文章编号：1003-0077(2009)01-0062-07

基于人工标注的个性化检索系统评测的研究

张 宇，范基礼，郑 伟，邹博伟，刘 挺

（哈尔滨工业大学 信息检索研究室 黑龙江 哈尔滨 150001）

摘 要：个性化信息检索可以根据用户的检索兴趣返回个性化的检索结果。该文构建了个性化检索标注系统和个性化检索评测系统，生成个性化检索系统所需的语料集；并提出了以用户为中心的基于人工标注的个性化检索评价方法。个性化检索评测系统采用了 NIST 所建立的评价体系，根据用户的标注结果对个性化检索系统的性能进行自动评价，并给出量化、直观的性能指标。

关键词：计算机应用；中文信息处理；个性化信息检索，以用户为中心，评价方法

中图分类号：TP391 **文献标识码**：A

Research on Evaluation of Personalized Information Retrieval Based on Manual Annotation

ZHANG Yu, FAN Ji-li, ZHENG Wei, ZOU Bo-wei, LIU Ting

(Information Retrieval Lab, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract: Personalized information retrieval can grasp the users' retrieval intention and find personalized results. A manual annotation system is designed in this paper to generate the corpus for evaluating personalized IR system. Then the User-centered manual annotation strategy is proposed for personalized IR evaluation. The evaluation system adopts the evaluation scheme provided by NIST performs an automatic evaluation according to the manually annotated results, and generates the quantified and straight-forward measurement results.

Key words: computer application; Chinese information processing; personalized information retrieval; user-centered; evaluation method

1 引言

随着网络的发展，Internet 上存在大量不同类型的信息资源，搜索引擎作为网络信息检索工具而被人们广泛使用，但是随着网络信息量的增加，面对用户个性化的信息需求，传统搜索引擎日益显现其不足，具体表现在以下几个方面：

（1）大部分搜索引擎是基于关键词匹配的，这种匹配完全是一种机械式的匹配，它既不能区分关键词的同形异义，也不能联想到相关同义词，更不可能考虑到关键词所处的特殊语言环境。因此，它不能有效地理解人们的查询请求。

（2）用户在选择关键词或构造查询表达式表达其信息需求时往往面临一些困难；面返回结果中，有许多无关信息，往往需要用户花大量时间浏览与选择。

（3）传统搜索引擎基本上都是“一个搜索适用所有用户”，对所有用户其检索结果都是一样，不能根据不同的用户给出相应的建议，无法满足用户的个性化需求。

由此可见，目前所广泛采用的信息检索技术无法满足不同背景、不同目的和不同时期用户的查询请求。Ask Jeeves 公司负责搜索和产品管理的副总裁拉哈尔—拉霍伊（Rahul Lahiri）称，“我们未来的产品不是由我们感觉用户需要什么而确定，而是由用户需要什么信息而确定。用户使用我们的产品是需要一种特定的答案，我们的重点也不再是增加诸

收稿日期：2008-08-21 **定稿日期**：2008-11-02

基金项目：国家自然科学基金资助项目（60736044，60503072，60435020）

作者简介：张宇（1972—），副教授、硕士导师，主要研究方向为自然语言处理、信息检索、自动问答；范基礼（1984—），本科，主要研究方向为信息检索；郑伟（1981—），硕士生，主要研究方向为信息检索。

如音频或是视频等新鲜的玩意儿，而是要提供用户真正需要的东西。”因此，如何提高搜索引擎检索结果的精度并向用户提供个性化服务已成为搜索引擎技术的一个新的发展方向和研究热点。

个性化信息检索是以用户为中心的信息检索技术，它获取以多种形式表达的用户需求（包括显式的、隐式的以及相关用户的需求），并综合利用这些用户信息，提高信息检索系统的性能。首先，不同的用户通过各种途径访问 Web 资源；其次，系统学习用户的特性，创建用户访问模型；最后，系统根据得到的知识调整服务内容，以适应不同用户的个性化需求[1]。个性化检索为不同用户提供不同的服务，并满足同一用户的不同时期的需求。个性化服务通过收集和分析用户信息来学习用户的兴趣和行为，从而实现主动推荐的目的。个性化服务技术能充分提高站点的服务质量和访问效率，从而吸引更多的访问者。

所以对个性化检索系统的建模效果及系统的评测、度量也是一个非常重要的问题。目前对个性化系统服务质量的评价，不同的系统采用不同的方式和测试数据，还没有一个通用的标准来客观的评价多个不同个性化系统服务质量的优劣。需要研究一种通用的性能指标并开发相应的 Benchmark 来评价各种不同的个性化检索技术。另外，目前的评价方法大多是基于人工评价的，缺乏自动评价的方法。所以，对个性化信息检索系统如何进行自动评价，也是目前需要解决的问题。本课题的研究不仅仅对个性化检索有着很重要的意义，而且对其他相关研究也有着较大的参考价值。

本文按照如下方式组织：第二节介绍个性化信息检索评测的相关研究；第三节介绍个性化检索评测系统的构建；第四节提出以用户为中心的信息检索评价方法；第五节给出了实验结果及分析，最后第六节总结全文并展望未来工作。

2　相关研究

由于个性化信息检索系统针对不同背景用户的查询会给出不同的检索结果，对于不同用户的检索结果很难给出统一的正确答案进行评价。因此个性化检索的评测一般都需要人工参与，通过用户对各个查询结果的正确性进行人工标注，综合用户对结果的正确性标注来评测个性化检索系统的性能。

（1）准确率评价方法。比萨大学的 Paolo Ferragina 和伊利诺伊大学香槟分校的 Xuehua Shen 分别在论文[2-3]中提到了使用排序靠前的检索结果的准确率（precision at N document，简称为 P@N）作为系统性能评测度量的人工评测方法。该评测方法只利用用户每次查询结果中排序靠前的结果来对系统性能进行评价，因此参与评测的用户只对自己每个查询返回的前 N 个结果的正确性进行标注。系统利用每个查询前 N 个结果中标注为正确的结果所占比例作为系统性能的评价指标，此值越高则说明系统性能越好。

P@N 方法与信息检索中普遍使用的准确率方法类似，其优点在于计算公式简单，根据普通用户使用检索系统的习惯只利用前 N 个结果的准确率可以突出重点而且减少了评测用户的工作量，实现起来较简单。此方法的缺点是用户的相关性标注缺乏指导、随意性较大。

（2）用户打分评测机制。汉诺威大学的 Paul Alexandru Chirita 在文献[4]中使用的利用用户对查询结果的打分来评测系统的性能也是使用比较广泛的方法，此方法仍然只对查询返回的前 N 个结果进行标注和评价。每个用户利用检索结果与自己所需信息的符合程度对前 N 个结果进行打分，最后利用所有用户对检索结果打分的平均值作为系统性能的评价。

用户给检索结果打分的评测机制由于可以将用户对结果的评价划分成很多不同的等级，并对每个等级事先给定将查询结果标注为此等级的详细依据，对用户的评价行为做出一定的指导，使用户的标注行为更加规范化。

（3）DCG 评测算法。麻省理工大学的 Jaime Teevan 在文献[5]中提出了利用人工对查询结果打分的方式结合 DCG（Discounted Cumulative Gain）公式来评测个性化检索系统的方法。此方法依据不同网页在检索结果中排序的不同给其赋予不同的重要度，排序越高的检索结果重要度越大，用户对有较高重要度的检索结果的打分对系统性能的影响也越大，因此利用 DCG 公式将用户对检索结果的打分与结果的排序位置结合，计算出的值作为系统性能的评测指标。

实际使用中，用户更加倾向于查看检索结果中排序靠前的网页，因此对于用户来说检索结果排序越靠前的网页对于系统性能的影响也越大，DCG 评测算法将用户对查询结果的打分与结果在系统检索中的排序结合的做法更加符合用户使用的习惯，对系统整体的评价更加符合实际情况。

3　个性化信息检索评测系统的建立

为了实现个性化检索系统的评价，首先需要让

评测用户采用特定的标注方式对结果进行标注，然后设计出合适的方法通过用户的标注计算出针对每个用户的系统性能指标，最后综合所有用户的指标给出系统整体性能。因此，需要解决以下问题：

(1) 用户的标注方式。个性化检索系统的特点使得它很难使用传统检索系统的测试集进行系统性能评测，因此需要采用用户参与的方式进行评测。评测的用户可以采用构建个性化测试集的方式进行人工参与。构建个性化评测集的方法是让每个用户根据自己的兴趣对固定的查询给出与此相关的文档集，这样每个查询就形成了针对不同用户的相关文档集。

(2) 综合不同用户标注结果评价系统。由于人工标注产生了针对不同用户的相关标注集合，因此怎样利用这些不同的相关标注集合来对系统进行评测是需要研究的一个问题。虽然个性化检索系统与传统检索系统评测方法不同，但是固定的一个用户给出了他的标注集合，针对一个用户对系统的评测方法和传统评测方法可以相同。因此我们可以首先针对每个用户和他给出的标注采用传统评测方法对系统进行评测，然后将针对不同用户对系统的评测指标综合起来作为系统的总体评价指标。

根据上述研究内容，需要构建两个子系统，如图1所示：个性化检索标注系统和个性化检索评测系统。其中个性化检索标注系统需要用户参与，系统会记录用户的检索行为和标注结果生成个性化检索系统所需的语料集；个性化检索评测系统会根据用户的标注结果对个性化检索系统的性能进行评价。

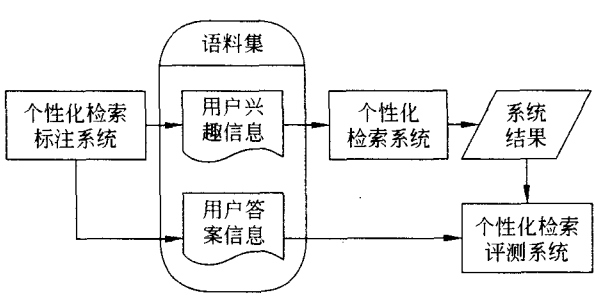

图1 个性化检索评测框图

本文对用户信息的收集分为两个方面，显式信息的采集和隐式信息的采集，如图2所示。

(1) 显式信息的收集，指由用户提供给系统来明确表达其兴趣、偏好、检索意图以及对检索结果做出的评价和反馈的信息。显式信息是由用户主观能动提供给系统的信息，这些信息可能包括：

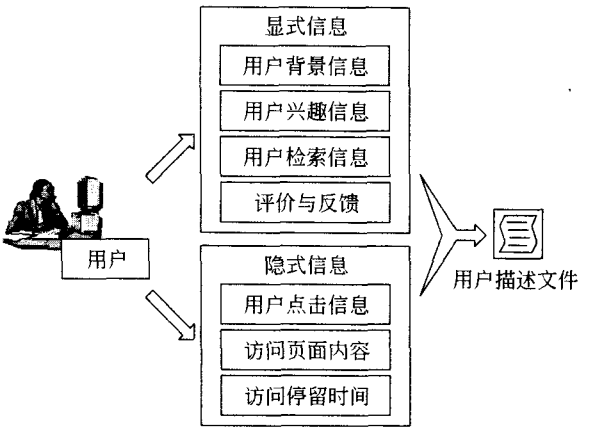

图2 个性化标注系统的用户信息采集

用户背景信息——主要包括用户提供的性别、年龄、学历、职业、收入等。这些信息有利于了解用户的兴趣背景，并针对某些特定的领域，使用统计信息对用户进行聚类或分类，来挖掘用户潜在的检索偏好和意图。

用户兴趣偏好——在特定的兴趣分类的体系框架下，由用户选定的自己感兴趣的信息类别。这种方法能够快速的收集用户的兴趣信息，比较准确地反映用户的需要和兴趣。

用户检索意图——用户的检索意图不再仅仅局限在一个关键字或几个关键字的逻辑组合，还可以引入更符合用户习惯的自然语言查询，增加用户表述其检索意图的途径，让用户提供尽可能多的语言信息。

评价与反馈——基于相关反馈的技术，通过用户对返回的部分结果进行标定来确定用户的兴趣，从而对搜索结果重新整理和排序。

(2) 隐式信息的收集，指通过对用户的浏览行为进行跟踪而得到的隐含信息。隐式信息的收集需要监视用户在Web页面的行为，如采集用户点击了哪些网页，在点击页面停留的时间、文档的长度、用户访问的URL地址、用户的翻页行为等数据，通过分析该日志文件总结出用户的特征数据，研究表明一定时间段的Web访问日志中蕴涵了用户的稳定兴趣。这种方法对用户透明，但用户数据的收集往往需要一段较长的时间。

4 以用户为中心的个性化信息检索评价方法

针对个性化信息检索的特点，可以利用用户人

工标注建立针对单个用户的标准评测集，利用用户标准评测集计算针对单个用户的系统性能指标，最后综合所有用户的系统指标计算出系统的总体性能评价指标。为了有效的评价检索系统，需要构造具有代表性的查询集。重点研究从大规模用户日志数据中进行有效采样，以描述尽可能全面和主流的用户查询需求。在此基础上，形成用户主题(Topic)，在主题中至少要包含用户查询及用户查询意图的详细描述。未来将根据用户查询意图进行相关性判断。

传统信息检索系统的评测需要利用经过人工标注的测试集作为评判标准，个性化检索系统的评测也可以利用人工标注的测试集来对系统的性能进行评测。但是由于传统的信息检索系统对不同用户的相同 query 检索结果固定不变，因此评测集的标注是依据同一标准综合不同标注者的标注结果对每个 query 给出固定的标准相关集。而个性化检索系统对于相同 query 返回给不同用户的结果是不同的，因此对测试集的标注和传统的标注方法有所不同，对于同一 query 需要每个标注者给出自己的标准相关集，之后分别利用每个标注者的相关文档集进行系统性能评测。

系统首先需要获取这些标注者的背景信息，包括显式和隐式获取两种方法。显式获取要求用户直接提交自己的兴趣爱好和背景信息，隐式获取要求用户事先使用系统一段时间使得系统可以通过用户的查询历史、浏览记录等获取用户的背景信息。用户的标注过程中如果直接让用户对测试集中的所有网页进行相关性标注会导致用户工作量太大。我们可以事先利用传统的检索系统在测试集中利用 query 进行检索，由于每个用户认为相关的网页必定是传统检索系统可以检索到的，每个用户的相关文档集必定包含在传统检索系统的检索结果中，因此可以直接让用户从传统检索系统检索出的结果中标注出自己的相关文档集。

通过上述方法构建了针对每个用户的相关文档集合后，依据每个标注者给出的标准相关文档集，利用评测公式对结果计算出针对每个用户的系统性能指标。个性化信息检索中对于单个用户的系统评测与传统信息检索的评测方法完全相同，可以采用传统信息检索的各项评价指标，如：错检率、漏检率、CDet 值等计算出针对单个用户的系统性能指标。

最后利用针对每个用户的系统性能指标的平均值作为整个系统的性能评价指标。图 3 给出了利用评测集对个性化信息检索系统评测的方法。

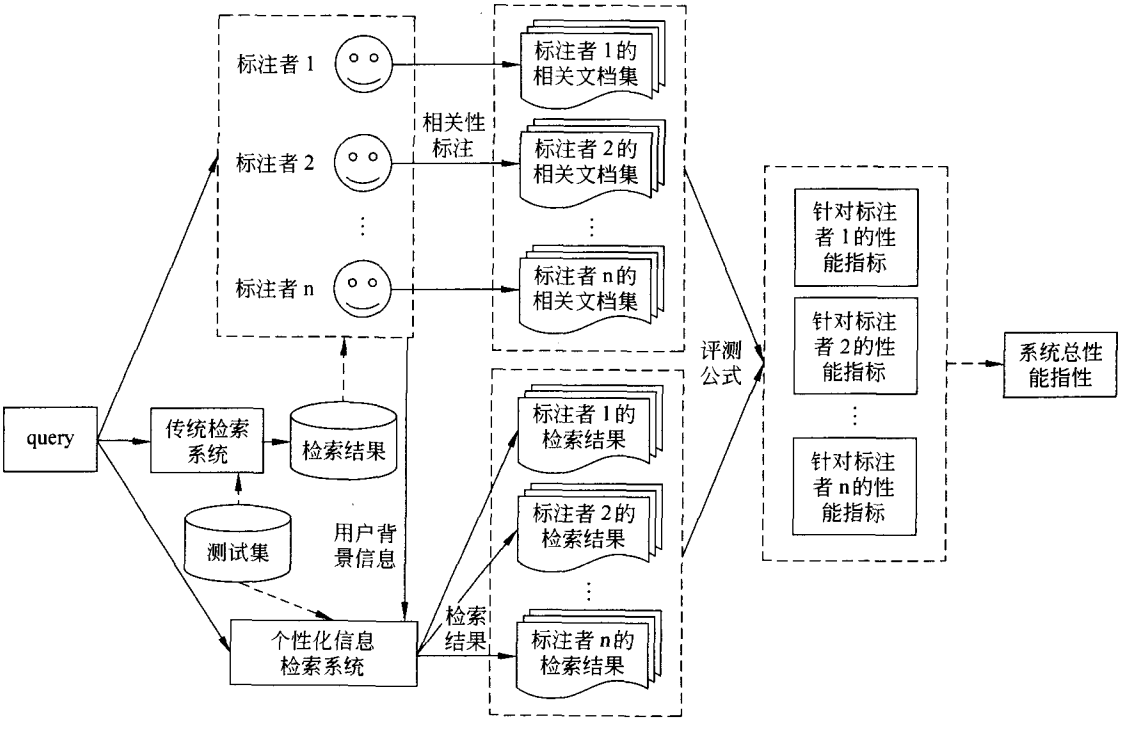

图 3　个性化信息检索评测框架

5 实验及结果分析

5.1 实验语料

针对个性化信息检索，我们建立了标准评测集，开发了基于天网100G语料的个性化评测语料标注辅助系统[6]，标注者利用此系统模拟正常的检索行为，系统记录下用户在检索过程中的各种隐式信息，包括query内容，检索结果，用户查看的结果网页，查看时间和翻页信息等。针对用户在每一个检索对象中进行的查询，按照检索对象的不同将query划分为不同段落，作为用户新兴趣发现任务的标准答案。同时，系统还将每个query返回的前20个网页，以及标注者在查询过程中点击过的网页提交给标注者，由标注者判断这些网页是否是检索目标的正确结果，标注后的结果作为个性化检索评价系统的标准答案。

利用个性化检索标注辅助系统，我们收集了9名标注者的标注结果。其中每个人对100个检索问题进行检索和标注，平均每个人进行了230次检索。用户兴趣发现任务的标准答案中，一个检索问题对应一个query段落，共100个query段落，每个query段落中query的平均个数为4.5个。本文将其中前50个query段落作为训练集，后50个query段落作为测试集。

5.2 评测方法

用户新兴趣发现任务的评价指标借鉴话题跟踪与检测(Topic Detection and Tracking，简称TDT)中的评价指标。因为在TDT评测中，Ptarget描述了正确答案在语料总数中的比例，可以更好的反映不同语料上的实验效果，而且DET曲线和(CDet)Norm值能够更准确地描述取不同相似度阈值时，系统性能的好坏。基于TDT2003的评测方法[6]，通过误检率和漏检率对系统性能进行评测。其计算公式如下：

$$P_{FA}=\frac{C}{A+C},\quad P_{Miss}=\frac{B}{B+D} \tag{1}$$

其中A、B、C、D如表1所示，A为系统判定是用户新兴趣的query且在标准答案也是新兴趣query的个数，B、C、D同理。P_{FA}、P_{Miss}是系统误检率和漏检率，值越小则系统性能越好。

表1 评测的参数

	系统判定相关	系统判定不相关
答案相关	A	B
答案不相关	C	D

之后，通过误检率和漏检率计算总的评价指标$(C_{Det})_{Norm}$，公式如下：

$$(C_{Det})_{Norm}=(C_{Miss}P_{Miss}P_{target}+C_{FA}P_{FA}P_{non-target})/\min(C_{Miss}P_{target},C_{FA}P_{non-target}) \tag{2}$$

其中，C_{Miss}是系统进行一次漏检的代价、C_{FA}是系统进行一次误检的代价，由于实际中，错误的段落划分和漏掉正确段落划分对个性化检索任务的影响基本等价，因此将C_{Miss}和C_{FA}都设为1；P_{target}是每个query为用户新兴趣的概率，$P_{non-targe}$是非新兴趣的概率，针对语料中的正确答案，将P_{target}和$P_{non-target}$分别设为0.435与0.565。$(C_{Det})_{Norm}$是系统性能损耗代价，此值越小则系统性能越好。

为了使系统性能得到更直观的体现，本文引入TDT的中的决策错误权衡曲线(Decision Error Tradeoff curve，简称DET曲线)评测系统性能。横坐标是误检率，纵坐标是漏检率，曲线越靠近图的左下角则性能越好，在图中还标出了最小性能损耗代价，此值越小则系统综合性能越好。个性化检索任务的评测采用相同的方法。

5.3 实验设计

个性化信息检索是以用户为中心的信息检索技术，它利用以多种形式表达的用户需求(隐式和显式信息)，个性化检索系统共分成：用户新兴趣发现、用户兴趣跟踪、相似用户群建立、个性化检索4个部分，每个部分既有相对独立的功能和输入输出，每一部分又是后一部分的输入，紧密联系成一个完整的系统，给出针对用户的个性化信息。

为了准确、详细的评价系统的性能，个性化检索评测系统也将会分成4个部分，分别针对个性化检索系统的4个部分进行评测，给出每个部分的独立性能评价。

5.3.1 用户新兴趣发现评测模块

用户新兴趣发现模块的主要功能：对query进行分析，发现用户新的检索需求，将检索对象相同的query划分为同一段落。例如用户依次输入query：数码相机、佳能相机、佳能A720、西藏旅游、进藏铁路，根据用户的检索需求，可以将数码相机、佳能相

机、佳能 A720 划分成一个段落，将西藏旅游、进藏铁路划分成一个段落，所以数码相机、西藏旅游就分别是用户两个新兴趣的开始，如图 4 所示：

图 4　新兴趣发现 query 划分图

用户新兴趣发现模块的输出(评测模块的输入)为：用户输入的 query 中所有相邻两个 query 的相似度。如图 5 所示，第一行为用户使用的 query 流，query 之间用 # 号隔开，第二行依次为相邻两个 query 之间的相似度。例，锻炼肌肉方法、哑铃锻炼方法两个 query 的相似度为 0.334 714。

5.3.2　用户兴趣跟踪评测模块

用户兴趣跟踪模块的主要功能：找到与当前 query 段落检索对象领域相同的其他段落。本模块为用户新兴趣发现模块的下一步工作，此时 query 流已经被划分成不同的段落，如图 6 中有 3 个 query 段落，系统会计算其中任两个段落之间的相似度，可知段落 3 和段落 1 有相同的检索需求，则段落 1 的历史信息能够帮助系统返回给用户更加合适的结果，同样能够为同一用户不同领域的兴趣分别建模、同一领域中相似用户群的建立提供帮助。

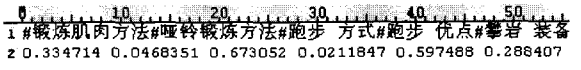

图 5　新兴趣发现输出数据图

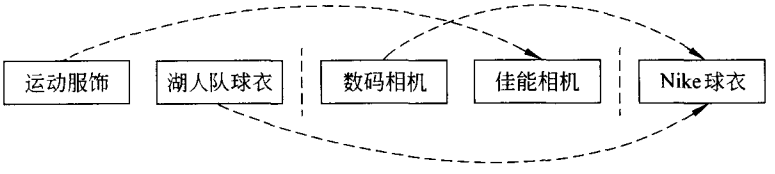

图 6　用户兴趣跟踪段落图

用户兴趣跟踪模块的输出(评测模块的输入)为：当前用户任两个 query 段落之间的相似度。如图 7 所示，第一行显示该用户的 query 段落数为 4，接下来 6 行中任两个 query 段落用 & 隔开，如段落 4.1.1 和段落 4.3.1 的相似度为 0.087 802 5。

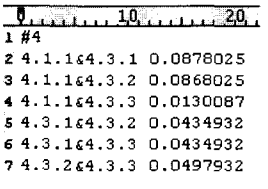

图 7　用户兴趣跟踪输出数据图

评测模块的算法描述：

(1) 读入用户新兴趣发现任务的结果值作为阈值集合。

(2) 针对当前阈值，大于等于阈值的 query 段落对判断为相关，对比正确答案，计算出该用户的错检率和漏检率。

(3) 针对当前阈值，计算出所有用户错检率和漏检率的平均值。

(4) 依次递增阈值，直至计算出在所有阈值时的用户的平均漏检率和错检率。

针对每一个阈值，系统都会得出相应的一对错检率(P_{FA})、漏检率(P_{Miss})，这时以错检率为横坐标、以漏检率为纵坐标，在图中画出所有阈值的点并连线，就得出了该模块的二维 DET 性能曲线。图 8 为用户新兴趣发现的评价性能曲线，用叉标记的点为

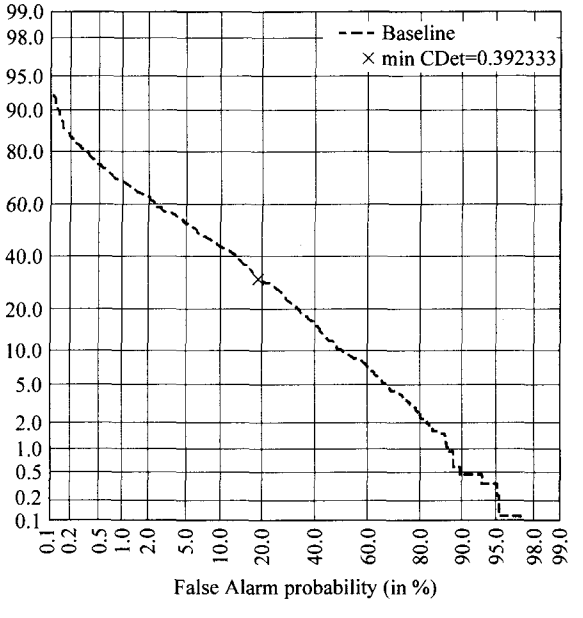

图 8　用户兴趣跟踪 DET 曲线

系统的最佳性能。

5.3.3 相似用户群建立评测模块

相似用户群模块的主要功能：在同一类别的检索对象下，找到与当前用户兴趣相同的用户，当此用户查询的此 query 块再次出现时，则可以针对此 query 块的用户群中其他用户的对应 query 块的隐式反馈信息用到此用户的当前 query 中。如图 9 所示，在旅游这一话题中，用户 1 和用户 2 具有相同的信息需求，则用户 2 的信息可以作为隐式反馈来辅助用户 1 的信息检索。

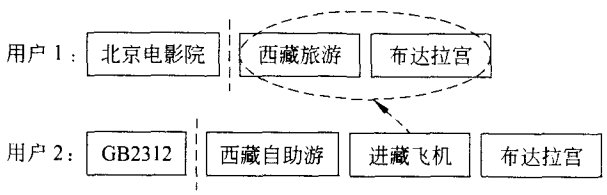

图 9 相似用户群 query 段落图

相似用户群的输出(评测模块的输入)为：在同一类别的检索对象下，任意两个用户之间的相似度。如图 10 所示，第一行为所有用户名，用户之间以空格隔开，余下的数据构成一个方阵，方阵的行数和列数依次对应不同用户，每一列数据表示其他用户和该用户的相似度。例 zw 和 hy 的相似度为 0.089 570 2。

```
0.....10.....20.....30.....
1 #hy thresh   zw   zwwayne
2 1 -0.0237798  0.0942146 0.121264
3 -0.0374968  0.990451  0.55056 0.393713
4 0.0895702 0.559464  0.995725  0.566256
5 0.133299  0.455338  0.563153  0.984518
```

图 10 相似用户群输出数据图

评测模块的算法描述：

(1) 读入用户新兴趣发现任务和用户兴趣跟踪任务的结果值作为阈值集合。

(2) 针对当前阈值，大于等于阈值的用户对判断为相关，对比正确答案，计算出该类别的错检率和漏检率。

(3) 针对当前阈值，计算出所有类别的错检率和漏检率的平均值。

(4) 依次递增阈值，直至计算出所有阈值的类别的平均漏检率和错检率。

针对每一个阈值，系统都会得出相应的一对错检率(P_{FA})、漏检率(P_{Miss})，这时以错检率为横坐标、以漏检率为纵坐标，在图中画出所有阈值的点并连线，就得出了该模块的二维 DET 性能曲线。图 11 为用户新兴趣发现的评价性能曲线，用叉标记的点为系统的最佳性能。

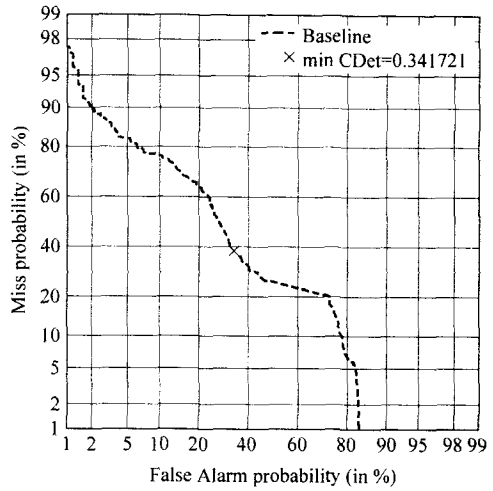

图 11 相似用户群 DET 曲线

5.3.4 个性化检索评测模块

评测模块的算法描述：

(1) 读入前三个任务的结果值作为阈值集合。

(2) 针对当前阈值，大于等于阈值的网页判断为相关，对比正确答案，计算出该 query 的错检率和漏检率。

(3) 再对 query 平均来计算每个检索对象的错检率和漏检率。

(4) 再对检索对象平均计算每个用户的错检率和漏检率。

(5) 最后对用户平均计算系统的错检率和漏检率。

图 12 为个性化检索的评价性能曲线，用叉标记的点为系统的最佳性能。

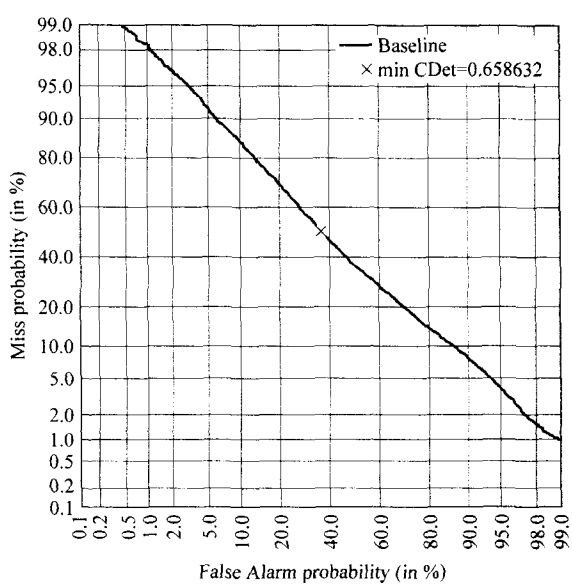

图 12 个性化检索 DET 曲线

(下转第 101 页)

2004). New York: ACM Press, 2004: 1～6.

[11] Ntoulas, A., Najork, M., Manasse, M., and Fetterly, D. Detecting spam web pages through content analysis. [C]//Proceedings of the 15th International Conference on World Wide Web (Edinburgh, Scotland, May 23-26, 2006). WWW '06. ACM Press, New York, NY, 2006: 83-92.

[12] Davison B. Recognizing nepotistic links on the Web. [C]//Artificial Intelligence for Web Search, pages 23--28. AAAI Press, July 2000. Presented at the AAAI-2000 workshop on Artificial Intelligence for Web Search, Technical Report WS-00-01.

[13] Amitay, E., Carmel, D., Darlow, A., Lempel, R., and Soffer, A. The connectivity sonar: detecting site functionality by structural patterns. [C]//Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia (Nottingham, UK, August 26-30, 2003). HYPERTEXT '03. 2003.

[14] Gyöngyi, Z., Garcia-Molina, H., and Pedersen, J. Combating web spam with trustrank[J]. In Proceedings of the Thirtieth international Conference on Very Large Data Bases-Volume 30. 576-587.

[15] Krishnan, V. and Raj, R. Web Spam Detection with Anti-Trust-Rank. [C]//the 2nd International Workshop on Adversarial Information Retrieval on the Web (Seattle, United States, August 2006). AIRWeb '06.

[16] Becchetti, L., Castillo, C., Donatol, D., Leonardi, S. and Baeza-Yates, R. Using Rank Propagation and Probabilistic Counting for Link Based Spam Detection. [C]//Proc. of WebKDD'06(Philadelphia, Pennsylvania, USA, August 20, 2006).

[17] Saracevic, T. 1995. Evaluation of evaluation in information retrieval. [C]//Proceedings of the 18th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Seattle, Washington, United States, July 09-13, 1995). E. A. Fox, P. Ingwersen, and R. Fidel, Eds. SIGIR '95. ACM Press, New York, NY, 138-146.

[18] Benczur, A., B? ro, I., Csalogany, K. and Sarlos T. 2007. Web spam detection via commercial intent analysis. [C]//Third International Workshop on Adversarial Information Retrieval on the Web (Banff, Alberta, Canada, May 8, 2007). AIRWeb '07. ACM Press, New York, NY, 89-92.

(上接第 68 页)

6　结论与未来工作

本文中，构建了个性化检索标注系统和个性化检索评测系统。其中个性化检索标注系统需要用户参与，根据系统给出的检索需求检索出答案，在检索过程中系统会自动记录用户的隐式信息和标注结果，生成个性化检索系统所需的语料集；个性化检索评测系统会根据用户的标注结果对个性化检索系统的性能进行自动评价。相比现有的个性化检索评价方法，本文中所采用的评价方法，对显示信息和隐式信息的获取更加全面、准确；使用了 NIST 建立的自动化评测方法，该评测标准是建立在检验系统漏检率和错检率的基础之上，不需要用户再进行干预，能够通过 CDet 值和 DET 曲线衡量系统性能。在下一步的研究工作中，可以继续研究个性化信息检索评估体系中的评价指标的量化，进一步增加语料的规模，实现大规模测试集的建立和自动评价体系。

参考文献：

[1] 李晓明，闫宏飞，王继民. 搜索引擎原理、技术与系统[M]. 科学出版社. 2005，212-215.

[2] P Ferragina, A Gulli. A Personalized Search Engine Based on Web Snippet Hierarchical Clustering[C]//International World Wide Web Conference, Chiba, 2005. New York, ACM Press, 2005: 801-810.

[3] X Shen, B Tan, CX Zhai. Implicit User Modeling for Personalized Search [C]//Proceedings of the 14th ACM international conference, Bremen, 2005. New York, ACM Press, 2005: 824-831.

[4] P. A. Chirita, W. Nejdl, R. Paiu, C Kohlsch tter. Using ODP Metadata to Personalize Search[C]//Proceedings of the 28th annual international ACM SIGIR, Salvador, 2005. New York, ACM Press, 2005: 178-185.

[5] J Teevan, ST Dumais, E Horvitz. Personalizing Search via Automated Analysis of Interests and Activities[C]//Proceedings of the 28th annual international ACM SIGIR, Salvador, 2005. New York, ACM Press, 2005: 178-185.

[6] R. W. White, J. M. Jose, C. J. van Rijsbergen I Ruthven. A simulated study of implicit feedback models[M]. Springer Berlin/Heidelberg, 2004: 311-326.