DOI:CNKI:11-2746/G2. 20111205. 1104. 002

交互式跨语言信息检索中用户行为研究*

吴　丹

摘　要　交互式跨语言信息检索是信息检索的一个重要分支。在分析交互式跨语言信息检索过程、评价指标、用户行为进展等理论研究基础上,设计一个让用户参与跨语言信息检索全过程的用户检索实验。实验结果表明:用户检索词主要来自检索主题的标题;用户判断文档相关性的准确率较高;目标语言文档全文、译文摘要、译文全文都是用户认可的判断依据;翻译优化方法以及翻译优化与查询扩展的结合方法在用户交互环境下非常有效;用户对于反馈后的翻译仍然愿意做进一步选择;用户对于与跨语言信息检索系统进行交互是有需求并认可的。用户行为分析有助于指导交互式跨语言信息检索系统的设计与实践。图4。表6。参考文献16。

关键词　跨语言信息检索　用户行为　用户交互　用户研究

分类号　G354

On User Behavior in Interactive Cross Language Information Retrieval

Wu Dan

ABSTRACT　Interactive cross-language information retrieval(Interactive CLIR) is an important topic in information retrieval. Based on reviewing the theoretic foundations of interactive CLIR, particular its process, evaluation metrics, and user behavior, we conducted a retrieval experiment that involves users in all stages of interactive CLIR. The study explores users' activities in six aspects, including query formulation, document selection, document examination, relevance feedback, translation re-selection, and others. The results show that query terms are mainly generated from the title of the search statement; users' relevance judgments are reliable; the full content of the original documents, their abstracts and their content translations all are the basis for users' judgments; translation enhancement and the combination of translation enhancement and query expansion are effective techniques in interactive search; users are willing to select new translations after feedback; and users have the requirement and satisfaction on the interaction with cross language information retrieval system. All these will benefit the design and the development of interactive CLIR systems. 4 figs. 6 tabs. 16refs.

KEY WORDS　Cross language information retrieval.　User behavior.　User interaction.　User study.

1　交互式跨语言信息检索概述

跨语言信息检索是信息检索的一个分支,指以一种语言查询检索出另一种语言文档信息的检索方法。跨语言信息检索与单语言信息检索一样,也是一个查询和文档匹配的过程,只不过查询和文档的表示语言不同。与单语言信息检索一样,在跨语言信息检索中,尽管用户的信息需求可用一个静态的查询来表达,但事实上,用户应该被集成到整个跨语言信息检索的过程中来,因为是用户提出并修改查询,同时,用户决定检索到的信息是否相关。因此,产生了一个重要的研究方向——交互式跨语言信息检索

*　本文系教育部人文社科研究项目"多语言信息获取中的用户相关反馈研究"(项目编号:09YJC870022)的研究成果之一。

通讯作者:吴丹,Email:danwoo@126. com

(Interactive Cross Language Information Retrieval，iCLIR)。在一个典型的跨语言信息检索过程中，从用户启动一个检索任务来满足其信息需求到用户找到相关信息满足了信息需求，这一包含用户的整个过程，被称为交互式跨语言信息检索[1]。

如图1所示，He 等[2]提出了一个完整的交互式跨语言信息检索过程。如6个正向箭头所指，查询表示、查询翻译、检索、文档选择、文档查看这5个步骤中，除了检索由系统自动完成外，其余4步均融入了用户的参与。同时，尽管查询是用户生成的，但用户往往不可能一次就构成一个正确的查询，而是需要经过多次检索才能明确其信息需求，因为用户的信息需求在检索过程中可能发生变化。这种情况下，在交互式跨语言信息检索中，可以利用反馈来进行重复优化(Iterative Refinement)。图1中，如6个逆向箭头所指，用户反馈可以形成6个路径，即在查询翻译完成后、文档选择之后、文档浏览完后均可以形成新的查询，文档选择之后、文档浏览完后均可重新选择翻译，文档浏览完后可以重新选择文档。

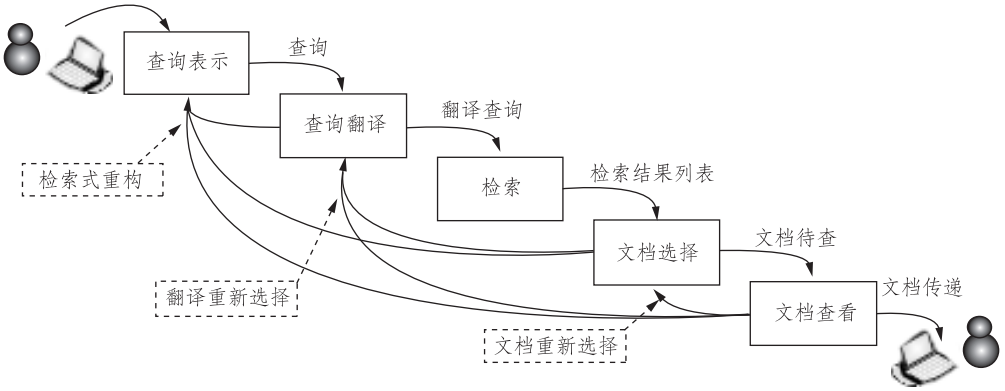

图1　交互式跨语言信息检索过程

在交互式信息检索评价方面，目前的检索任务已不仅仅是在交互式检索中如何检索到相关文献，而是更加注重关于检索者行为、用户满意度和检索过程等定性和定量的指标，使用了摄像、边做边说、过程录音、系统过程日志、详细的检索过程描述等方法[3]。此外，信息检索系统评价本身是一个相关性判断的过程，以往评价一个检索系统的性能时，所用的往往是二值相关性(Binary Relevance)标准，即将文档区分为要么相关要么不相关，并用相关文献作为标准去评价一个检索系统的查全率和查准率。然而，从用户角度来判断相关性时会发现，相关性往往是多维的，二值判断背离了现实情况。因此，多级相关性(Graded Relevance)被许多学者推荐作为交互式信息检索评价的新标准，即针对文档的相关性分为若干等级，如高度相关、一般相关、不相关等[4]。其中 Järvelin 等[5]提出的标准化折扣累积增量(Normalized Discounted Cumulative Gain，NDCG)就是一个基于用户的评价指标。该指标将用户的多级相关性判断因素考虑在内，能够用来衡量最相关文献在检索结果列表中是否优先被检出，其计算公式如下：

$$N_q = M_q \sum_{j=1}^{K} (2^{r(j)} - 1) / \log(1 + j) \qquad (1)$$

其中，M_q是一个常量，j 是检索结果的位置，r(j)是多级相关性判断的分值，例如，不相关0分，有一点相关1分，高度相关2分，等等。

2　交互式跨语言信息检索中用户行为研究进展

用户和系统交互对于跨语言信息检索很重要，因为它们产生了互补优势。系统适合做明确的、指定的、重复的工作，而人可以提供创造

性的、具有模式识别功能的贡献,因此,交互式跨语言信息检索一直以来都是研究热点。在理论研究上,国际上一些重要的信息检索和跨语言信息检索会议都将用户交互作为一个重要研究方向。国际著名文本检索会议 TREC 从 1995 年第四次会议(TREC-4)起开始设置“交互式研究”(Interactive Track)方向,致力于分析检索系统、检索者、检索主题分别对检索结果的影响。著名的跨语言信息检索会议——欧洲跨语言评价论坛(CLEF)自 2001 年起每年设立交互式主题(iCLEF),研究如何在用户辅助下提高查询翻译质量。在应用研究上,许多研究机构开发了交互式跨语言信息检索系统并进行了相关的用户实验。如美国 Maryland 大学的 MIRACLE 系统[2],荷兰 Twente 大学的 Twenty-One 系统[6],英国 Sheffield 大学的 Clarity 系统[7] 等,它们主要侧重在用户辅助查询翻译、用户辅助文档选择、用户辅助短语翻译等交互式功能方面。

在交互式跨语言信息检索中用户行为研究方面,国际上已有学者作了一些研究。如 Orengo 和 Huyck[8] 进行了英语和葡萄牙语间交互式相关反馈实验,由 27 位葡萄牙语志愿者分别对检索结果列表的前 10 篇文档的三种形式(英文原文、由 SYSTRAN 机器翻译系统翻译的葡萄牙文文档、人工翻译的葡萄牙文文档)进行二值相关性判断,再由系统进行查询扩展。实验结果显示,用户对机器翻译和人工翻译的文档来进行相关性判断时,其结果的准确性高于由用户直接对原文进行判断的方式,用户对人工翻译的文档和对机器翻译的文档进行相关性判断时所能达到的准确性几乎一样。Ostenero 等[9] 开发了一个西班牙语和英语的交互式跨语言检索系统 UNED,并研究了交互式跨语言信息检索过程中用户在短语翻译(Phrase Translation)和短语反馈(Phrase Feedback)时的行为特征。Oard 等[10] 在其开发的交互式跨语言信息检索系统 MIRACLE 上进行了用户辅助查询翻译和用户辅助文档选择的研究,主要分析了用户在系统的帮助下进行查询翻译和文档选择过程中的行为特征。Petrelli 等[6] 分析了在设计一个以用户为中心的跨语言信息检索系统过程中用户的需求、任务、系统界面测试等全过程,旨在设计一个以用户为中心的跨语言信息检索系统。国外在系统开发方面研究较多,而在深入分析用户行为方面略显不足。国内对于交互式跨语言信息检索的研究还很少,用户行为研究方面就更缺乏了。

综观上述研究进展,尽管有一些研究工作分析了交互式跨语言信息检索,但大部分与跨语言信息使用相关的挑战还没有解决。目前这方面的研究主要集中在开发帮助用户选择翻译工具上,而对用户在交互式检索过程中对信息需求、信息系统、文档集的正确理解方面还比较欠缺。鉴于此,本文将通过一个用户全程参与的跨语言信息检索实验来研究交互式跨语言信息检索全过程中的用户行为。

3 交互式跨语言信息检索的用户实验设计

本研究设计了一个让用户参与跨语言信息检索全过程(包括如图 1 所示的正向检索过程与逆向反馈过程)的实验,以了解用户在交互式跨语言信息检索中的真实行为。

3.1 实验资源

实验所用系统是研究人员自己开发的一个交互式跨语言信息检索系统 ICE-TEA[11],界面如图 2 所示。该系统可以让用户操作的地方包括:输入查询、选择词典的累积概率阈值、进行翻译并检索、对检索结果进行多级相关性判断、翻译优化、翻译优化后重新选择翻译、查询扩展、查询扩展后重新选择扩展词、先翻译优化再查询扩展、查看检索结果及摘要和全文等。

需要指出的是,翻译优化与查询扩展是我们实现的两种跨语言信息检索相关反馈方法。翻译优化[12] 是根据用户判断的相关文献及其译文(用户通常在译文上进行判断),从“相关文献对”中抽取检索词及其翻译关系,通过估计检索词在相关文献对中的翻译概率来改进初始查询翻译,以使查询中检索词的翻译更加贴近用户当前的检索。改进包括:修改初始查询翻译权值,给予相关文献对中所出现翻译更高的权重,

或者引入初始词典中没有、但出现在相关文献对中的翻译，或者去除初始翻译中噪音较大的翻译等。查询扩展[13]是在原始查询式中根据特定算法增加一些相关检索词。按照发生在多语言信息检索过程的先后位置，查询扩展方法可以分为翻译前查询扩展、翻译后查询扩展，以及翻译前与翻译后查询扩展的结合。本系统实现的是翻译后查询扩展，它是三种跨语言查询扩展方法中效果最好的[14]。

实验所用语料库为语言资源联盟（Linguistic Data Consortium，LDC）开发的 TDT4 和 TDT5 文本语料中的 83，627 篇中文文档，来源于新华社、联合早报等 11 个新闻机构[15]。

图 2　交互式跨语言信息检索系统 ICE-TEA

笔者直接从系统所用的 TDT4 和 TDT5 语料库中选取了 9 个英文检索主题，供用户检索实验用。图 3 是其中一个检索主题示例，每个检索主题包含编号、标题、事件元素（What，Who，Where，When，4W）、主题描述（Topic Explication）、主题解释（On Topic）。

41012. Trouble in the Ivory Coast

Seminal Event

WHAT：Presidential election

WHO：Laurent Gbagbo，Alassane Ouattara，Ivory Coast voters

WHERE：Ivory Coast

WHEN：October 25，2000

Topic Explication

On October 25，Laurent Gbagbo，head of the Ivorian Popular Front，declared himself president，as early polls showed him in the lead. Alassane Ouattara called the election unfair，but then conceded，though tens of thousands of his supporters took to the streets.

On topic：A recent history of power struggle that led to the current election. Disputes concerning the election including violence by the opposition groups.

图 3　检索主题示例

3.2 实验对象

在实验中,笔者从武汉大学信息管理学院找到愿意参加实验的志愿者54名。实验开始之初,通过问卷了解其背景,以便反映实验的代表性。具体统计结果如下:

①参与者中男性占40.7%,女性占59.3%。

②参与者中90.7%的专业是文科,7.4%的是理科,1.9%的是其他专业。

③参与者中年龄在20岁以下占3.7%,20-25岁占79.6%,26-30岁占13%,30岁以上占3.7%。

④参与者中本科学历占85.2%,硕士占7.4%,博士占7.4%。

⑤参与者中38名通过了大学英语四级考试,37名通过了大学英语六级考试。

⑥参与者中90.7%上过《信息检索》相关课程,9.3%未上过。

⑦参与者中1.8%对"跨语言信息检索技术"非常了解,85.2%对该技术仅听说过但了解不深,13%完全没有听说过该技术。

⑧参与者中7.4%每天上网时间低于1小时,44.4%每天上网查资料的时间为1-2小时,35.2%上网时间为2-3小时,13%上网时间达3小时以上。

⑨参与者中61.1%经常使用搜索引擎,并对搜索引擎非常熟悉;38.9%偶尔使用搜索引擎,并对搜索引擎比较熟悉。

⑩参与者中37%没有使用过跨语言搜索引擎,61%使用过Google Translated Search,2%使用过除Google外的其他跨语言搜索引擎。

⑪参与者中13%认为自己对搜索引擎的检索结果能够准确判断,并非常有信心;85.2%认为自己对搜索引擎的检索结果基本能够判断,并比较有信心;1.8%认为自己对搜索引擎的检索结果不能判断,没有信心。

⑫参与者中46.4%每天花在收听、阅读或观看新闻上的时间低于1小时,50%所花时间为1-2小时,1.8%所花时间2-3小时,1.8%所花时间在3小时以上。

综合而言,本次实验参与者的特征在性别上基本平衡,在专业上以文科背景用户为主,在年龄上以20-30岁的用户为主,在学历上以本科生为主。参与者的知识在掌握英语能力上基本没有问题,在信息检索能力上受过专业培训,对跨语言信息检索技术基本了解。参与者的上网时间普遍在每天1-3小时,在搜索引擎的使用上非常熟悉,在跨语言搜索引擎的使用上大部分用过Google Translated Search,对检索结果的判断比较有信心,在新闻语料所花的时间上基本为每天0-2小时。

3.3 实验方法和过程

实验用户的任务是在规定时间内,对笔者提供的检索主题分别用三种方法(基准跨语言信息检索、翻译优化、翻译优化与查询扩展的结合)进行跨语言信息检索,并对检索结果进行相关性判断与反馈。本实验采取三级相关性判断方法:高度相关、一般相关、不相关。

笔者采取"用户内实验设计"(Within-Subject Design),即每个用户均用相同的9个检索主题进行检索,且每个用户均用3种方法:Baseline,即没有任何相关反馈的基准跨语言信息检索;TE,即在基准跨语言信息检索基础上,进行翻译优化;Combined,即在基准跨语言信息检索的基础上,进行翻译优化与查询扩展的结合。其中,每种方法均用3个检索主题进行检验。为了避免检索主题顺序及检索方法顺序所造成的影响,在实验中采用如表1所示的拉丁方阵方法对每个用户的检索主题、所用方法及检索顺序进行轮排。9个检索主题轮转一圈有9种方式,3种方法有6种全排列,共有$9\times6=54$种组合。因此,54个用户尽管所用的检索主题和系统一样,但其顺序均不同。

在用户实验过程中,笔者利用后端程序对整个实验过程进行了日志记录。每个参与者的实验历时120分钟,过程如下:①听取ICE-TEA交互式英汉跨语言信息检索系统功能及使用方法的介绍,并填写一个关于其背景资料调查的问卷——5分钟;②用检索主题40004进行系统试运行——5分钟;③用第一种方法检索3个检索主题——每个检索主题10分钟,共30分钟;④填一份调查问卷,对第一种方法进行评

表 1　用户对应的检索主题

用户编号	检索主题编号,所用方法,检索顺序								
s1	1b	2b	3b	4t	5t	6t	7c	8c	9c
s2	1t	2t	3t	4c	5c	6c	7b	8b	9b
s3	1c	2c	3c	4b	5b	6b	7t	8t	9t
s4	2b	3b	4b	5t	6t	7t	8c	9c	1c
s5	2t	3t	4t	5c	6c	7c	8b	9b	1b
s6	2c	3c	4c	5b	6b	7b	8t	9t	1t
s7	3b	4b	5b	6t	7t	8t	9c	1c	2c
s8	3t	4t	5t	6c	7c	8c	9b	1b	2b
s9	3c	4c	5c	6b	7b	8b	9t	1t	2t
s10	4b	5b	6b	7t	8t	9t	1c	2c	3c
s11	4t	5t	6t	7c	8c	9c	1b	2b	3b
s12	4c	5c	6c	7b	8b	9b	1t	2t	3t
s13	5b	6b	7b	8t	9t	1t	2c	3c	4c
s14	5t	6t	7t	8c	9c	1c	2b	3b	4b
s15	5c	6c	7c	8b	9b	1b	2t	3t	4t
s16	6b	7b	8b	9t	1t	2t	3c	4c	5c
s17	6t	7t	8t	9c	1c	2c	3b	4b	5b
s18	6c	7c	8c	9b	1b	2b	3t	4t	5t
s19	7b	8b	9b	1t	2t	3t	4c	5c	6c
s20	7t	8t	9t	1c	2c	3c	4b	5b	6b
s21	7c	8c	9c	1b	2b	3b	4t	5t	6t
s22	8b	9b	1b	2t	3t	4t	5c	6c	7c
s23	8t	9t	1t	2c	3c	4c	5b	6b	7b
s24	8c	9c	1c	2b	3b	4b	5t	6t	7t
s25	9b	1b	2b	3t	4t	5t	6c	7c	8c
s26	9t	1t	2t	3c	4c	5c	6b	7b	8b
s27	9c	1c	2c	3b	4b	5b	6t	7t	8t
s28	1b	2b	3b	4c	5c	6c	7t	8t	9t
s29	1t	2t	3t	4b	5b	6b	7c	8c	9c
s30	1c	2c	3c	4t	5t	6t	7b	8b	9b
s31	2b	3b	4b	5c	6c	7c	9t	9t	1t
s32	2t	3t	4t	5b	6b	7b	8c	9c	1c
s33	2c	3c	4c	5t	6t	7t	8b	9b	1b

续表

用户编号	检索主题编号,所用方法,检索顺序								
s34	3b	4b	5b	6c	7c	8c	9t	1t	2t
s35	3t	4t	5t	6b	7b	8b	9c	1c	2c
s36	3c	4c	5c	6t	7t	8t	9b	1b	2b
s37	4b	5b	6b	7c	8c	9c	1t	2t	3t
s38	4t	5t	6t	7b	8b	9b	1c	2c	3c
s39	4c	5c	6c	7t	8t	9t	1b	2b	3b
s40	5b	6b	7b	8c	9c	1c	2t	3t	4t
s41	5t	6t	7t	8b	9b	1b	2c	3c	4c
s42	5c	6c	7c	8t	9t	1t	2b	3b	4b
s43	6b	7b	8b	9c	1c	2c	3t	4t	5t
s44	6t	7t	8t	9b	1b	2b	3c	4c	5c
s45	6c	7c	8c	9t	1t	2t	3b	4b	5b
s46	7b	8b	9b	1c	2c	3c	4t	5t	6t
s47	7t	8t	9t	1b	2b	3b	4c	5c	6c
s48	7c	8c	9c	1t	2t	3t	4b	5b	6b
s49	8b	9b	1b	2c	3c	4c	5t	6t	7t
s50	8t	9t	1t	2b	3b	4b	5c	6c	7c
s51	8c	9c	1c	2t	3t	4t	5b	6b	7b
s52	9b	1b	2b	3c	4c	5c	6t	7t	8t
s53	9t	1t	2t	3b	4b	5b	6c	7c	8c
s54	9c	1c	2c	3t	4t	5t	6b	7b	8b

注:表中 1 -9 代表 9 个检索主题,b = Baseline,t = TE,c = Combined。

价——5 分钟;⑤用第二种方法检索 3 个检索主题——每个检索主题 10 分钟,共 30 分钟;⑥填一份调查问卷,对第二种方法进行评价——5 分钟;⑦用第三种方法检索 3 个检索主题——每个检索主题 10 分钟,共 30 分钟;⑧填一份调查问卷,对第三种方法进行评价——5 分钟;⑨最后对三种方法的实验结果进行比较评价,填一个调查问卷——5 分钟。

4 交互式跨语言信息检索实验的用户行为分析

完成实验后,笔者对交互式跨语言信息检索过程中的查询形成、文档选择、文档查看、相关反馈、翻译重新选择和其他 6 个方面的用户行为进行分析。

4.1 用户的查询形成分析

参与者在每种检索方法上做三个检索主题,每做完一个检索主题,即让其对该检索主题进行分析,每个检索主题共有三个相同的问题:①实验前对该检索主题是否熟悉?(1 分"完全不熟悉";2 分"不熟悉";3 分"一般";4 分"熟悉";5 分"非常熟悉");②语言问题是否影响你对该主题检索结果的相关性判断?(1 分"完全不影响";2 分"基本不影响";3 分"还好";4 分

“有一点影响”;5 分“非常影响”);③对于该检索主题,你输入的检索词主要来源于(可多选)?(A. 所给材料的标题(Title);B. 所给材料的事件元素(4W);C. 所给材料的主题描述(Topic Explication);D. 所给材料的主题解释(On Topic);E. 没有基于所给材料,是自己对该主题的理解)。笔者对用户回答按照检索主题进行统计,结果见表 2。

从表 2 看出:①对于问题一,9 个检索主题的差别不大,基本平均得分都在 2 - 3 分之间,说明参与者对笔者所选检索主题的熟悉程度一般,且不会由于用户对检索主题熟悉程度的差异而影响其检索。②对于问题二,由于参与者的母语为中文,因此笔者担心会因为语言而影响其相关性判断,但是用户的回答打消了这一顾虑。9 个检索主题的平均得分都在 2 - 3 分之间,说明参与者对自己的英语能力比较自信,认为语言基本上不影响其检索和判断这些检索主题。③对于问题三,由于是多选题,笔者计算每个选项被选的次数。9 个检索主题的结果也非常一致:选项 A,即 TDT 语料检索主题的标题(Title)是用户产生查询词的主要来源;选项 B、C、D 其次,且较均衡,即所给材料的事件元素、主题描述和主题解释是用户产生查询词的有益补充;而选项 E 被选次数非常少,说明用户在构造查询时还是以所给材料为依据。

表 2　用户对检索主题打分

检索主题编号	问题 1:对检索主题的熟悉度(平均分)	问题 2:语言是否影响该主题判断(平均分)	问题 3:检索词的来源(选项被选次数)
40007	2. 37	2. 85	A:41　B:34　C:34　D:23　E:2
40019	2. 56	2. 44	A:42　B:34　C:24　D:21　E:1
40028	2. 85	2. 57	A:46　B:33　C:23　D:18　E:2
40039	2. 15	2. 63	A:45　B:23　C:21　D:24　E:2
40043	2. 51	2. 60	A:42　B:33　C:18　D:23　E:1
41018	2. 39	2. 81	A:46　B:34　C:24　D:19　E:2
41025	2. 26	2. 78	A:42　B:28　C:30　D:22　E:1
41027	2. 57	2. 80	A:44　B:30　C:29　D:26　E:1
41035	2. 57	2. 63	A:47　B:30　C:23　D:23　E:2
平均	2. 47	2. 68	A:44　B:33　C:25　D:22　E:2

此外,笔者还对用户输入的查询长度进行了统计,得到用户输入查询的平均长度为 5 个词。通常,在自动反馈实验中,研究者会按照类似 TREC 检索主题中的主题标题(Title)、主题问题(Description)、主题描述(Narrative)字段生成几种不同长度的查询:由 Title 字段组成的短查询,称为 T;由 Title 和 Description 字段组成的中查询,称为 TD;由 Title、Description、Narrative 字段组成的长查询,称为 TDN。在本实验中,笔者将 TDT 的检索主题改写成 TREC 的检索主题格式:将原 TDT 主题的编号作为 TREC 主题中的 number 字段,将原 TDT 主题的标题作为 TREC

主题中的 Title 字段,将原 TDT 主题的事件元素作为 TREC 主题中的 Description 字段,将原 TDT 主题的主题描述与主题解释作为 TREC 主题中的 Narrative 字段。经统计,得到短查询 T 的平均长度为 4 个词,中查询 TD 的平均长度为 27 个词,长查询 TDN 的平均长度为 127 个词。笔者将用户实际输入的查询长度与自动生成的查询长度进行比较,结果见图 4。可见,真实用户输入的查询通常较短,与检索主题的标题长度基本一致。因此,在进行模拟用户参与的信息检索实验时,用检索主题的标题来查询通常比较接近真实的用户检索行为。

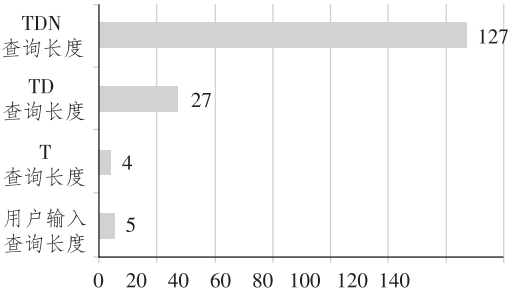

图 4 用户输入查询与自动生成查询长度比较

综上得出结论,在查询形成方面的用户行为特征有:用户平均输入的查询长度为 5 个词,检索词主要来自检索主题的标题。

4.2 用户的文档选择分析

笔者利用 TDT 语料库的“标准答案”(Ground-Truth)对用户判断的结果进行评价。由于是多级相关性判断,笔者采用严格相关(Strict Relevance)评价和松散相关(Loose Relevance)两种评价方法[16]。严格相关评价是指,将“一般相关”的文献当作“不相关”的文献来计算;松散相关评价则相反,是将“一般相关”的文献当作“相关”的文献来计算。评价的指标是计算判断准确率(P)、判断完全率(R),以及 F 均值,计算方法如下:

$$P = \frac{|S_{user} \cap S_{truth}|}{|S_{user}|} \quad (2)$$

$$R = \frac{|S_{user} \cap S_{truth}|}{|S_{truth}|} \quad (3)$$

$$F_a = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}} \quad (4)$$

其中,S_{user}是用户判断的相关文献集合,S_{truth}是“标准答案”给定的相关文献集合,α 为权重,在本实验中取值为 0.8,称为 F08。F08 更强调判断准确率 P。实验结果见表 3。

表 3 三种方法与所得检索结果的相关性判断

		Baseline	TE	Combined
松散相关	Precision	0.7322	0.7310	0.6941
	Recall	0.2436	0.4321	0.4924
	F08	0.4675	0.5861 *	0.5900 *
严格相关	Precision	0.8530	0.8378	0.8522
	Recall	0.1638	0.2353	0.2353
	F08	0.4019	0.4921	0.4842 *

* 表示具有统计性显著差异,且 0.01 < p - value < 0.05。

对于“松散相关”,与使用基准跨语言信息检索方法(Baseline)的相关性判断相比,使用翻译优化方法(TE)或使用翻译优化与查询扩展相结合的方法(Combined)的用户在相同给定时间内,分别选择了明显更多的相关文档(判断完全率 R 值较高),同时选择相关文档的准确率保持不变(判断准确率 P 值类似),因此与基准跨语言信息检索相比,翻译优化和翻译优化与查询

扩展的结合都在 F08 上取得统计上的显著提高。对于“严格相关”，同样与使用基准跨语言信息检索方法的相关性判断相比，使用翻译优化方法或使用翻译优化与查询扩展相结合的方法的用户在相同给定时间内，分别选择了明显更多的相关文档（判断完全率 R 值较高），同时选择相关文档的准确率保持不变（判断准确率 P 值类似），但是只有翻译优化与查询扩展的结合和基准跨语言信息检索的 F08 值之间有显著差异，翻译优化和基准跨语言信息检索的值之间没有显著差异。这个结果说明，翻译优化方法和翻译优化与查询扩展的结合方法使用户有更多机会发现相关文献。

综上得出结论，在文档选择方面的用户行为特征主要有：用户判断文档相关性的准确率较高。

4.3　用户的文档查看分析

在所有检索实验完成之后，笔者让用户对 ICE-TEA 系统进行综合评价，表 4 显示了这部分问题的平均得分。

表 4　用户对系统的综合评价

No.	问题	平均分
1	是否依赖英文摘要来判断检索结果的相关性	3.26
2	是否依赖英文全文来判断检索结果的相关性	2.67
3	是否依赖中文全文来判断检索结果的相关性	3.98
4	关键词的高亮显示是否有助于你判断检索结果的相关性	4.33
5	高度相关、一般相关、不相关三个级别是否容易确定	3.78

注：1 分“完全否定”；2 分“否定”；3 分“中立”；4 分“肯定”；5 分“非常肯定”。

根据表 4 所获数据，得到三点启示：①从用户相关性判断的依据上看，目标语言文档的全文、目标语言文档的译文摘要、目标语言文档的译文全文都是被参与者认可的判断依据。当然，在本实验中，由于参与者的母语都是中文，因此他们更倾向于中文全文，即目标语言文档的全文。②在用户相关性判断过程中，关键词的高亮显示被参与者普遍认为肯定有帮助。③多级相关性判断还没有被广泛用于信息检索，主要是研究者们担心会给用户造成判断的不确定性。但是，多级相关性判断能够更真实地反映用户对信息的理解，为提高信息检索的效率提供更多、更详细的信息。本实验的参与者肯定了多级相关性判断的可行性。当然，对于级数的确定还需根据具体实验环境再进一步研究，并非级数越多越好。

综上得出结论，在文档查看方面的用户行为特征有：目标语言文档的全文、译文摘要、译文全文都是用户认可的判断依据。关键词的高亮显示被认为肯定有帮助。大多数用户认可多级相关性判断。

4.4　用户的相关反馈分析

由于本实验用户所进行的相关性判断为多值判断，因此笔者采用 NDCG 值作为评价指标对各种方法的反馈效果进行了检验。在公式（1）中，笔者对 r(j) 的取值分别为：“高度相关”文档的权值设为 4；“一般相关”文档的权值设为 1；“不相关”文档的权值设为 0。除了 Baseline 方法外，笔者对采用 TE 方法前后各个用户在各个检索主题上获得的平均 NDCG 值和采用 Combined 方法前后各个用户在各个检索主题上获得的平均 NDCG 值进行计算，结果见表 5。同时，笔者还对采用 TE 方法前后、Combined 方法前后的 NDCG 值进行了统计检验，检验其是否具有显著性差异。

表 5 TE 和 Combined 方法与反馈前 NDCG 值比较

	NDCG 均值
TE 前	0.79
TE 后	0.84 **
Combined 前	0.78
Combined 后	0.84 **

** 表示相关反馈后比反馈前具有统计性显著差异,且 p-value < 0.01。

从表 5 看出,无论是 TE 方法还是 Combined 方法,都获得了比其相关反馈前更高的 NDCG 值,且两组的差异都非常显著。说明 TE 方法和 Combined 方法都是非常有效的相关反馈方法。相比 TE 和 Combined 这两种方法,它们得到的 NDCG 值是一样的,都是 0.84,且二者之间没有显著差异。说明这两种方法的效果相当,几乎没有区别。

综上得出结论,在相关反馈方面的用户行为特征主要有:翻译优化方法以及翻译优化与查询扩展结合的方法在用户交互环境下非常有效,但二者并无实质差别。尽管从理论上说,系统每经过一次相关反馈,其检索结果应有所提高。但从用户的实际检索结果来看,翻译优化方法与翻译优化和查询扩展的结合方法的效果相当。说明在真实的用户参与环境下,用户的反馈行为并非次数越多越好。

4.5 用户的翻译重新选择分析

翻译优化的作用是通过相关反馈来改进查询翻译的质量,但经过优化的翻译不可避免地会存在一些噪音。因此 ICE-TEA 系统在翻译优化之后设计了一个交互过程,即让用户能够对翻译优化后的查询翻译进行重新选择。系统在翻译优化后的默认设计是全选所有经过优化的查询翻译,用户需要做的是把那些他认为是不正确的翻译删除。通过挖掘用户日志,笔者对 54 个参与者使用 TE 和 Combined 两种方法的所有查询在翻译优化后的改动幅度和改动强度进行了统计,结果见表 6。研究翻译优化后用户对翻译的选择有助于我们根据用户习惯来确定优化后翻译的个数。

表 6 翻译优化后用户对翻译的改动统计

改动幅度	查询总数	被改动的查询数	被改动查询占总查询的百分比(%)		
	330	277	84		
改动强度(针对被改动的 277 个查询)	检索主题	被删除翻译数的平均值	被删除翻译数的最大值	被删除翻译数的最小值	标准差
	40007	2.9	7	1	
	40019	3.6	11	1	
	40028	4.9	14	1	
	40039	3.4	8	1	
	40043	3.3	7	2	
	41018	3.6	10	1	
	41025	2.9	6	1	
	41027	2.8	16	1	
	41035	4.4	16	1	
	平均	3.6	16	1	2.5

如表6所示，从参与者的改动幅度上看，在全部330个查询中，被改动的查询数为277个，占84%。可见，参与者在实验过程中对优化后的翻译仍然不够满意，并且喜欢用这个交互功能，希望人工再对优化的翻译作进一步筛选。

我们进一步对277个被改动了的查询按照检索主题进行改动强度统计。在9个检索主题中，40028的改动强度最大，平均被删除的翻译数为4.9个；41027的改动强度最小，平均被删除的翻译数为2.8个；其他检索主题的改动强度在这个范围内，但各不相同。可见，检索主题是影响跨语言信息检索相关反馈的一项重要因素。全部9个检索主题平均被删除的翻译数为3.6，最大值为16，最小值为1，标准差为2.5，说明参与者在对优化后的翻译进行筛选时，其改动强度具有一定的稳定性。

综上得出结论，在翻译重新选择方面，用户行为特征有：用户对翻译的改动幅度为84%，改动强度为平均删除3.6个翻译，用户对于反馈后的翻译仍然愿意作进一步选择。

4.6 其他问题分析

问卷中我们还要求参与者对三种方法进行排序。44.44%的参与者认为Combined方法最好，TE方法其次，Baseline第三。20.37%的参与者认为TE方法最好，Combined方法其次，再次是Baseline。不论TE还是Combined，很明显参与者更喜欢有相关反馈技术的跨语言信息检索系统，说明参与者愿意与系统进行交互。

5 结论

交互式跨语言信息检索是信息检索一个重要的研究方向，其理论依据已被研究者们提出并受到关注。在查询表示、查询翻译、检索、文档选择、文档查看等主要步骤及其逆过程中，大多均可融入用户的参与。在本研究中，笔者设计了一个用户全程参与的跨语言信息检索正向检索与逆向反馈实验，了解用户在整个检索过程中的行为特征。通过对交互式跨语言信息检索过程中的查询形成、文档选择、文档查看、相关反馈、翻译重新选择等几个方面用户行为的分析发现，从正向检索来看，用户的查询输入不长，一般选自能够概括其检索需求的标题词，对于检索结果的相关性判断倾向多级且准确率较高，在查看检索结果时往往依赖目标语言文档的全文、译文摘要、译文全文；从逆向反馈来看，用户乐于参与相关反馈，他们使用翻译优化和查询扩展这两种相关反馈方法都能够获得较满意的检索效果，且实验证明，这两种方法结合使用的效果与翻译优化方法单独使用的效果相当，都能够提高检索结果，但二者并无显著差异，说明两次相关反馈并不能比一次相关反馈获得更好的检索结果，相关反馈的次数需要根据用户的检索行为与需求特点适当制定。然而，调查问卷反映大多数用户更倾向于两种相关反馈方法的结合，认为其效果最佳。以上这些用户行为分析将有助于指导交互式跨语言信息检索系统的设计与实践。

此外，在研究中笔者也发现交互式跨语言信息检索用户行为研究还有进一步深入的空间。在今后的研究中，笔者将选择背景更多元化的用户参与实验，让用户自己在给定的检索任务下选择检索主题，并根据用户的信息行为来设计交互式界面以帮助用户更好地完成检索任务。

参考文献：

[1] He D Q, Wang J Q. Cross-language information retrieval[M]//Göker A, Davies J. Information Retrieval: Searching in the 21st Century. UK: A John Wiley and Sons, Ltd, Publication, 2009: 233-254.

[2] He D Q, Oard D W, Wang J Q, et. al. Making MIRACLEs: Interactive translingual search for Cebuano and Hindi[J]. ACM Transactions on Asian Language Information Processing, 2003, 2(3): 219-244.

[3] 张秀坤. TREC人机交互检索评价项目研究[J]. 图书情报工作, 2006(1): 72-75. (Zhang Xiukun. Study of the TREC interactive IR evaluation track[J]. Library and Information Service, 2006(1): 72-75.)

[4] Kekäläinen J, Järvelin K. Using graded relevance assessments in IR evaluation[J]. Journal of American Society for Information Science, 2002, 53(13):1120 - 1129.

[5] Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques[J]. ACM Transactions on Information Systems, 2002, 20(4):422 - 446.

[6] Petrelli D, Beaulieu M, Sanderson M, et al. Observing users, designing clarity: A case study on the user-centered design of a cross-language information retrieval system[J]. Journal of American Society for Information Science, 2004, 55(10):923 - 934.

[7] Kraaij W, Hiemstra D. Cross-language retrieval with the twenty-one system[C]. Proceedings of the Sixth Text REtrieval Conference (TREC 1997). Gaithersburg, MD, USA, 1997:753 - 760.

[8] Orengo V M, Huyck C. Relevance feedback and cross-language information retrieval[J]. Information Processing and Management, 2006, 42(5):1203 - 1217.

[9] Ostenero F L, Gonzalo J, Verdejo F. UNED at iCLEF 2003: Searching cross-language summaries[C]. Proceedings of the 3rd Cross-Language Evaluation Forum (CLEF'03). Trondheim, Norway, 2003:450 - 461.

[10] Oard D W, He D Q, Wang J Q. User assisted query translation for interactive cross-language information retrieval[J]. Information Processing and Management, 2008, 44(1):181 - 211.

[11] 吴丹. 英汉交互式跨语言检索系统设计与实现[J]. 现代图书情报技术, 2009(2):89 - 95. (Wu Dan. Design and implementation of an English-Chinese interactive cross-language information retrieval system[J]. New Technology of Library and Information Service, 2009(2):89 - 95.)

[12] He D Q, Wu D. Enhancing query translation with relevance feedback in translingual information retrieval[J]. Information Processing & Management, 2011, 47(1):1 - 17.

[13] 吴丹, 何大庆, 王惠临. 基于伪相关反馈的跨语言查询扩展[J]. 情报学报, 2010(2):232 - 239. (Wu Dan, He Daqing, Wang Huilin. Cross-language query expansion using pseudo relevance feedback[J]. Journal of the China Society for Scientific and Technical Information, 2010(2):232 - 239.)

[14] Ballesteros L, Croft W B. Phrasal translation and query expansion techniques for cross-language information retrieval[C]. Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval. Philadelphia, PA, USA, 1997:84 - 91.

[15] Cieri C, Strassel S, Graff D, et al. Corpora for topic detection and tracking[M]//Allan J. Topic detection and tracking. Norwell, MA, USA: Kluwer Academic Publishers, 2002:33 - 67.

[16] He D Q, Wang J Q, Oard D W, et al. Comparing user-assisted and automatic query translation[C]. Proceedings of the 3rd Cross-Language Evaluation Forum(CLEF'02). Rome, Italy, 2002:400 - 415.

吴 丹 武汉大学信息管理学院博士, 副教授。通讯地址: 武汉市珞珈山武汉大学信息管理学院。邮编: 430072。

(收稿日期: 2011-07-14; 最后修回日期: 2011-08-25)