DOI:10.3969/j.issn.1673-4785.201203024
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20120925.1009.001.html

# 面向浏览推荐的网页关键词提取

闫兴龙[1,2,3],刘奕群[1,2,3],马少平[1,2,3],张敏[1,2,3],茹立云[1,2,3]
(1.清华大学 计算机科学与技术系,北京 100084;2.清华大学 智能技术与系统国家重点实验室,北京 100084;3.清华大学 清华信息科学与技术国家实验室(筹),北京 100084)

**摘 要**:在网页浏览推荐任务中,如何利用网页内容选取合适的推荐关键词是具有挑战性的研究热点.为了实现有效的关键词推荐方法,利用大规模的真实网络用户浏览行为数据,以及相关提取算法和新词发现算法实现并比较了基于领域关键词提取技术和基于查询词候选集合的关键词推荐方法.实验结果证明,2种方法都能够有效地表征用户信息需求,而第1种推荐方法的准确率更高,具有更好的推荐性能.

**关键词**:浏览推荐;关键词推荐;关键词提取;网页关键词
**中图分类号**:TP391 **文献标志码**:A **文章编号**:1673-4785(2012)05-0398-06

# Study on website keyword extraction for browsing recommendation

YAN Xinglong[1,2,3], LIU Yiqun[1,2,3], MA Shaoping[1,2,3], ZHANG Min[1,2,3], RU Liyun[1,2,3]
(1. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China; 2. State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China; 3. Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China)

**Abstract**: It is very challenging when conducting research and it is especially difficult as it pertains to website browsing and recommendation system task for selection of suitable keyword usage. This research study will focus on proper use of website browsing and recommendations on how to select keywords for conducting research. The challenge is to leverage user behavior features, as well as develop an effective keyword's recommendation content page. The implementation of a comprehensive user browsing data, relevant extraction algorithm and algorithm finding methods for new keywords were examined in the research study. The research study also proposed additional, keyword recommendation methods utilizing large-scale and related algorithm approaches for domain-specific keyword extraction technology and a query keyword candidate set were compared. The experiment results confirm both methods demonstrate that they satisfy users' information demand. However, the keyword recommendation methods show a significant performance improvement in effectiveness. The keyword recommendation method has a higher accuracy and better recommendation performance.

**Keywords**: browsing recommendation; keyword recommendation; keyword extraction; web keywords

自21世纪以来,随着互联网的不断发展,我国网民人数稳定增加.据中国互联网信息中心测算,截至2011年12月,中国网民规模达到5.13亿,全年新增网民5 580万,互联网普及率较2010年提升4个百分点,达到38.3%.越来越多的人通过互联网获取各种信息,网络资源正逐渐成为人们获得知识的主要渠道,截至2011年12月底,中国网页数量为866亿个,比2010年同期增长44.3%.从网民数和网页数来看,现在网民从互联网获取信息存在信息过载的问题.过量的信息呈现在用户面前,用户很难快速地获取自己需要的信息,这样就降低了用户的信息使用效率.现在用户提高信息使用效率,过滤无

收稿日期:2012-03-29. 网络出版日期:2012-09-25.
基金项目:国家自然科学基金资助项目(60736044, 60903107, 61073071);高等学校博士学科点专项科研基金资助项目(20090002120005).
通信作者:闫兴龙. E-mail: yan-xinglong@163.com.

用信息的主要方法是通过门户网站和搜索引擎. 门户网站和搜索引擎在一定程度上满足了用户信息过滤的需求,但是门户网站和搜索引擎都有其存在的问题,门户网站的主要问题就是网页的过滤是通过人工的方法进行,这样会费时费力,而且并不能满足每个人的信息需求. 搜索引擎是当前非常重要的用户获取信息的途径[1],其主要的问题有 2 个方面:1)无法提供用户的个性化需求;2)用户需要较为繁琐地提供需求来获取信息. 为了给用户提供更好、便捷和个性化的服务,推荐系统应运而生. 推荐系统和搜索引擎的主要区别在于:1)搜索引擎面向的是所有的用户,提供主流的结果,推荐系统更重要地是研究用户模型,利用用户的历史记录或者社交网络提供用户的个性化服务;2)搜索引擎是用户主导的,需要用户主动提供和修改查询词,推荐系统是由系统主导用户浏览,能够提供更好的推荐结果. 高质量的推荐系统能够使用户更加依赖该系统,提高用户的忠诚度.

## 1　相关工作

当前网页推荐系统基本上可以分为 3 种方法:基于日志挖掘的推荐方法、基于知识的推荐方法和基于内容的推荐方法.

1)基于日志挖掘的推荐方法. 基于日志挖掘的推荐方法[2-6]主要是根据用户的 Web 访问日志信息,划分出用户的会话,通过模式匹配以及关联规则等数据挖掘的方法,推荐出用户需要的网页. 这种方法很好地利用了用户行为,能够更好地实现个性化需求,但是由于互联网的扩散性和数据的稀疏性,这种方法只能应用于小规模的封闭集合.

2)基于知识的推荐方法. 该方法更多的是利用知识工程的方法对网页进行分析,在某种程度上可以看成是一种推理技术. 它主要是通过语义 Web 的分析[7-11],得到各个网页之间的关系,从而由系统推荐出网页.

3)基于内容的推荐方法. 该方法[2,12]是当前网页推荐系统最主要的方法,它首先提取网页中用户的信息需求,然后通过一系列的数据挖掘方法得到推荐的对象. 所提取的用户信息需求特征主要通过关键词来表示,关键词的质量是影响这种方法最主要的因素. 当前基于内容的关键词提取主要通过以下 2 种方法实现:①基于已有的分词程序中的词语集合[2];②基于已有的词语词典[9]. 但是上述 2 种方法同样存在各自的问题,在第 1 种方法中,分词程序中的词语往往很短,无法得到更能反映用户需求的长关键词;第 2 种方法并不针对网页文本,网页文本和书面文本存在一定的差异,并不一定能表征用户的信息需求. 本文对这 2 种方法得到的关键词候选集进行对比实验,结果表明,基于用户行为信息的领域关键词提取技术有更好的效果.

## 2　基于领域关键词提取技术的关键词推荐方法

领域关键词抽取方法的流程如图 1 所示.

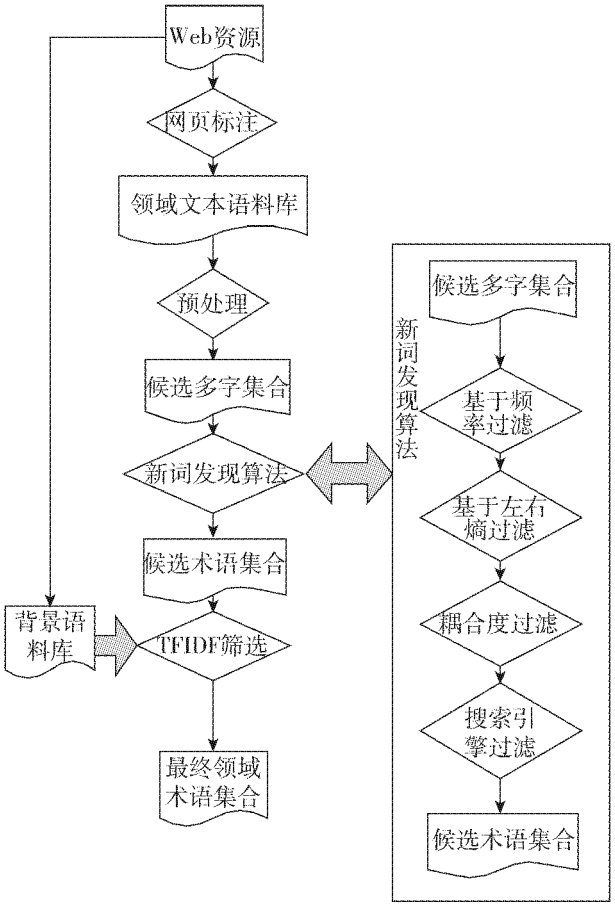



**图 1　基于 Web 资源领域关键词提取方法框架**

**Fig. 1　Framework of domain-specific term extraction based on Web resource**

本文主要通过 4 步运算,可以得到最终领域关键词的集合.

1)网页标注. 网页标注主要通过归纳总结找到某个领域的网页 url 的规律和特点,最终总结出基于 url 的网页筛选方法. 通过该方法可以得到某领域相关的 Web 资源. 大型新闻门户网站中某领域网页的 url 均是在某个子域名下,而某领域专业网站下的网页一般为该领域的相关文本.

2)预处理. 预处理为新词发现算法处理语料库,对原有的网络文本进行整理,如对网页正文进行抽取,以

及对原有文本不规则内容进行整理,然后对句子进行切分,得到多字集合,用于新词发现算法处理.

3)新词发现. 新词发现基于上一步得到的候选多字集合. 该方法首先统计候选多字集合中每个候选多字出现的频率,将低于某频率阈值的多字滤除出候选集合;然后分别计算每个候选多字的左右信息熵,将低于某熵值的多字滤除.

假设词语 $w$ 属于候选集,另外,$A=\{a_1,a_2,\cdots,a_m\}$ 和 $B=\{b_1,b_2,\cdots,b_k\}$ 分别为该词语对应的左右单字集合,则左右熵的定义为:

$$E_{\mathrm{L}}(w)=-\frac{1}{n}\sum_{a_i\in A}C(w,a_i)\log\frac{C(w,a_i)}{n},$$

$$n=\sum_{a_i\in A}C(w,a_i);$$

$$E_{\mathrm{R}}(w)=-\frac{1}{m}\sum_{b_i\in B}C(w,b_i)\log\frac{C(w,b_i)}{m},$$

$$m=\sum_{b_i\in B}C(w,b_i).$$

式中:$C(w,a_i)$ 和 $C(w,b_i)$ 分别是词语 $w$ 的左单字 $a_i$ 和右单字 $b_i$ 出现的次数.

对于一个实际存在的词而言,如果它的出现频率较高且左右单字集的频率也很高,则可以通过其左信息熵和右信息熵的方法进行过滤.

通过上一步的过滤,仍然有部分非词语无法过滤,如“化股份”这个词,从语义的角度来讲,该候选词中的“股份”应该和“化”分开,之前没有分开的原因是由于该词的左信息熵过大,这样依据上一步的规则无法被滤除. 根据已有的信息熵和候选词的出现频率,提出基于递推的耦合度过滤算法,具体算法如下.

①对于字长为3的 $w$,如果存在 $w_1\in T_2$($T_2$ 为长度为2的候选词集合),$w$ 可分解为 $p+w_1$,$p$ 为单字. 计算 $p$ 和 $w_1$ 的耦合度为

$$C_o(p,w_1)=\left(\frac{C(w)}{C(w_1)}<\lambda\right)\cap\left(E_{\mathrm{L}}(w)<E_{\mathrm{L}}(w_1)\right)\cap\left(E_{\mathrm{L}}(w)<\gamma\right).$$

如果存在 $w_1\in T_2$($T_2$ 为长度为2的候选词集合),$w$ 可分解为 $w_1+p$,$p$ 为单字. 计算 $p$ 和 $w_1$ 的耦合度公式为

$$C_o(p,w_1)=\left(\frac{C(w)}{C(w_1)}<\lambda\right)\cap\left(E_{\mathrm{R}}(w)<E_{\mathrm{R}}(w_1)\right)\cap\left(E_{\mathrm{R}}(w)<\gamma\right).$$

式中:$\gamma$ 和 $\lambda$ 为参数阈值. 如果耦合度的值等于1,则认为 $w$ 不应该为词.

②对于字长为4的 $w$,如果存在 $w_1\in T_3$($T_3$ 为长度为3的候选词集合),$w$ 可分解为 $p+w_1$,$p$ 为单字. 计算 $p$ 和 $w_1$ 的耦合度公式为

$$C_o(p,w_1)=\left(\frac{C(w)}{C(w_1)}<\lambda\right)\cap\left(E_{\mathrm{L}}(w)<E_{\mathrm{L}}(w_1)\right)\cap\left(E_{\mathrm{L}}(w)<\gamma\right).$$

如果存在 $w_1\in T_3$($T_3$ 为长度为3的候选词集合),$w$ 可分解为 $w_1+p$,$p$ 为单字. 计算 $p$ 和 $w_1$ 的耦合度公式为

$$C_o(p,w_1)=\left(\frac{C(w)}{C(w_1)}<\lambda\right)\cap\left(E_{\mathrm{R}}(w)<E_{\mathrm{R}}(w_1)\right)\cap\left(E_{\mathrm{R}}(w)<\gamma\right).$$

式中:$\gamma$ 和 $\lambda$ 为参数阈值. 在耦合度计算中,如果交集中的每个不等式都成立,则耦合度的值等于1,否则耦合度为0. 如果耦合度的值为1,则认为 $w$ 不应该为词,将 $w$ 滤除.

对于参数的估计,采用最小二乘法实现,首先抽取已过滤候选集中的1 000个样本,对样本进行标注,根据抽取出样本的数据,计算出 $\frac{C(w)}{C(w_1)}$ 值和 $E_{\mathrm{R}}(w)$ 或者 $E_{\mathrm{L}}(w)$ 值,对已得到的数据进行组合,得到候选参数集合,通过计算每对候选参数所对应的样本正确率,将最高正确率的参数对作为估计出的参数. 实验表明,该算法可以有效地滤除候选集合中的非词语,并保留实际存在的词语.

以此类推,可以得到更长长度的词. 进行耦合度过滤之后,将得到的结果放入搜索引擎进一步过滤,最后得到候选关键词集合.

4)TF/IDF 筛选. TF/IDF 是一种常用的计算某个词在某篇文档或部分文档集合中重要程度的方法. 基于 TF/IDF 筛选是为了更好地得到与领域相关的关键词,通过计算每个候选关键词在文本语料库中的 TF/IDF 值,得到每个候选关键词在领域文本语料库的重要程度. TF/IDF 值的定义如下.

对于文档 $d$,候选关键词 $w$ 对应的词频及文档频度倒数特征的计算公式为:

$$\mathrm{tfidf}(d,w)=\mathrm{tf}\times\mathrm{idf},$$

$$\mathrm{tf}=\frac{f(w)}{\sum f(w)},$$

$$\mathrm{idf}=\log\left(\frac{|D|}{|\sum d|}+0.01\right).$$

式中:$f(w)$ 为词 $w$ 在文档 $d$ 中出现的次数,$\sum f(w)$ 为一篇文档的总词数,$|D|$ 为语料库中的文件总数,$|\sum f(w)|$ 为包含词语 $w$ 的文件数目. 使用类别的 TF/IDF 作为候选词的评价函数,其公式为

$$\mathrm{tfidf}(w)=\sum_{d\in D}\mathrm{tfidf}(d,w).$$

以上述领域关键词提取技术得到的领域关键词为候选集合,对于每个候选词,提取其在网页中的以下特征:标题中出现的次数、标题中第一次出现的位置、正文中出现的次数和在正文中第一次出现的位置,并且结合关键词本身的一些特征,如关键词的长度、其在领域关键词提取时的频率以及 TF/IDF 值. 依据这 7 个特征,利用线性拟合的方法得到各个特征的参数,根据这些特征的特征值,得到排序较高的前 3 位结果作为推荐的关键词,即用户的兴趣点所在.

## 3　基于用户查询集合的关键词推荐方法

### 3.1　用户查询集合的候选集选取

当前用户在浏览网页时获取所需信息的重要方式便是通过搜索引擎提交查询词,所以用户提交的查询词是反应用户兴趣的重要信息,以查询词为候选集合,对用户进行关键词推荐,能够很好地表征用户的信息需求.

基于用户查询词集合的关键词推荐方法,首先需要对查询词的候选集合进行选取,采用的查询词选取方法的步骤分为以下 3 步(如图 2 所示).

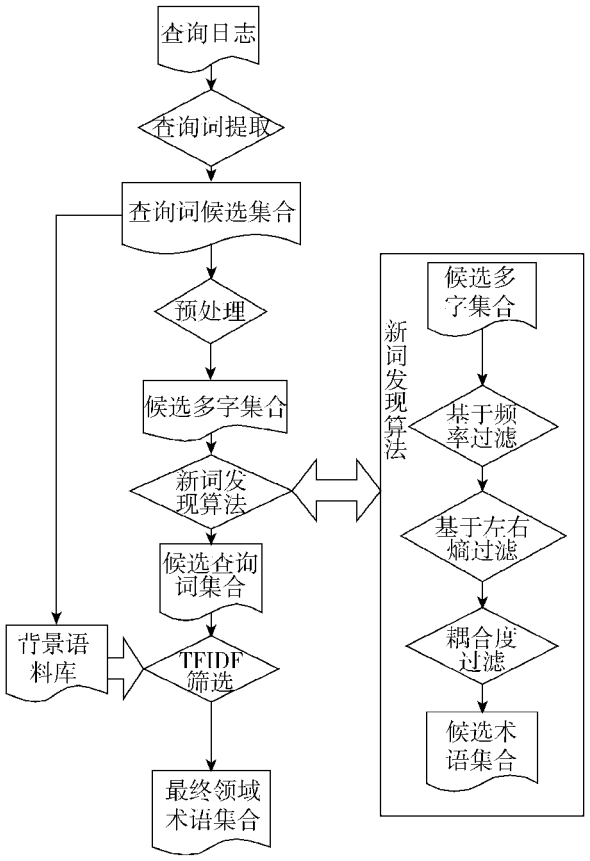



**图 2　用户查询词候选集合选取方法**

**Fig. 2　Framework of selecting query candidate**

1)预处理. 因为查询词中有部分噪音的存在,比如标点符号、不成词的查询. 首先去除存在的标点符号,因为一般正常的查询都不存在符号. 接下来去除查询频率过低和过高的查询词,频率过低的查询词一般不是真实存在的词语,而频率过高的查询词又没有真正的区分度,用户往往只是为了利用搜索引擎引导到某个网站,如“新浪”、“搜狐”等.

2)利用领域关键词提取技术中的新词发现算法,主要针对查询中小于等于 4 个字的结果,去除查询词中非词语,从而提高检索的速度,该算法能很好地滤除非词语.

3)建立基于长查询词和短查询词的映射关系. 一方面,由于查询词的集合过于庞大,为了提高查询词匹配的速度,所以采用 2 级索引的方法. 另一方面,短查询词反映的含义简单扼要,但是有部分查询词无法全面反映用户的意图,所以采用 2 级索引,能更加全面地反映用户的信息需求.

### 3.2　基于用户查询集的关键词推荐方法

利用上述方法得到的短查询词集合为候选集合,对于每个候选词,提取其在网页中的以下特征:标题中出现的次数、标题中第一次出现的位置、正文中出现的次数和在正文中第一次出现的位置,并且结合查询词本身的一些特征,如查询词的长度、其查询的频率以及 TF/IDF 值. 根据这些特征,利用线性拟合的方法,得到各个特征的参数值,通过计算得到排序靠前的查询候选词.

将得到的排序靠前的短查询词索引下的长查询词作为长查询词候选集合,对于这个候选集合中的候选词,以对应的短查询词的得分为基础,加入长查询词的查询频率这个特征. 最后根据上述 2 个特征值,排序得到推荐的长查询词. 从实验的结果可以看出,长查询词往往是与文本内容相关的信息内容.

## 4　实　验

关键词的质量是影响基于内容的网页推荐系统效果的重要因素,使用计算不同准确率的方法评价关键词的推荐方法,准确率 $P$ 的定义为

$$P = \frac{\text{推荐出相关关键词的网页数}}{\text{推荐的网页数}}.$$

### 4.1　实验数据

由于领域关键词提取技术只能提取出单个领域的关键词,因此本文主要针对单个领域的网页进行推荐,下面以财经领域的网页为例进行实验.

利用用户浏览行为信息,抽取用户在浏览过程中点击的锚文本信息,以该信息作为提取关键词的背景语料库. 锚文本是指由网页制作者编写的,用于描述对应的超链接网页内容的文本样式. 数据是在

“用户体验改进计划”中抽取的,数据收集经过了用户的同意,并删除了用户的IP、用户名等个人信息.查询集合采用某商业搜索引擎18 d(2010-10-08—10-25)的查询日志,对于锚文本,采用同一时期的用户浏览日志信息.

随机选取10 000个网页url,从中提取出财经领域的网页,并且筛选出不合格的网页,共提取出134个与财经相关的网页,利用领域关键词提取方法对相关网页进行推荐,然后通过人工标注,计算出该方法的前1位和前3位的准确率.

基于用户查询集合的关键词推荐方法的实验则是从10 000个url中随机抽取1 000个进行推荐,并计算相应的前1位和前3位的准确率.

### 4.2 实验结果及分析

通过标注,得到基于领域关键词提取技术和基于查询词集合的关键词推荐方法在不同网页下的实验结果,如表1所示.从实验结果可以看出,这2种方法得到的关键词推荐都能得到较好的推荐效果,但是基于领域关键词提取技术的关键词推荐效果更为显著,具有更高的准确率.

表1 2种方法的实验结果

Table 1 Results of two methods %

| 方 法 | $P_{@1}$ | $P_{@3}$ | $P$ |
|---|---|---|---|
| 基于领域关键词提取技术 | 97.0 | 91.0 | 97.0 |
| 基于查询集合 | 76.2 | 72.3 | 77.3 |

对实验的结果进行具体的分析,造成基于领域关键词提取技术推荐错误的主要原因在于不存在相对应的候选关键词.例如:网页标题为“主力研究”,正文为“沪深两市依旧是小盘股全面活跃,大盘股不涨反跌.虽然不少投资者……”,基于领域关键词提取技术的推荐方法得到的结果为“主力”.通过分析,候选集合中没有“主力研究”这个词,但在该网页中,只有“主力研究”能够很好地反映该网页的内容.对于基于查询集合的关键词推荐方法而言,导致推荐结果不对的原因主要是某些关键词在查询词中并没有出现,导致候选集合中并没有这些关键词.有如下的例子:网页标题为“Q友乐园”,网页正文为“Q友乐园,专注分享精品头像与个性素材的专业性网……”,查询词候选集合中,并没有“Q友乐园”这个候选词,所以最终的推荐结果中,只推荐了“乐园”这个词.由于领域关键词提取技术主要针对单个领域的词进行相应的过滤和提取,所以能够更好地获取某个领域的关键词.而基于用户查询集合的关键词推荐方法则主要依据用户提交的查询词,造成查询词候选集合词汇不足的主要原因有:1)用户提交的查询词无法涵盖所有关键词;2)由于查询词集合过大,对长尾查询词进行过滤,导致丢失了部分有用的查询词数据.

## 5 结束语

基于领域关键词提取技术的关键词推荐方法可以更好地把握用户的信息需求,但是其有一定的局限性,只能在单个领域中发挥较好的作用.而基于查询词集合的关键词推荐方法可以在各个领域推荐出用户需求的信息,虽然在准确率和召回率方面有一定的缺陷,但是其普适性对于推广该方法有很大的帮助.接下来的工作中,将结合这2种方法的优缺点,得到更高效、准确的关键词推荐方法.

## 参考文献:

[1]许海玲,吴潇,李晓东,等.互联网推荐系统比较研究[J].软件学报,2009,20(2):350-362.
XU Hailing, WU Xiao, LI Xiaodong, et al. Comparison study of internet recommendation system[J]. Journal of Software, 2009, 20(2): 350-362.

[2]张培颖.基于Web内容和日志挖掘的个性化网页推荐系统[J].计算机系统应用,2008,17(9):9-12.
ZHANG Peiying. Personalized web page recommendation system based on web content and log mining[J]. Computer System and Applications, 2008, 17(9): 9-12.

[3]YANG Qingyan, FAN Ju, WANG Jianyong, et al. Personalizing web page recommendation via collaborative filtering and topic-aware Markov model[C]//IEEE International Conference on Data Mining. Sydeny, Australia, 2010: 1145-1150.

[4]SUMATHI C P, VALLI R P, SANTHANAM T. Automatic recommendation of web pages in web usage mining[J]. International Journal of Computer Science and Engineering, 2010, 2(9): 3046-3052.

[5]刘强,郭景峰.基于用户访问路径分析的页面推荐模型[J].计算机技术与发展,2007,17(1):151-154.
LIU Qiang, GUO Jingfeng. A web page recommendation model based on analyzing user access pattern[J]. Computer Technology and Development, 2007, 17(1): 151-154.

[6]WU Y H, CHEN Y C, CHEN A L P. Enabling personalized recommendation on the web based on user interests and behaviors[C]//Proceedings of the 11th International Workshop on Research Issues in Data Engineering. Washington, DC, USA: IEEE Computer Society, 2001: 17-24.

[7]邵华,高凤荣,邢春晓,等.基于VSM的分层网页推荐算法[J].计算机科学,2006,33(11):85-88,105.
SHAO Hua, GAO Fengrong, XING Chunxiao, et al. A hi-

erarchical webpage recommendation algorithm based on vector space model[J]. Computer Science, 2006, 33(11): 85-88, 105.

[8]杨学明,蒋云良. 基于语义的自适应个性化网页推荐[J]. 情报理论与实践, 2009, 32(3): 93-96.
YANG Xueming, JIANG Yunliang. Based on semantic adaptive personalized pages recommendation[J]. Information Studies: Theory and Application, 2009, 32(3): 93-96.

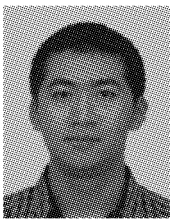
[9]梁邦勇,李涓子,王克宏. 基于语义 Web 的网页推荐模型[J]. 清华大学学报:自然科学版, 2004, 44(9): 1272-1276, 1281.
LIANG Bangyong, LI Juanzi, WANG Kehong. Web page recommendation model for the semantic web[J]. Journal of Tsinghua University: Science and Technology, 2004, 44(9): 1272-1276, 1281.

[10]袁燚,张璟,李军怀. 基于网页关键词的个性化 Web 推荐算法[J]. 西安理工大学学报, 2007, 23(1): 59-61.
YUAN Yan, ZHANG Jing, LI Junhuai. A personal web recommendation algorithm based on web page key words [J]. Journal of Xi'an University of Technology, 2007, 23(1): 59-61.

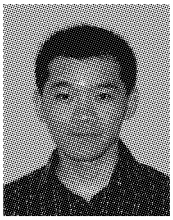
[11]杨学明. 基于本体学习的个性化网页推荐[J]. 情报杂志, 2009, 28(3): 171-174, 198.
YANG Xueming. Personalized web recommending based on ontology learning[J]. Journal of Intelligence, 2009, 28(3): 171-174, 198.

[12]赵银春,付关友,朱征宇. 基于 Web 浏览内容和行为相结合的用户兴趣挖掘[J]. 计算机工程, 2005, 31(12): 93-94, 198.
ZHAO Yinchun, FU Guanyou, ZHU Zhengyu. User interest mining of combining web content and behavior analysis [J]. Computer Engineering, 2005, 31(12): 93-94, 198.

**作者简介:**



闫兴龙,男,1986 年生,硕士研究生,主要研究方向为信息检索、推荐系统.



刘奕群,男,1981 年生,助理研究员,博士,中国人工智能学会知识工程专委会副秘书长. 主要研究方向为网络搜索与性能评价、面向搜索引擎的用户行为分析. 2010 年获"钱伟长中文信息处理科学技术奖"青年创新一等奖. 申请专利 13 项,其中 6 项已获得授权. 发表学术论文 50 余篇,出版教材 1 部.

马少平,男,1961 年生,教授,博士生导师,博士,中国人工智能学会副理事长、知识工程专业委员会主任,中国中文信息学会理事、信息检索与内容安全专业委员会副主任. 主要研究方向为智能信息处理,包括模式识别、文本信息检索、图像信息检索、中文古籍的数字化与检索等. 作为项目负责人承担"973"计划、"863"计划、国家自然科学基金和国际合作项目多项. 发表学术论文 70 余篇,出版专著 2 部.