N + y_N}{x_N + y_N + cz_N} > \frac{x_N + y_N}{x_N + y_N + z_N} = \omega \end{aligned}$$

从以上内容可以看出，一旦使用了推荐系统，用户就会浏览更多的有效页面，更快的到达目标页面，从而网站的准确率和效率均会提升。

4.2.2.2 构建智能提示系统

为了探究智能提示系统对于用户输入关键词的效率，本文使用附件给出的关键词列表进行离线测试计算。假设用户想要输入的关键词为 x_i ，该词处于系统关

关键词词典 D 的第 i 个位置中（若用户输入的词不在词典中，则该智能提示系统不能够提高用户输入的效率）。假设用户从关键词的前缀开始输入（非前缀时类似考虑），设智能提示系统的提示列表长度为 K 项，假设 x_i 的长度为 $Length(x_i)$ ，则假设输入长度为 $Input(x_i)$ ($Input(x_i) \leq Length(x_i)$) 时，单词 x_i 首次出现在备选项。因此，可以定义输入提升效率 IMP 为：

$$IMP = 1 - \frac{\sum_{x_i \in D} Input(x_i)}{\sum_{x_i \in D} Length(x_i)}$$

下表为当列表长度 K 取不同的值是，IMP 值的变化：

表 9 提示列表长度 K 与 IMP 的变化

提示列表长度 K	0	1	3	5	8	10	12	15	20
IMP (%)	0	13.7	22.6	26.0	29.1	29.9	30.6	31.7	32.3

将这种变化绘成折线图如下：

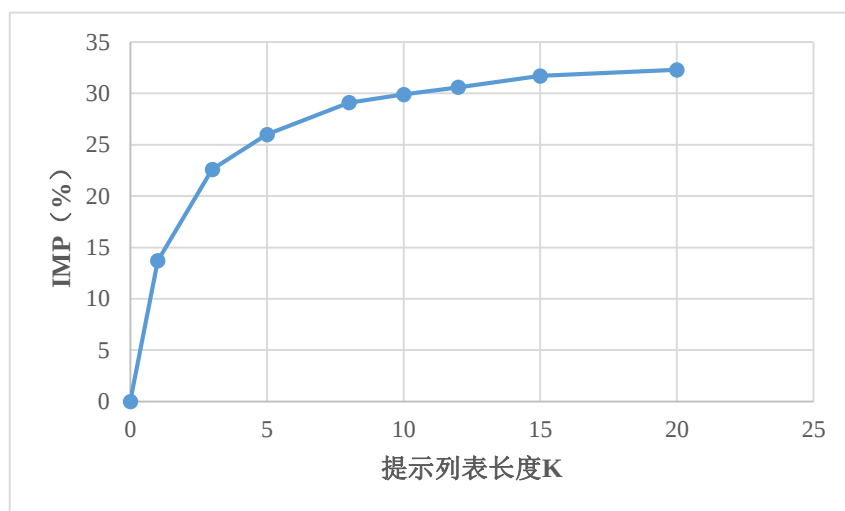


图 15 提示列表长度 K 与 IMP 关系图

由结果可知，当智能提示系统的备选项个数 K 为 0，即没有智能提示系统时，， $IMP=0$ ，这与实际情况相符，即此时系统无法给用户进行推荐，提升度即为 0；随着 K 的增加，IMP 显著提升，由折线图可以看出，提示列表长度由 0~8 增长过程时 IMP 上升速度很快，当 $K>8$ 时，IMP 上升速度缓慢， $K>15$ 后，曲线变化平缓。此外，考虑到当提示列表长度过高时占屏幕过高，给用户造成了不便，影响用户体验。因此，本文认为提示列表长度为 5~10 时效果较为理想。

5 结论

在问题一中，基于日志数据对用户行为的分析如下，首先对用户行为进行单维度分析，得到如下结论：

- 日志数据来自于 2015 年 2 月 1 日 00:00:06 到 2015 年 4 月 29 日 17:11:03 的用户浏览网站的记录, 包含 230149 位用户的浏览记录, 通过对每个用户浏览网站的行为分析, 本文发现用户浏览网站的行为有长尾效应;
- 用户来自于 13 个地区, 其中编号为 140100 地区的用户记录最多;
- 用户使用的操作系统中, Windows 系列系统所占比例高达 86%, 同时也发现了有部分用户使用移动端浏览网页;
- 用户浏览法律咨询网站所使用的入口最多的是直接在地址栏输入网址, 同时使用搜索引擎的用户也占很大一部分比例, 使用最多的搜索引擎是百度和好搜;
- 用户最经常浏览的页面类型为咨询内容页和知识首页;
- 用户最常搜索的关键词是法律咨询、律师、婚姻法等;
- 在用户所咨询或查找的法律信息类型最多的是定罪量刑、故意伤害、妇女权益、劳动合同纠纷等;

而后对用户行为进行双维度分析, 主要分析的是地区和标题类型两个因素, 可以得到结论: 总的来说, 用户们最关注的几个领域是定罪量刑、妇女权益、故意伤害、劳动合同纠纷等。但每个地区内用户们对这几个问题的关注程度是有细微差别的, 如在 140103 地区, 用户最关心的是故意伤害方面的法律知识, 在 140105 和 140106 区域, 除了定罪量刑这个方面, 用户们分别对妇女权益和劳动合同纠纷方面也有很高的关注度。

在问题一中的第二部分, 本文定义了有效页面、无效页面、过渡页面和终止页面四个关键的概念, 并基于这四个概念计算用户在针对特定问题找到相关界面或者专业律师的成功率和效率。最终得到的结果是, 在用户查询离婚类别问题时, 成功率为 12.60%, 效率为 56.81%; 在查询非离婚类别问题时, 成功率为 22.38%, 效率为 93.87%。出现这样的结果是由于离婚类别问题的数据相较其他类别问题的数据而言, 数据量更大; 且由于系统并未给用户信息提示或推荐, 用户在面对更多的数据信息时依然近似进行了随机选择, 这样在最终的页面中用户获取到的无关信息将会增多, 用户访问的无信息页面也会增多, 最终导致了准确率和效率的降低现象。也因此, 根据用户行为数据的分析进而给用户合理的信息和推荐, 对于用户在面对大量信息时提高访问的准确率和效率是非常重要的

在问题二中, 针对当前网页服务效率低, 本文从构建智能推荐系统、优化页面结构、构建智能提示系统三个方面对网站进行优化, 从而提高用户浏览网站的成功率和效率。

- 其中, 本文利用隐语义模型 (Latent Factor Model, LFM) 对网页的隐形反馈数据进行分析, 定义了预测精度 Precision、召回率 Recall、覆盖率 Coverage 来刻画推荐的效果。通过设置随机推荐的空白参照组, 证明了算法提升了系统的效率。然后结合问题一中定义的准确率和效率, 定性的分析了模型对于两者均有提升。
- 优化页面结构。本文针对无行为历史数据的新用户和有历史数据的老用户分别考虑。针对新用户, 提出了要利用用户基本信息, 如地区、使用的操作系统、社交资料等作为粗粒度个性化推荐的依据, 针对老用户则进一步考虑用户的历史数据。最后结合信息对网络运营商产生经济效益

的能力，共同构建了多目标的信息重要度函数，并据此对网页进行了简要设计，信息重要度高的信息更加突出的显示；

- 智能提示系统。本文利用 Levenshtein 距离构建了模型，用户在信息检索时，系统能够根据用户的实时输入进行智能推荐，这样弥补了法律知识匮乏的用户难以定位信息的问题。此外，通过离线测试计算，本文验证了算法对于提高检索精度的效果显著，且结合效率提升度 IMP 和提示列表长度，本文认为提示列表长度为 5~10 时效果比较好。

最后，根据本文的分析，结合当前 lawtime 网站的实际情况，本文对网站的进一步做出以下建议：

1. 增加内容质量：当前的用户在界面的停留时间普遍偏短，直接体现了用户对于网页内容不敢兴趣。网站内容是网站生命力的源泉，因此建议网站能增强网站的内容质量，这样才能确保网站用户不流失。
2. 推荐系统：由于法律咨询网站的使用者中，很多用户缺乏相应的法律知识，单纯依靠用户自主辨识网页展示内容是不可行的，首先是效率低，效果差，其次是用户体验度高，用户难以获取有用信息。因此研发适应网站知识背景的推荐系统是十分紧迫的。本文所采用的 LFM 算法在效率上优于协同过滤算法，能有效的提高推荐的准确率和效率。
3. 显性反馈：网站的发展离不开用户，加强与用户的互动能帮助开发展更好的掌握用户的需求，及时对网站维护进行调整。当前网页主要采用的是隐性反馈，我们认为适当的增加用户的显性反馈能更好的反应用户的需求，且能够使得推荐系统的效率得到更大的提升。
4. 智能提示：由于网页界面的版面有限，而且要保证展示信息的质量，开发者不能展示所有的信息，而只能展示部分信息，这会导致马太效应。而且针对用户的特殊需求，应加强用户的自定义查找，但为了保证查找的效率更高，应增加智能提示，帮助用户快速获取有用信息。
5. 结构优化：当前网页主要是针对关注领域的热度进行信息排序，重点信息不突出，而且展示的内容中有大量的长文本标题，对于用户而言很难获取有价值信息。因此建议网站对用户进行个性化推荐，而且针对重要的信息进行重点展示，尽量保证页面的简洁。

6 模型的优点与缺点

6.1 模型的优点

1. 充分利用数据，合理分析了用户的行为特征；
2. 利用了 LFM 推荐算法，分类依据完全来源于数据，避免了人为的定义推荐标准；
3. 针对当前网页的不足，提出了信息重要度的概念，结合了用户的基本信息、历史数据和经济效益等诸多因素对网页结构进行了优化。

6.2 模型的缺点

1. 由于停留时间没有给出单位，因此对于用户页面停留时间没有合理考虑，而定性的对其进行了分类；
2. 对于网页结构的设计只是给出了概念性的设计，缺乏定量的数据分析和模型仿真；
3. 网页关于律师的信息不完整，因此粗略的将律师信息与页面信息做了整合；
4. 在构建智能系统方面，若用户输入的词不在词典中，则该智能提示系统不能够提高用户输入的效率。

7 模型的进一步讨论

本文通过构建智能推荐系统、智能提示系统以及优化页面结构来提升用户的体验，主要体现在提升用户寻找信息的成功率和使用效率。如今，随着人工智能技术的不断发展，智能问答系统（Question Answering）已慢慢进入人们的视野。智能问答系统通过人机交互，精确定位用户所需要的知识。

智能问答系统最大的优点是省去了用户自行查找信息的时间，全面提升了网站的用户体验。智能问答系统通过对网站知识的建模，将网站知识结构化，从而智能的选择并回答用户的提问。在下一步的研究工作中，我们将知识建模、知识图谱、智能问答系统的各项技术应用于法律服务问答网站中去，令法律咨询服务网站的用户体验进入一个全新的阶段。

参考文献

- [1] 刘建国, 周涛, 汪秉宏. 个性化推荐系统的研究进展[J]. 自然科学进展, 2009, 19(1): 1-15.
- [2] 王国霞, 刘贺平. 个性化推荐系统综述[J]. 计算机工程与应用, 2012, 48(7).
- [3] Linden G, Smith B, York J. Amazon. com recommendations: Item-to-item collaborative filtering[J]. Internet Computing, IEEE, 2003, 7(1): 76-80.
- [4] Koren Y. Collaborative filtering with temporal dynamics[J]. Communications of the ACM, 2010, 53(4): 89-97.
- [5] Adomavicius G, Tuzhilin A. Context-aware recommender systems[M]//Recommender systems handbook. Springer US, 2011: 217-253.
- [6] Rendle S. Factorization machines with libFM[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2012, 3(3): 57.
- [7] Chin W S, Yuan B W, Yang M Y, et al. LIBMF: A Library for Parallel Matrix Factorization in Shared-memory Systems[J].
- [8] Amatriain X, Lathia N, Pujol J M, et al. The wisdom of the few: a collaborative filtering approach based on expert opinions from the web[C]//Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2009: 532-539.

- [9] 项亮. 推荐系统实践[M]. 人民邮电出版社, 2012.
- [10] 詹尼士. 推荐系统[M]. 人民邮电出版社, 2013.
- [11] Pan R, Zhou Y, Cao B, et al. One-class collaborative filtering[C]//Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. IEEE, 2008: 502-511.