基于检索日志的检索词推荐研究

边鹏 [1,2]　苏玉召 [1,2]

(1. 中国科学院文献情报中心，北京，100190; 2. 中国科学院研究生院，北京，100049)

[摘要] 为了满足检索用户对推荐服务日益迫切的需求，本文研究推荐理论，结合检索词推荐需求，基于三种典型推荐方法——基于内容的过滤、基于规则的过滤和基于协作的过滤，提出一种检索词的混合推荐方法，并基于检索日志构建了一种“脱机预处理和挖掘、联机推荐”的检索词推荐模型。最后，在 NSTL 嵌入式系统上进行实证研究，基于检索日志数据，以简单检索方式下的检索词推荐为突破口，设计一套原型系统，验证了检索词的推荐效果。同时，原型系统采用 BWP 方法确定最佳聚类数，提高原型系统的自动化水平，并且在原型系统上的检验了一种改进的 BWP 方法的效果。

[关键词] Web 日志挖掘；推荐系统；个性化；最佳聚类数

[分类号] TP311, G350

Research on the Recommendation of Retrieval Words Based on Retrieval Log

Bian Peng[1,2]　Su Yuzhao[1,2]

1 (National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

2 (Graduate University of Chinese Academy of Sciences, Beijing 100049, China)

[Abstract] Based on the theory research on the Recommendation of Information Service, three typical recommendation methods are induced into the recommendation of the retrieval words in order to meet users' increasing and pressing demand. And a combined recommendation method is provided in this essay and a recommendation model of retrieval words based on the retrieval log is formed which is 'preprocessing and mining off line, recommending on line'. At last, the lab is conducted on the NSTL Embedded Resource Services System. A prototype has been designed base on the retrieval log. This prototype aims on the simple retrieval recommendation, and proves good effect. In the meanwhile, the BWP method has been applied in the prototype system, which improved the automatic level of the prototype system. And this prototype also inspects the effect of an optimized BWP method.

[Keywords] Web Log Mining; Recommendation System; Personalization; Optimal Cluster Number

1. 引言

为了满足用户对检索信息的推荐服务日益迫切的需求，本文研究检索词推荐方法，基于三种典型推荐方法——基于内容的过滤、基于规则的过滤和基于协作的过滤，提出一种检索词的混合推荐方法，并从用户体验角度出发，针对现有研究的不足，构建一套基于检索日志的检索词推荐模型，以 NSTL 嵌入式系统为实验平台，设计一套推荐效率较高、推荐效果较

好的原型系统，提供相似检索词推荐、关联检索词推荐、同类用户检索词推荐三大功能，通过实验检验原型系统。

2. 检索词推荐方法研究

推荐依据的原理是数据挖掘理论。因此，推荐使用的方法主要来自数据挖掘理论。通常，推荐分为基于规则过滤、基于内容过滤、基于协作过滤的方法、以及两种方法混合的推荐方法。根据不同的个性化推荐方法，采用的挖掘算法也各不相同[1]。本文所述的推荐以检索词作为推荐对象，同时考虑三种推荐方法，满足信息检索系统用户丰富的个性化信息服务需求。

2.1 基于内容的过滤

基于内容的过滤推荐技术特点是[2]，根据用户过去选择项目的特点，系统为其推荐相似的项目。基于内容的过滤系统最大的缺点是用户模型的建立过度依赖于用户以前选择和点击的具体项目。研究显示在线推荐系统对用户最有用的价值是为其推荐意想不到的项目[3]，但是，如果只是采用内容相似性方法可能会丢失一些重要实用的关系，这些关系存在于 Web 对象之间，例如特定任务上下文中常用和补充的应用 Web 对象关系。

2.2 基于规则的过滤

Forsati 等人提出一种用于 Web 个性化基于权重的关联规则算法，该算法是对传统的关联规则算法的扩展，允许交易中的每一个项目分配一个权重以反映用户对该项目的兴趣度。在结果关联规则集里每一个项目都对应一个权重参数，根据用户的兴趣程度，为每个用户访问的 Web 页面分配一个时间权重和访问频率权重。实验结果表明，与传统的关联规则方法相比较，这种方法能够客观地、更有效的表示预测结果，对推荐系统效率有很大改进[4]。但基于关联规则的过滤对挖掘稀有信息的效率不高。

2.3 基于协作的过滤

基于内容过滤的推荐系统根据商品内容的相似性进行推荐，而协作过滤推荐系统利用了用户的相似性进行推荐[1]。但是，协作过滤技术也有其潜在的严重不足，最大的缺点是缺乏可伸缩性[1]。

2.4 当前检索词推荐研究的不足

当前解决基于内容过滤和协作过滤不足的研究热点是采用混合推荐算法，目的是提高推荐的精度。Burke提出的方法是混合基于内容和协作过滤技术，通过丰富变量的方法生成推荐系统，旨在提高推荐的质量[5]。有的混合推荐系统，例如，Ardissono 等人的用户建模和个性化推荐技术研究[6]，通过收集多种用户偏好的信息，采用多种异构推荐技术的方法实现。这种方法越来越多地被用于各种个性化服务中，例如，Nima等人关于Q-learning的Web推荐系统研究[7]，Chen等人的混合手机新闻推荐系统的普适访问研究[8]，李秦等人混合基于内容和基于规则的检索推荐系统研究[9]。但是，仅采用上述一种或两种方式的推荐无法满足用户日益增长的个性化信息服务需求。

2.5 一种同时采用上述三种方法的混合推荐方法

考虑到上述三种推荐方法各自存在一定局限性，为了使信息检索系统的个性化推荐更加有效，尽可能向用户提供丰富的个性化服务，本文提出一种同时采用上述三种推荐的混合方法。在用户检索时，信息检索系统同时向用户推荐基于内容的、基于规则的、基于协作的三种过滤结果，即相似检索词、关联检索词、同类用户检索词，为信息检索系统用户提供丰富的推荐信息。

对于信息检索系统用户而言，基于内容的检索词过滤是指用户在使用检索服务时，输入各种感兴趣的检索词，这些检索词中彼此有些是相似度较高的，可以推荐给用户。涉及到的技术包括数据预处理、聚类和推荐。首先要将检索词从服务器日志中清洗出来，进行中英文

分词后，得到检索词的最小词集合，停用其中的中英文小品词、标点符号等无用的词，得到有效的词集合。为了方便聚类，还需要将有效词集合转化成文本向量，这样就将字符串数据转化成了实数数据。然后，对向量进行聚类，将聚类的结果与原始的检索词对应起来存到数据库中。当用户输入检索词与数据库中某类检索词相同，就向该用户推荐该类检索词中出现频率最高的其他检索词。

基于规则的检索词过滤是指同一用户在使用检索服务时，输入的检索词可能具有内在联系，当用户再输入某个检索词时，可以推荐同时出现频率较高的其他检索词，涉及到的技术包括数据预处理、关联规则和推荐。注册用户使用检索服务时，登录后其用户身份可以被识别，然后再检索，数据的预处理对象选择用户登录日志，这就需要对日志依次进行清洗、用户识别、会话识别，处理成关联规则可以操作的字符串（检索词）集合。其中，用户识别采用了用户 ID 作为标识，会话识别采用通用的 30 分钟[10]。然后，用关联规则分析会话，挖掘出强关联规则存入数据库。当用户输入检索词与数据库中某条规则中的检索词相同，就向该用户推荐该条规则中其他检索词。

基于规则的检索词过滤是指用户在使用检索服务时，根据其检索词作为用户的兴趣，建立用户模型，并将相似用户聚类。一旦用户登录，可以向用户推荐同类用户感兴趣的其他检索词。涉及到的技术包括数据预处理、聚类和推荐。首先，从服务器日志中将检索词按照注册用户 ID 作为索引清洗出来，以<用户 ID，检索词 1，检索词 2，…，检索词 n>，以检索词为属性，形成用户向量。在聚类后，将聚类结果以<用户 ID，频繁检索词 1，频繁检索词 2，…，频繁检索词 n>存储到数据库中。当注册用户登录时，将该用户 ID 到数据库去匹配已经完成聚类建模的用户模型库，推荐相似的其他用户使用的频繁检索词。

3. 基于检索日志的检索词推荐模型研究

本文讨论的检索词推荐基于数据挖掘技术，着眼于用户在使用信息检索服务时记录在服务器日志中的检索信息。然而，目前的检索词推荐模型在用户使用效率上还存在改进空间。

文献[11]采用的检索推荐是基于用户一段时期内的检索请求形成的，虽然在一定程度上可以减轻用户的负担，但向用户推荐时的算法需要匹配的信息较多，耗时较长。CyrusShahabi 和 Yi-Shin Chen 设计了一种 Yada 系统[12][13]，它混合使用多种技术，如聚集、内容分析、协同过滤等。这个系统主要是采用两个阶段进行推荐服务的。在脱机处理阶段,Yoda 通过对客户端的用户使用信息进行聚集和内容分析，产生聚类的推荐列表，这种方法通过预处理工作有效地解决了大量数据处理耗时的问题。在联机处理阶段,Yoda 依据用户当前的访问行为，产生定制化的推荐信息。Yoda 系统这种无需用户提供很多附加信息，而是利用用户进一步访问形成的导航信息来调整可信度值的计算，固然在推荐精度上有一定优势[14]，但依然需要在推荐阶段进行一定复杂运算，牺牲了推荐时的用户效率。文献[15][16]也是在推荐阶段，每次都需要对检索词进行相似性计算。还有很多的检索推荐模型[17]-[22]中未说明挖掘和推荐是联机的，还是脱机的。

为了保证用户的体验，使检索效率不会因为新增加的推荐功能而出现明显下降，本文构建的基于检索日志的检索词推荐模型（见图 1），将推荐过程独立于其它过程——日志预处理和模式挖掘过程。与从图 1 中，可以看出，检索词推荐系统的预处理模块和模式挖掘模块都是通过脱机处理完成的，而用户检索词匹配和推荐过程是联机完成的。这种“脱机预处理和挖掘、联机推荐”的设计是为了尽可能保障推荐的用户体验，减少用户在使用推荐时的时间延迟，将主要耗时的工作安排到后台完成，使推荐服务不影响正常的检索服务。

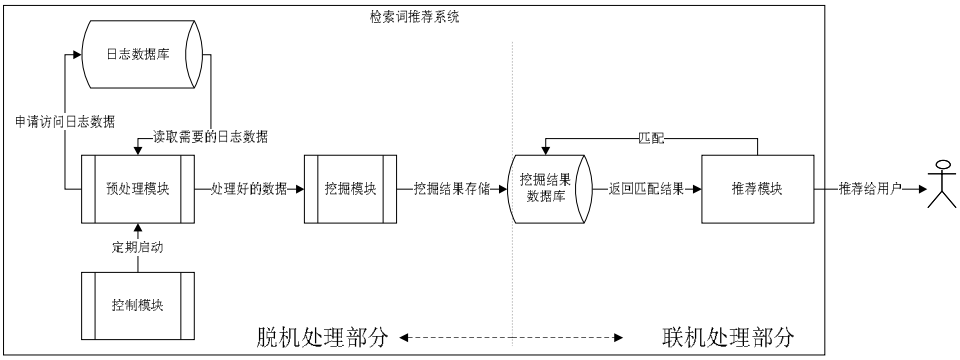

图 1 基于检索日志的检索词推荐模型

在基于检索日志的检索词推荐模型中，首先是控制模块定期触发预处理模块，经过日志数据的预处理，形成可供挖掘的数据格式，再通过机器学习的挖掘方法，建立起个性化推荐的模型（含关联的规则、关联的内容和相似兴趣的用户），并存储到数据库中；推荐模块可以单独有人工触发运行，也可以作为接口程序被外部程序调用，按照基于规则的、基于内容的和基于协作的三种模式分别去匹配推荐的模型，并返回匹配的检索词。

4. 关键算法的选择

本文关键算法的选择着眼于通用性和实用性，服务于上文提到的混合推荐方法和基于检索日志的检索词推荐模型，主要涉及到聚类算法和关联规则算法的选择。

4.1 聚类算法

文献[23]引入了两种特征来表示用户兴趣向量，可以节省大量运算时间,提高了聚类具有相似兴趣的用户的准确性。文献[24][25]在协同过滤中引入语义信息。但是，这几种寻找近邻的算法都需要提前在用户兴趣向量表示上做工作，在推广方面还存在一定局限性。吉雍慧在数字图书馆的检索推荐[26]中采用 K-means 算法计算相似度，具有较好的通用性。

本文选择 K-Means 算法对检索词的文本向量进行聚类，原因是该算法通用性较好，容易实现对大规模数据集的聚类[27]，且其简单直观和易于理解，应用广泛。K-means 算法是由 Mac Queen[28]提出，该算法是一种基于划分的聚类算法，它通过不断的迭代来实现聚类，当算法收敛到结束条件时就终止迭代过程，得出聚类结果[27]。但是，K-Means 算法[29]的缺点也比较明显，如需要预知类数目[30]。为了提高系统的自动化程度，减少人工干预，本文采用了文献[31]中的 BWP4 方法，借鉴文献[32]提出的类间类内划分(Between-Within Proportion，简称 BWP)指标，并修正了该指标，扩大了聚类数的搜索范围，自动确定最佳聚类数。

4.2 关联规则算法

蔡伟杰等人[33]将关联规则挖掘算法分为经典频集方法、其他频集挖掘方法、多层和多维关联规则的挖掘。考虑到作为经典频集方法的 Apriori 是一种较常见的关联规则算法[34]，且可以比较有效地产生关联规则[35],易于理解和实现，通用性强，本文采用 Apriori 算法进

行关联规则挖掘。Weka 的 Apriori 算法要对每一个会话的检索词集合向量化，导致一个向量的属性激增，且算法要考虑检索词集合中未出现的词是否出现时的频率，既增加了运行开销，又不方便之后的推荐匹配，因此，本文使用经典的 Apriori 算法[36]。

5. 实证

本文以 NSTL 嵌入式系统为实验平台，在原有的嵌入式系统中增加了图 1 所示的功能。实验过程中，邀请中国科学院国家科学图书馆的研究生参与使用该原型系统进行文献检索，历时 2 周，根据这些数据进行了实验研究，因此，控制模块的触发周期设为 14 天。使用的日志格式见本专辑“信息检索系统日志建模研究”一文，格式如表 1 所示。

表 1 NSTL 嵌入式系统的 Web 日志数据格式

序号	属性	属性值类型	说明	序号	属性	属性值类型	说明
1	User_ID	Varchar	用户标识	5	Operation_Status	Integer	操作状态
2	Operation_Type	Integer	操作类型	6	Object_Type	Integer	内容类型
3	Operation_Time	DateTime	操作时间	7	Object	Varchar	内容
4	IP	Char	IP 地址				

考虑到收集的用户数据多为学术期刊的简单检索，本文挖掘的日志数据在表 1 的基础上限定于学术期刊的简单检索，检索词推荐需要获得用户标识、操作类型（简单检索）、内容类型（学术期刊）、内容和操作时间，涉及到的日志数据如表 2。表 2 中的 Object 根据操作类型的不同，可以是文献标识号、检索词或关键词、或者订单号，本文需要使用期刊检索词，因此，Object_Type（即内容类型）选取 1，此时，Object 代表学术期刊，在从这个 Object. Object_Type（子内容类型）选取 5，此时，Object.Object 代表学术期刊的检索词。

表 2　检索词推荐涉及到的日志数据格式

序号	属性	说明
1	User_ID	用户标识
2	Operation_Type	操作类型
3	Operation_Time	操作时间
4	Object_Type	内容类型
5	Object	内容

笔者在原型设计实现中使用了是 JAVA、MySQL Server5.1 和 Weka3.6 等工具，硬件平台环境是 Intel（R）Core（TM）2 Duo CPU 2.4GHz 和 2G 内存，操作系统是 WindowXP。实验结果如下：

5.1 相似检索词推荐

从服务器日志中清洗出用户（含登录与非登录的用户）检索期刊时使用的 3308 个检索词，经过分类、去除停用词、去重和向量化，形成 123 个样本 187 维的数据集合，样本分布较为稀疏。BWP 方法是将取得最大 BWP 值时的类数作为最佳聚类数，从图 2 的实验结果可以看到，使用 BWP 方法，聚类搜索范围[2，Int（$\sqrt{n}$）]（n 为样本总数）,最佳聚类有 2 类，每类中的检索词相差甚远，没有起到聚类的作用；若放开聚类搜索范围，最佳聚类有 120 类，只得出 2 个有效类，各含两个样本，其他 119 个样本都分到一类中，聚类的效果较差；当使用改进后的方法 BWP4 时，最佳聚类数为 101 类，由于单一样本类不适用于推荐，本文主要考查非单一样本类，其聚类结果见表 3，找到 12 个非单一样本类，其中，除了类别

号为 2 的类外，其他 11 个类比较适合检索词推荐。如用户检索“数字图书馆”时，就会推荐“数字图书馆的理论与实践”作为其近似词，以便检索到更加精确的文献。经与人工分类相比，准确率达到 88.23%。由于用户输入的检索词越多、越相近，聚类效果就越好，考虑到有限的样本数量，目前的聚类效果已经比较明显。

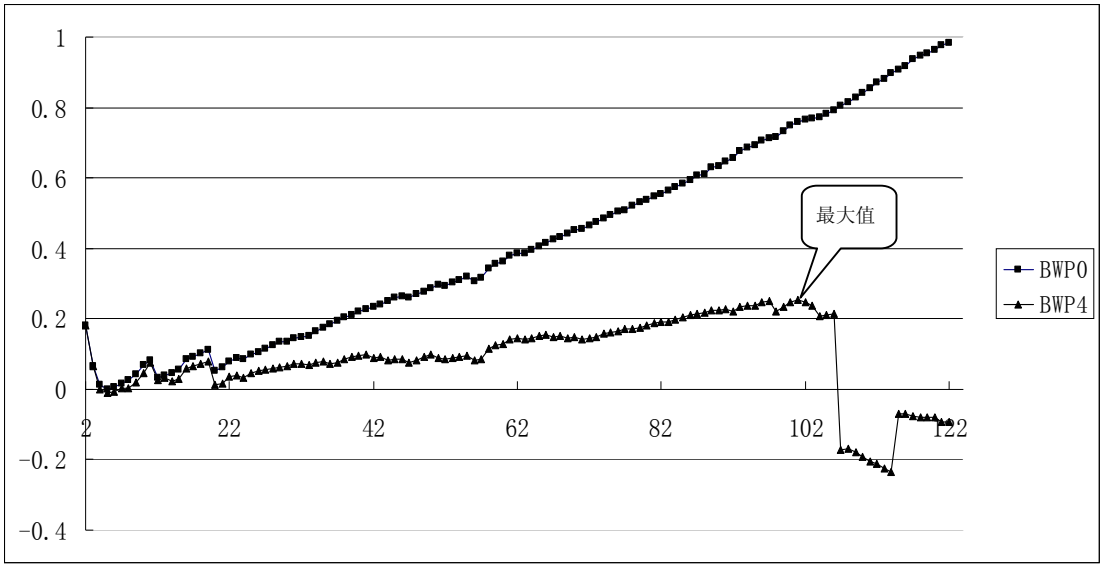

图 2 最佳聚类数指标 BWP 随类数的变化

表 3 相似检索词的聚类结果（非单一样本类）

序号	登录和非登录用户的检索词	类别	人工分类
1	数字图书馆发展趋势	1	1
2	数字图书馆发展趋势　最新	1	1
3	0926-3373	2	2
4	ontology	2	13
5	0956-7151	2	2
6	CN1387751	2	14
7	1573-3998	2	2
8	1871-5206	2	2
9	低碳	2	15
10	development	3	3
11	nano	3	16
12	ALL , [development] ;ALL , [nano] ;ALL , [ ] ;ALL , [ ]	3	3
13	computer	4	4
14	computer　administrator	4	4
15	computer　administrator111	4	4
16	e　gov	5	5

17	e gov 2010	5	5
18	JTITLE，[web] ;JTITLE，[search] ;JTITLE，[ ] ;JTITLE，[ ]	6	6
19	TITLE，[web] ;TITLE，[search] ;JTITLE，[ ] ;JTITLE，[ ]	6	6
20	TITLE，[nano] ;TITLE，[Development] ;JTITLE，[ ] ;JTITLE，[ ]	7	7
21	JTITLE，[nano] ;TITLE，[development] ;ALL，[ ] ;ALL，[ ]	7	7
22	ALL，[nano] ;TITLE，[development] ;ALL，[ ] ;ALL，[ ]	7	7
23	TEM STUDY OF DIFFUSIONAL α -Fe INTERFACE （ NANO-SCALE-CHARACTERIZATION	8	8
24	TEM STUDY OF DIFFUSIONAL α -Fe INTERFACE	8	8
25	JTITLE，[数字图书馆] ;AUTHOR，[李广建] ;ALL，[ ] ;ALL，[ ]	9	9
26	数字图书馆的理论与实践	9	9
27	数字图书馆	9	9
28	data mining	10	10
29	data	10	10
30	web crawler	11	11
31	crawler	11	11
32	web	12	12
33	web retrivel	12	12
34	web crwaling	12	12

相似检索词推荐的预处理和挖掘过程耗时 127 秒。

5.2 关联检索词推荐

78 条注册用户关于期刊类的检索记录，经过数据预处理，形成 48 个会话。由于实验数据不多，支持度分别设为 0.1，0.05，0.03，信度分别设为 0.1，0.01，得到强关联规则条数见表 4。

表 4 不同支持度、信度下的强关联规则数量

支持度	信度	强关联规则条数
0.1	0.1	2
0.1	0.01	2
0.05	0.1	2
0.05	0.01	2
0.03	0.1	5
0.03	0.01	8

为了尽可能获得最多的推荐规则，笔者选择支持度大于 0.03，信度大于 0.01，在该条件下可以找到 8 条强关联规则，具体见表 5。基于规则的过滤发现了基于内容的过滤所没有发现的检索词关系，如“library”和“web”之间的关系，这是相似检索词检索无法办到的。

表 5：检索词的强关联规则

序号	强关联规则	信度	支持度
1	crawling => web	1	0.041666666666666664
2	search => web	0.8571428571428571	0.125
3	web => search	0.2857142857142857	0.125
4	web => crawler	0.09523809523809523	0.041666666666666664
5	crawler => web	0.2857142857142857	0.041666666666666664
6	library => web	0.2222222222222222	0.041666666666666664
7	web => library	0.09523809523809523	0.041666666666666664
8	web => crawling	0.09523809523809523	0.041666666666666664

关联检索词的预处理时间小于 1 秒，挖掘过程耗时 1 秒。

5.3 同类用户检索词推荐

对注册用户的日志数据进行相似兴趣建模，使用 4.1 相同的聚类算法得到用户聚类的结果，见表 6，为保护隐私，隐去真实用户 ID，以用户序号代替。易见，同类别的用户经常使用相同或相似期刊类检索词，但其中也有不同的检索词可供推荐使用，如对使用“web crawler”检索的 10 号用户就会推荐“crawler”，以获得更广泛的检索结果。聚类结果显示，与人工分类相比，准确率达到 79.17%。

表 6 同类用户聚类结果

用户序号	用户使用过的检索词	类别	人工分类
1	web document document library web search web search web web search web retrivel web search web search web search library	0	0
2	Library	1	1
3	Library	1	1
4	Library	1	1
5	Library	1	1
6	air crawler web2.0 web web web library library	10	0
7	computer document 机构 库 development [Ljava.lang.String;@49481c web search web2 web search web2 web library	11	0
8	web web crawler	13	13
9	web crawler	13	13
10	search crawler crawler crawler crawler crawler crawler web web web web web service web service	14	13

11	文献 管理	12	12
12	web crawling	2	13
13	web crawling	2	13
14	computer computer	3	3
15	Computer	3	3
16	Computer	3	3
17	Computer	3	3
18	Web	4	4
19	Web	4	4
20	清华	5	5
21	Eclipse	6	6
22	ROUTING AND WAVELENGTH ASSIGNMENT OF SCHEDULED LIGHTPATH DEMANDS IN A WDM OPTICAL TRANSPORT NETWORK	7	7
23	data mining e gov e gov 2010 education computer data intergration data intergration 2010 data intergration 2009 data intergration 2009 ALL	8	8
24	web search	9	0

同类用户聚类的预处理和挖掘过程共耗时 60 秒。

表 7 个性化推荐的效率情况汇总表

推荐内容	相似检索词	关联检索词	同类用户检索词	三类检索词
工作内容	预处理+挖掘	预处理+挖掘	预处理+挖掘	推荐
耗时	127 秒	2 秒	60 秒	1 秒
准确率	82%	-	76%	-

从表 7 可以看出，混合推荐在原来 NSTL 嵌入式系统检索结果返回后 1 秒钟就可以完成，推荐取得了良好的效果。而包含预处理和挖掘过程在内，原型系统针对实验数据的一次整体脱机工作时间为 189 秒，但由于是脱机工作，所以不影响用户体验。

上述推荐结果的用户端展示，见本专辑“一种个性化的信息检索服务界面的设计与实现”一文。

6. 结语

本文在研究个性化信息服务理论的基础上，对基于内容的过滤、基于规则的过滤和基于协作的过滤三种典型的个性化推荐方法在检索词推荐方面进行分析，提出一种结果丰富的推荐方法，并针对现有检索推荐模型的不足，构建了适用于检索日志的检索词推荐模型。然后，对新提出的检索词推荐方法和推荐模型以 NSTL 嵌入式系统为实验平台进行实证研究，取得了良好的推荐效果。同时，原型系统由于采用了 BWP4 方法确定最佳聚类数，减少了人工干预，提高了原型系统的自动化水平，检验了 BWP4 方法的有效性，在原型系统的实验效果优于原 BWP 方法。但是，该原型的聚类结果仍然存在少量聚类错误的情况。下一步，可以进一步提升原型系统的建模准确性。

参考文献：

[1] 苏玉召,赵妍.个性化关键技术研究综述[J].图书与情报,2011,137(1):59-65.

[2] Billsus, D., Pazzani, M.: A personal news agent that talks, learns and explains[C]. In: Proc. 3rd Int. Conf. on Autonomous Agents (Agents ' 99), Seattle, WA (1999) 268 - 275.

[3] Rashmi Sinha, Kirsten Swearingen, Comparing recommendations made by online systems and friends[C], Proceedings of the DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries, Dublin, Ireland, June 2001.
[4] Forsati R., Meybodi M. R., Neiat A.G.. Web Page personalization Based on Weighted Association Rules[C]. International Conference on Electronic Computer Technology, 2009,130-135.
[5] Burke, R.: Hybrid Web recommender systems[C]. In Brusilovsky, P., Kobsa, A., Nejdl, W., eds.: The Adaptive Web: Methods and Strategies of Web Personalization, Lecture Notes in Computer Science, 2007,4321:377-408.

[6] Ardissono, L., Gena, C., Torasso, P., Bellifemine, F., Difino, A., Negro, B. User modeling and recommendation techniques for personalized Electronic Program Guides[C]. In: Personalized Digital Television. Targeting Programs to Individual Users. Kluwer Academic Publishers , 2004, 1:3-26.

[7] Nima Taghipour, Ahmad Kardan.A hybrid web recommender system based on Q-learning[G]. Proceedings of the 2008 ACM symposium on Applied computing, 2008:1164-1168.
[8] Wei Chen, Li-jun Zhang, Chun Chen, Jia-jun Bu. A Hybrid Phonic Web News Recommender System for Pervasive Access[G]. International Conference on Communications and Mobile Computing, 2009:122-126.
[9] 李秦，郑宏，基于用户行为的全文检索系统个性化研究[J]，图书馆杂志，2008，11（27）：25-28，34.
[10] Hiroshi Ishikawa, Manabu Ohta, Shohei Yokoyama, Junya Nakayama, and Kaoru Katayama，On The Effectiveness of Web Usage Mining for Page Recommendation and Restructuring[J]，Lecture Notes in Computer Science, 2003, Volume 2593/2003, 253-267.
[11] 张振亚,陈恩红,王 进,王煦法，RealCC在文本信息检索的个性化推荐中的应用研究[J]，数据采集与处理，2004，3（9）：338-342.
[12] Cyrus Shahabi,Yi Shin Chen.Web Information Personalization:Challenges and Approaches.In Proceedings of the Third International Workshop on Databases in Networked Information Systems,2003 2.
[13] Shahabi,C,Banaei-Kashani F,Chen Y S.,McLeod D.Yoda:An Accurate and ScalableWeb-based Recommendation System.In Proceedings of Sixth International Conference on Cooperative Information Systems,2001.
[14] 丁大可，李树青，徐侠，Web信息检索系统中的个性化技术[J],情报杂志,2006,10:51-53.
[15] 孙静宇,余雪丽,李鲜花，面向语义搜索的推荐模型研究[J]，广西师范大学学报:自然科学版,2008,9(3):202-205.
[16] 吉雍慧，数字图书馆中的检索结果聚类和关联推荐研究[J],现代图书情报技术,2008,(2):69-75.
[17] 吴良杰,刘红祥,张立堃,况振东，个性化服务中网页推荐模型的研究[J]，计算机应用研究，2005，6：83-85.
[18] 朱鲲鹏,刘文涵,王晓龙,刘远超,基于日志挖掘的检索推荐系统[J],沈阳建筑大学学报(自然科学版)，2009，2（3）：366-370.
[19] 李秦，郑宏，基于用户行为的全文检索系统个性化研究[J]，图书馆杂志，2008，11（27）：25-28，34.
[20] 董兵，吴秀玲，基于语义扩展的个性化知识推荐系统[J]，图书馆学研究，2008，11：44-49.
[21] 李珊，何建敏，厉浩，基于知识的协同过滤推荐系统研究[J]，情报学报，2008，27（3）：357-362.
[22] 马文峰,高凤荣,王珊,论数字图书馆个性化信息推荐系统[J],现代图书情报技术,2003,(2):16-18.

[23] 吴良杰,刘红祥,张立堃,况振东，个性化服务中网页推荐模型的研究[J]，计算机应用研究，2005，6：83-85.
[24] 董兵，吴秀玲，基于语义扩展的个性化知识推荐系统[J]，图书馆学研究，2008，11：44-49.
[25] 李珊，何建敏，厉浩，基于知识的协同过滤推荐系统研究[J]，情报学报，2008，27（3）：357-362.
[26] 吉雍慧， 数字图书馆中的检索结果聚类和关联推荐研究[J],现代图书情报技术,2008,(2):69-75.
[27] 谢娟英，蒋帅等，一种改进的全局K-均值聚类算法[J]，陕西师范大学学报（自然科学版），2010,38(2):18-22.
[28] MACQUEEN J. Some methods for classification and analysis of multivariate observations[C] Proceedings 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, 1967：281-297.
[29] 李飞，薛彬，黄亚楼，初始中心优化的K-M eans聚类算法[J]，计算机科学，2002,29(7):94-96.
[30] 姜园，张朝阳等，用于数据挖掘的聚类算法[J]，电子与信息学报， 2005,27(4):655-662.
[31] 边鹏，赵妍，苏玉召，一种改进的 K-means 算法最佳聚类数确定方法[J]，现代图书情报技术，2011(9):34-40.
[32] 周世兵,徐振源,唐旭清,K-means 算法最佳聚类数确定方法[J],计算机应用,2010,30(8)1995-1998.
[33] 蔡伟杰，张晓辉，朱建秋,朱扬勇，关联规则挖掘综述，计算机工程，2001，(5)：31-33，49.
[34] 蔡伟杰，张晓辉，朱建秋，朱扬勇，关联规则挖掘综述[J]，计算机工程，2001, 27(5):31-34.
[35] 胡吉明，鲜学丰，挖掘关联规则中 Apriori 算法的研究与改进[J]，计算机技术与发展，2006，(4)：99-104.
[36] AGRAWAL R, IMIELINSKI T, SWAMI A. Mining association rules between sets of items in large database[A]. In Proc. of the ACM SIGMOD Intl Conf. on Management of Data[C]. Washington, D.C., 1993:207-216.

作者简介：
边鹏，男，（1978--），中国科学院文献情报中心，中国科学院研究生院博士研究生，已发文 4 篇。
主要研究方向：网络信息管理技术与信息系统、Web 日志挖掘。
地址：北京中关村北四环西路 33 号
邮政编码：100190
联系邮箱：bianpeng@mail.las.ac.cn
电话：13641395632

苏玉召，男，（1975--），中国科学院文献情报中心，中国科学院研究生院博士研究生，已发文 5 篇。
主要研究方向：网络信息管理技术与信息系统、Web 日志挖掘。