一种新型最优检索结果的发现与论证

洪 宇 康杨杨 姚建民 朱巧明 周国栋
(苏州大学计算机科学与技术学院 江苏 苏州 215006)

摘 要 该文基于量化数据证明如下观点:检索结果中,本源正确的检索结果应为最佳(即蕴含的信息符合客观逻辑和自然规律的检索结果).在此基础上,提出了一种新型基于正确性的信息检索评测方法.文中讨论的信息正确性是一种事实性的客观正确性,不随用户主观判断产生正确性的变化,与依赖用户满意度的相关结果具有显著差异.当前,信息检索方向的研究尚未关注检索结果本源正确性的自动检测与应用,且尚未提出相应的排序优化算法.文中即针对这一问题进行量化的科学验证,并给出相关研究的评测框架.

关键词 信息检索;满意度;正确检索结果;评测标准
中图法分类号 TP391 **DOI号** 10.3724/SP.J.1016.2013.00643

Discovery and Illustration of Novel Optimal Retrieval Result

HONG Yu KANG Yang-Yang YAO Jian-Min ZHU Qiao-Ming ZHOU Guo-Dong
(*School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu* 215006)

Abstract Based on the large-scale data, in this paper we attempt to prove that originally correct search results (viz., the results whose content obey objective logic and nature laws) is optimal. Relying on this, we propose a new correctness based evaluation metric for information retrieval. The correctness in this paper is a kind of factual and objective truth which never changes with subjective cognition of users. It is very different from relevant search results which rely on the satisfactoriness of users. Until recently, there hasn't been related research on the original correctness detection and application in the field of information retrieval, and also the corresponding method for optimizing search result ranking has never been proposed. This paper focuses on illustrating the issue with quantitative evidences, and further offering a reasonable evaluation framework.

Keywords information retrieval; satisfaction; correct retrieval results; evaluation metric

1 引 言

传统搜索引擎借助查询(query)与信息内容层面的语义一致性或相关性获取和推荐信息,辅助用户获取知识或对未知事物的理解[1].个性化搜索引擎[2-3]则通过学习用户的搜索行为,掌握用户的检索意图,借以更有针对性地满足用户的信息需求.但是,上述研究忽视了一个重要问题:信息本源的正确性.本源正确性(以下简称"正确性")是一种信息本身固有的客观属性,且不由用户或任何外界因素左右.比如,"人类不能跳入太空源于万有引力的存在"

收稿日期:2012-06-27;最终修改稿收到日期:2012-11-08.本课题得到国家自然科学基金(61003152,61272259,60970056,60970057,90920004)、教育部博士学科点专项基金(20093211110006,20103201110021)、江苏省自然科学基金(BK2011282)、江苏省高校自然科学基金重大项目(11KIJ520003)以及苏州市自然科学基金(SYG201030)资助.**洪 宇**,男,1978年生,博士,副教授,主要研究方向为个性化信息检索、话题检测、信息抽取和观点挖掘.E-mail:hongy@sdua.edu.cn.**康杨杨**,男,1989年生,硕士研究生,主要研究方向为个性化信息检索.**姚建民**,男,1971年生,博士,研究员,主要研究领域为机器翻译.**朱巧明**,男,1963年生,博士,教授,研究领域为中文信息处理.**周国栋**(通信作者),男,1968年生,博士,教授,中国计算机学会(CCF)会员,研究领域为自然语言处理.E-mail:gdzhou@suda.edu.cn.

是一种源自客观规律的信息,任何破坏这一规律的因素都将导致信息组成的错误与传播的误导.

信息的正确性不同于相关性和满意度.正确性反映了信息与客观事实的一致性,比如一段文字是否客观描述了特定事件的真实过程;相关性仅仅反映信息间共有成分的含量或联系的强弱,比如两段文字是否描述了同一主题;满意度则反映特定认知能力和判定标准下,人对于信息的喜好和偏爱程度,其往往因认知角度和水平、知识底蕴以及衡量事物标准的差异,呈现出显著的个体独立性.换言之,任何个性化的用户对信息给予的"满意"、"认可"甚至"正确"的断言仅能体现其个体满意度,并不能绝对反映信息本身的正确性.例如,针对查询"什么是中国龙",如何区分如下检索结果(即相关、满意或正确的结果):

(1)你想知道关于中国龙的介绍吗?请看《中国的历史与文化》一书;

(2)近来一些研究证实中国龙实际上是鳄鱼;

(3)中国龙是中华民族的图腾.

显然,结果1与查询相关,但不是用户满意的信息,原因在于结果1本身并未令用户直接获取到"中国龙"的知识(仅仅提供了一种弱强度的联系).相反,结果2往往令用户满意,尽管本质上是一种错误和片面的论断.事实上,导致这一错误信息受用户"青睐"的原因在于其直接回答了查询,使用户以最小的时空消耗获取到意图所指的知识,尤其结果的文字体现出一定权威性,具有一定置信度.相对地,如果用户对中国的传统文化有足够了解,那么用户应对结果3满意,否则,尽管这一结果正确,但用户对它并不满意.原因在于特定用户的知识背景和积淀无法支持它对结果3的认知.

通过上述例子,本文作出以下假设:用户之所以使用搜索引擎是源于其"不知道且想知道"特定事物或知识.也因此,针对未知事物,用户往往具有很少甚至没有任何先验知识,借以辅助其判断信息的正确性.如果这一假设成立,那么目前以用户满意度为核心的搜索引擎正在陪伴用户"一起犯错",原因在于所有借助用户反馈(显式和隐式)的机器学习方法,如基于点击(click-through)[4]或视觉(eye-tracking)[5]行为的排序学习算法(Behavior-based Learning to Rank)[6],都是以本不具备判定对错能力的用户及其赋予的粗糙反馈为标杆,实现检索结果推荐或排序的"优化".简言之,这类方法本身是一种弱指导的"伪优化".

针对这一问题,本文提出了一种全新的观点:最佳的检索结果应为本源正确的信息.为了验证上述观点,本文结合维基百科(Wikipedia)中的百科知识假设为本源正确的信息源,开发了基于Google搜索引擎的标注平台,并设计了一种涉及用户体验的评测标准.实验结果表明用户对信息的满意度和正确度的评测标准有很大差异,并且正确检索结果对用户情绪的正面影响较大.

本文第2节介绍相关工作;第3节通过一种视觉游戏引出基于正确性检索的基本理论;第4节介绍标注平台、系列标注结果及分析、面向信息正确性的评价标准,并利用这一标准对现有基于用户行为的排序策略进行评测与分析;第5节总结全文.

2 相关工作

传统搜索引擎,如Google、Baidu、Yahoo和Bing等,为用户提供了大量相关检索结果,但是在追求返回更多信息的同时,很难兼顾检索结果的准确率.针对用户信息需求较为集中、分类更加精细的情况,传统搜索引擎往往对检索信息定位不精确,返回的结果冗余过大.随着检索模型[7]的优化,尤其是借助自然语言理解的辅助,如利用语义关系识别[8]、上下文语境匹配[9]、词义消歧[10]等技术,有效提高了文本相关性的度量精度,从而大大改进了检索结果与查询之间语义级的拟合程度,使检索精度显著提升.

个性化搜索引擎是指在传统搜索引擎的基础上,引入用户的背景、兴趣和偏好等特征,针对不同用户提供个性化的检索服务.比如,文献[11]在对搜索引擎的排序算法和用户行为深入研究的基础上,通过隐式方法收集用户检索过程中的行为信息,构建用户长期兴趣模型、短期兴趣模型、时段兴趣模型等.从而,个性化信息检索在内容、语义和语用等纯粹的文本处理和应用的基础上,进一步对信息关联关系的判定过程引入了用户认知习惯这一关键特征.由此,信息检索在通用性和个性化方面分别形成了各自的理论模型.

然而,尽管现有信息检索研究在获取相关信息和满足用户认知偏好方面有了长足进展.但截止目前,尚无针对检索结果的本源正确性进行判定和评估的研究.信息的正确性直接关系到认知过程的正确性,而信息检索是目前人们基于互联网实现知识共享和认知进步的重要平台.从而,信息的正确性是信息检索研究无法忽视的关键问题.为验证这一观点,本文第3节借助一种简易的"视觉游戏"直观地体现信息正确性的客观存在及其认知难度的普遍

性，并在第 4 节提供系列标注数据及量化指标作为理论依据.

3　视觉游戏

在探讨检索结果正确性之前，本文通过一项简易的“视觉游戏”呈现一种非常有趣的思维过程. 我们相信这种思维过程是支撑本文观点（正确的检索结果为最佳）的有力佐证.

“视觉游戏”的第 1 步是观测图 1 后回答如下问题：图中的环状图形是一种凸出物（如“平顶山丘”）或是凹陷物（如“弹坑”）？部分观测者首次看到该图片时往往认为其显示了一种凸出物，且这一论断存在不确定性. 但当浏览图 2 时，观测者将相对容易地判断出图示为月球表面的陨石坑. 以上的判断过程反映出人们在对事物缺少整体认知和全局概念的时候，往往难以对事物的客观属性给出正确判定，且该情况在人们认知新事物时经常发生.

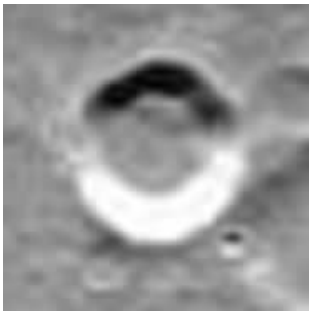

图 1　“视觉游戏”步骤 1

“视觉游戏”的关键步骤为第 2 步. 在此之前，要求观测者对自身给予暗示，使之认定图 2 子图(1)中的环状物为凸出物（可参考图 1 的直觉）. 在此基础上进行第 2 步，观测图 2 子图(2). 部分观测者可能发现其显示出一片凸出的环状物. 如果这一错觉持续作用到图 2 子图(3)中，观测者将会发现整个月球表面布满凸点. 因此，对于一个不了解任何月球常识的儿童而言，上述错觉的衍生过程，很可能导致其成为月球全新地质结构的发现者.

图 2　“视觉游戏”步骤 2

第 2 步反映了人们认知不熟悉或未知事物的思维过程，即从简单单一到复杂综合. 在整个思维过程中，初始认知显得特别重要，它将直接影响后续的认知甚至全部认知的正确性. 现实中，人们总是对初始认知印象深刻，并且这种印象将持续很长时间. 因此，如果初始认知有误，那么将很难被纠正. 这就类似于“儿童被告知存在圣诞老人，那么十六岁之前的每个圣诞节，他们都会在壁炉旁挂上袜子”.

返回本文论点所指的核心对象：信息检索，可发现知识的检索过程与“视觉游戏”有如下相似性：

(1) 探索未知并且寻找答案；

(2) 缺乏或甚至没有对目标事物的先验认知；

(3) 需要一个过程来理解和接受新事物，在这个过程中，每一步的正确性都难以预知.

实际上，在信息检索领域，上述现象发生在用户浏览检索结果列表并深入了解其细节的过程中. 但由于无法确定检索结果的质量，用户在此过程中的每一步都可能被错误的信息误导.

但用户是否可能在初始阶段即被误导呢？考虑到信息检索和“视觉游戏”的相似性，答案是肯定的. 为了解释这种现象，可回顾目前搜索引擎对检索结

果进行排序的原因.由于视觉范围的限制(源自 PC 屏幕的尺寸),人们的视觉只能捕捉到很小比例的检索结果(即检索列表在屏幕尺寸内的显示区域).因此位于检索列表顶部的结果将首先进入用户的视觉捕捉区域.所以,根据相关度和满意度对检索结果进行排序,并将最相关或最满意的结果置于顶部显得十分必要.然而,任何事物都有两面性,如果不正确的信息被置于检索列表顶端,也将会在用户浏览检索结果的开端即被捕获.所以,如同在"视觉游戏"中的分析,错误的初始认知会误导初学者后续的学习.尤其当不正确的信息是与用户需求相关,甚至满意度较高时,将会因其博取了用户的信任或偏爱,使得对用户后续认知过程的误导更为严重.

因此,考虑到人们认知未知事物的特性,本文提出了一种全新观点,搜索引擎检索的最佳结果应当是本源正确的信息,并且这类信息应当被排列靠前,置于检索结果列表的顶部.换言之,相关性和满意性不应当作为衡量检索结果质量的唯一标准.下一节将给出量化的数据支持这一观点.

4 标注系统和实验结果

本节首先介绍利用维基百科(Wikipedia)中的百科知识构建查询及其正确结果的方法以及基于 Google 搜索引擎建立结果正确性与满意度的混合标注平台的方法;其次,给出标注结果并利用量化数据进行分析;最后,介绍一种面向检索结果正确性及其用户体验的评价标准,即情绪波动代价系数 $MCost@n$,并利用这一标准检验现有基于用户行为,实现检索结果重排序算法的性能.

4.1 标注平台

标注平台包含 3 个组成部分:搜索引擎、交互接口和数据库.搜索引擎通过远程访问 Google 提供信息检索服务.交互接口用来显示搜索结果和获得用户体验.数据库用来保存预先的查询词条、相应的正确结果和用户的实时体验.本节主要介绍如下两项关键问题:

(1)如何获得含有相应正确结果的查询词?

(2)如何收集用户体验?

考虑到第 1 个问题的必要性,并且确保结果的正确性,本文将中文百科知识作为(百科知识的歧义性较低,且包含大量符合自然规律的信息)正确查询结果及相应查询的数据源.本文从维基百科中提取出 279 576 个中文条目作为查询词,并抽选与查询词(即百科词条)对应的维基百科、百度百科和第三方百科知识库的网页以及针对这类知识库资源进行拷贝和转载的网页,作为正确的检索结果.此外,对导致正确结果排列靠后的普通和较长的条目进行过滤后,最终得到 3719 个查询词.

为了确保用户参与实验的可行性和真实性,本文在交互接口中直接显示上述查询词,以便用户选择感兴趣的查询词进行检索.查询词列表可以实时点击更新,以提高用户获取感兴趣查询词的可能性.为获得用户体验的实时数据,标注平台提供了两项表单记录检索结果的正确度、满意度以及用户点击每个结果时的情绪.第 1 个表单(FBC)包括 6 个问题,在点击检索结果前弹出(见表 1).第 2 个表单(FAC)包括 8 个问题,在点击检索结果后弹出(见表 2).表单 FBC 用来记录针对任一特定检索结果时的用户初始体验.当用户深入了解该检索结果的细节后,表单 FAC 记录其后续体验.由此,形成针对某一特定查询的 FBC-click-FAC 表单记录并提交数据库.完整的标注过程如下所示.

1. 从交互接口中选择最感兴趣的词条;
2. 浏览检索结果列表,选择其中一项;
3. 填写表单 FBC 并且点击检索结果;
4. 填写表单 FAC;
5. 继续步 2 或退出.

表 1 点击检索结果之前需要回答的问题列表(Form Before Click,即 FBC 表单)

Form *before* Click (FBC)
Q1:这是第 1 个检索结果吗?(是/否)
Q2:你是否有习惯不看第 1 个结果或是直接点击当前位置的结果?(是/否)
Q3:在这之前,你查看了前面的几个检索结果吗?(是/否)
Q4:你觉得这个检索结果的满意度怎么样?(单选按钮)
A. 非常满意(5) B. 感觉一般(4) C. 只有一点满意(3) D. 很难说(2) E. 不喜欢(1)
Q5:你为什么选择这个检索结果?(单选按钮)
A. 一定是正确的(5) B. 应该是正确的(4) C. 看起来是正确的(3) D. 排名高(2) E. 随机选的(1)
Q6:你现在感觉怎么样?(单选按钮)
A. 感觉很棒(5) B. 感兴趣(4) C. 挣扎着进行(3) D. 焦急的(2) E. 乏味的(1)

表 2　点击并浏览检索结果之后需要回答的问题列表(Form After Click，即 FAC 表单)

Form *after* Click(FAC)
Q7：你对这个结果满意吗？(单选按钮)
A. 非常满意(5)　B. 感觉一般(4)　C. 只有一点满意(3)　D. 很难说(2)　E. 不喜欢(1)
Q8：你对这个检索结果的正确性有多肯定？(单选按钮)
A. 非常肯定(5)　B. 肯定(4)　C. 也许(3)　D. 很难说(2)　E. 它是不正确的(1)
Q9：你对这个检索结果的错误性有多肯定？(单选按钮)
A. 非常肯定(5)　B. 肯定(4)　C. 也许(3)　D. 很难说(2)　E. 它是正确的(1)
Q10：什么使你相信它是正确的？(单选按钮)
A. 高度权威性(5)　B. 来源可靠(4)　C. 内容可靠(3)　D. 看上去合理的(2)　E. 我有同样的想法(1)
Q11：什么使你满意？(单选按钮)
A. 提供了我需要的一切(5)　B. 能够学到一些知识(4)　C. 有趣的(3)　D. 网页漂亮(2)　E. 其它(1)
Q12：请将你查看的所有检索结果按照正确性排序.
Q13：请将你查看的所有检索结果按照满意度排序.
Q14：你现在感觉怎么样？(单选按钮)
A. 感觉很棒(5)　B. 感兴趣(4)　C. 挣扎着进行(3)　D. 焦急的(2)　E. 乏味的(1)

4.2　标注结果及分析

在实验过程中，共 19 名志愿者在上述平台上进行了标注. 其中，1 名志愿者为该研究的直接参与和设计者，2 名为具有信息检索研究背景的硕士研究生(包括前者)，另有 9 名自然语言处理方向的硕士研究生，8 名其它专业的本科、硕士学生(包括 1 名外国语专业的本科生). 19 名志愿者在标注过程中体现的数据分布特征较为一致且量级相似. 因此，下文所涉实验结果的量化指标以 19 名标注者的均值予以度量.

标注过程的周期为一个月，期间共提交了针对 719 个查询词的 1659 个 FBC-click-FAC 表单记录. 经过对误操作引起的不完整表单和错误表单的过滤，最终获得了针对 539 个查询词的 1238 个表单记录.

在讨论主要的标注结果之前，本节重点给出实验中的术语定义：

(1) FBC-click-FAC：从 FBC 到 FAC 过程中的点击；

(2) FBC 表单：用户点击检索结果之前填写的表单；

(3) FAC 表单：深入了解检索结果后填写的表单；

(4) 耐心层：标注某查询过程中，用户的耐心指标.

实验采用用户完成的 FBC-click-FAC 表单的个数来说明耐心层. 比如，给出一个查询，如果用户点击了两个检索结果，并且成功提交相应的 2 个 FBC-click-FAC 表单，那么耐心层的值即为 2.

此外，为了量化用户对检索结果正确性、满意度的意见和 FBC-click-FAC 表单的标注数据，标注平台根据强度，对表单中各个提问的结果分别制定了由“1 分”至“5 分”的分差(见表 1 和表 2). 最后，本文后续章节所提定量数据和趋势曲线，都产生自 FBC-click-FAC 表单中的特定问答(Question and Answer，QA)或问答组合的相应标注数据. 因此，后续章节标题中增加了形如 Q7＋Q8 的符号组合，借以区分数据分析结果的产生源.

4.2.1　基本数据分布(Q1＋Q2＋Q3)

根据 FBC-click-FAC 表单数据，一项统计结果显示仅仅有 20.52%查询的正确检索结果被用户获取到(被点击并且被浏览). 但是，根据数据库中的先验记录，至少有 50.09%的查询在检索结果列表的前 10 项结果中具有正确检索结果，并且所有志愿者均报告其至少浏览了检索结果的第 1 页(即浏览范围包含前 10 项结果). 因此，有充分的理由相信志愿者遗漏了 29.57%查询的正确检索结果.

此外，通过检查不同耐心层的召回率(见表 3)，本文发现了两个有趣现象：

(1) 用户应该有耐心去查看更多检索结果(见“点击数量”这一列，给出了不同耐心层下的点击数量)；

(2) 随着耐心层的提高，正确检索结果的召回率上升(见图 3).

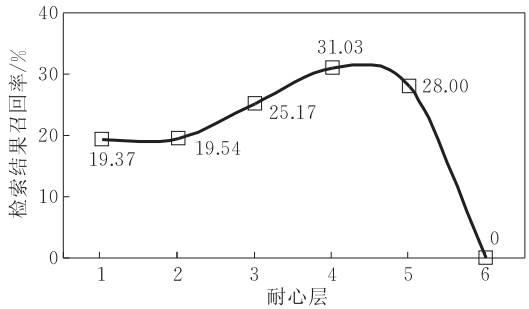

图 3　召回率随着耐心层增大的变化趋势

上述现象说明如果用户有耐心查看和学习更多知识，他能够逐步提升认知正确检索结果的能力. 然而，事实上绝大多数用户仅查看 1～3 个检索结果(见表 3)，因此，接近正确检索结果和成功认知的概

率都很低,尤其当正确结果远离检索结果列表顶部的时候,上述现象更为明显.

表 3 不同耐心层的召回率

耐心层	召回率/%	点击数量
1	19.37	191
2	19.54	266
3	25.17	429
4	31.03	116
5	28.00	125
6	0.00	90

4.2.2 认知转变(Q12)

数据的基本分布显示出一个特例,当耐心层增加到 5 时召回率下降.如果该异常不是由于对应耐心层上稀疏的数据引起,那么存在这样一个问题,用户的认知是否会被超量的学习所干扰.为了回答这一问题,本文使用 FBC-click-FAC 表单中 Q12 的数据来检查在不同的耐心层上的"正确排序"情况.实验专注于以下两个问题:

(1)耐心层增加时,相对的排名是否会改变?

(2)用户学习到新事物时,排名是否会下降?

针对第 1 个问题,本文对每一个耐心层计算:新点击事件发生后相对排名改变的平均概率(见图 4).其中,"相对排名的改变"是指志愿者接受到新知识刺激后,对先前检索结果正确性线性顺序的改变.例如,点击第 i 个和第 j 个检索结果后,用户给出的初始正确性顺序为$\{i,j\}$(即结果 i 的正确性优于结果 j),当用户再次点击第 k 个检索结果(点击通常意味着用户对新结果的摄入操作)并重新排列上述所有已观察的检索结果时,初始$\{i,j\}$的顺序将发生两类变化.其中,诸如 $\{i,k,j\}$,$\{i,j,k\}$,$\{k,i,j\}$的排序情况中,$\{i,j\}$的相对顺序不改变,而$\{j,k,i\}$,$\{j,i,k\}$和$\{k,j,i\}$的排序情况中,$\{i,j\}$的相对顺序(即相对排名)发生变化.

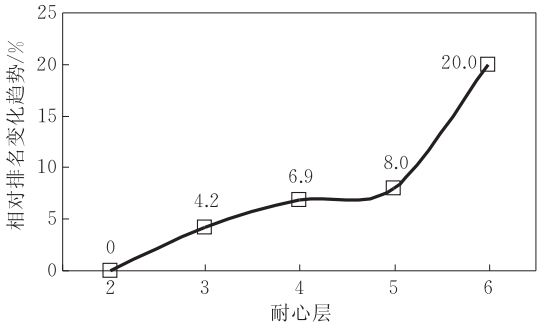

图 4 相对排名改变的趋势

该测试是用来检测新知识对于用户认知正确性的干扰程度.如图 4 所示,干扰不仅存在,而且随着耐心层增加,干扰相应提高.这种现象说明,当用户面临很多可能正确的信息时会感到困惑.因此,用户这种缺乏耐心和易受超量学习干扰的特性,初步说明了正确信息在检索结果列表中前置的必要性.相对地,正确信息后置则容易造成用户错失正确信息或迷失在冗余的信息当中.然而,正如 4.2.1 节中的统计记录,实际上只有 50.09%的正确信息出现于开始的 10 项检索结果中.由此,几乎一半的正确信息都存在于用户耐心难以达到的排序位置,超量学习也将难以避免.

对于第 2 个问题,本文计算了不同耐心层上每个点击事件排名下降的概率(见图 5).此处,"排名下降"表示志愿者点击新检索结果后,先前点击的检索结果排名下降的现象.例如,第 i 个和第 j 个点击的原始顺序是$\{i,j\}$,第 k 个点击被插入到线性顺序中变成$\{i,k,j\}$,其中,第 j 个点击对应的检索结果排名下降,而第 i 个点击未下降.

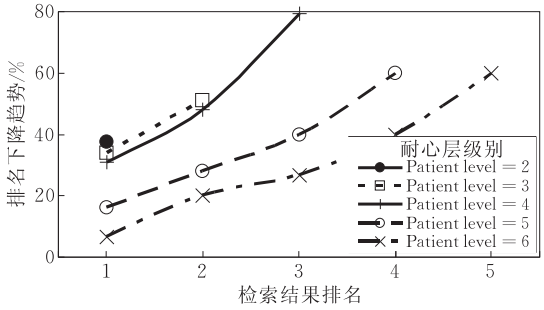

图 5 排名下降概率

该测试用来检测用户相信其判断正确的坚定程度.如图 5 所示,无论哪一个耐心层,其上的第 1 个点击排名下降的可能性最低.这一标注结果表明,用户总是坚持认为其初始点击的结果最为正确.然而,实际上第 1 个点击的正确率仅有 15.4%.此外,如果用户在第 1 次点击时不能成功获取正确结果,那么在接下来的浏览过程中,其将很难精确识别正确结果(如表 4 中所有准确率).这一现象支持了本文在"视觉游戏"中提出的观点,即用户总是轻易地接受初始认知,并持续这种认知很长时间.然而,目前的搜索引擎基本未考虑用户在整个学习过程中认知未知事物和初始错误可能引起的消极影响.

表 4 不同点击的准确率

第 i 次点击	准确率/%	第 i 次点击	准确率/%
1	15.40	4	1.39
2	6.32	5	0.00
3	5.12		

4.2.3　数据的错误修正(Q12)

通过在 4.2.1 节中的分析，用户如果有耐心浏览和学习更多检索结果(如图 3 所示，高耐心层有更高的召回率)，用户能够逐步提升获取正确信息的能力.然而，事实上很难预计用户在获得正确点击之前将犯多少错误.为此，实验检测了每一个耐心层上错误修正的概率(见表 5).此处，“错误修正”是指用户错误地降低正确结果的排名.实际上，除了错误修正，检索过程中的错误还包括用户在成功获得正确检索结果之前的不正确点击.但是本文将错误修正看成更为严重的问题，因为其发生在用户已经接触到(甚至深入了解)正确知识之后，丧失了对正确结果的信任.这一数据分析能用于检测用户关于正确检索结果的信心度.

表 5　不同耐心层的错误修正

耐心层	错误修正/%	耐心层	错误修正/%
2	26.92	5	28.57
3	30.56	6	0.00
4	33.33		

如表 5 所示，几乎在所有的耐心层都存在错误修正，并且当耐心层达到 4 级之前，相应的错误修正概率持续增加(见图 6).这说明用户对于正确性的初步判断依赖于直觉或间接的相关知识，使其难以确认信息的正确性，从而引起后续点击准确率降低(见表 4).有趣的是，在第 2 层的错误修正率相当低.这是由于用户更加相信其首次认知，且很少修改.从而进一步验证了初始正确知识的摄入，对用户认知正确性的重要作用.由此，如何基于内容识别正确信息并减少用户判定正确信息时的顾虑，应是检索领域全新的挑战.

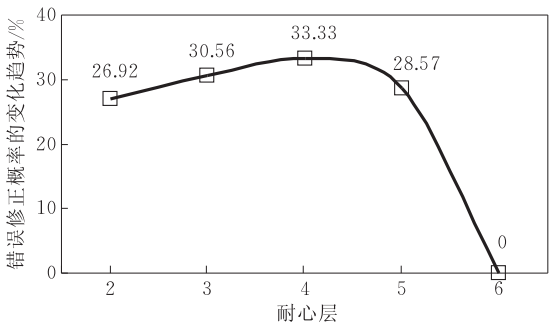

图 6　错误修正概率

4.2.4　认知一致性(Q4+Q5)&(Q7+Q8)

通过在 4.2.1 节、4.2.2 节、4.2.3 节的分析，可总结出用户在认知本源正确检索结果时呈现的行为特点：

(1)比较低的准确率和召回率；

(2)容易被错误的检索信息迷惑；

(3)坚持相信最初的选择；

(4)对每一个正确的决定没有把握.

因此，一个非常有必要去思考的问题是用户为什么接受和相信不正确的信息以及为什么会被迷惑.正如 4.2.1 节中的讨论，一种可能的原因是用户关于正确性和满意性的认知不同.具体而言，用户可能总是关心信息是否满足其需求，而忽略信息本源的正确性，甚至用户无法区别正确性和满意度.为了验证这一论断，本文检查了 FBC-click-FAC 表单中 Q7，Q8，Q12 和 Q13 的答案，借以调查用户在正确性和满意度上存在的差异.

首先，本文分别检查第 1 个、第 2 个和第 3 个点击结果在点击之前和点击之后的平均满意度得分，见表 6(点击之前得分 Score-before-Click，SBC；点击之后得分 Score-after-Click，SAC).实验结果表明随着点击数量的上升，平均满意度 SBC 不断升高，而平均满意度 SAC 不断降低.这说明在点击之前和点击之后，用户的满意度往往不同.根据这一差异，至少可以认定用户总是清楚地知道其需求是什么.原因在于，用户深入了解详细信息之后，完全可以否定先前对于信息是否令人满意的意见.

表 6　不同点击的满意度得分

第 i 次点击	SBC	SAC	差值
1	3.183	3.646	0.463
2	3.357	3.214	0.143
3	4.000	2.556	1.444

正确性的情况则完全不同.本文检查了第 1 个、第 2 个和第 3 个点击在点击之前(SBC)和点击之后(SAC)的平均正确性得分，见表 7.实验结果表明随着点击数量的上升，SBC 和 SAC 的平均正确性都不断降低.尤其是 SBC 和 SAC 之间的差别很小.因此，如果用户只通过浏览标题或网页快照不能判断检索结果正确性这一假设为真，那么 SBC 和 SAC 之间的微小差别在一定程度上能说明用户常常意识不到什么是正确信息.

表 7　不同点击的正确性得分

第 i 次点击	SBC	SAC	差值
1	3.366	3.556	0.190
2	3.286	3.203	0.083
3	2.333	2.258	0.075

其次，本文检查了第 1 个、第 2 个和第 3 个点击结果关于正确性和满意性的 SBC 和 SAC 的差别.结果显示了 SBC 截然不同的趋势(见图 7).满意度

上升的趋势反映如下规律:如果信息看起来有趣、相关且符合用户偏好,用户往往很容易满足.相反,正确性下降的趋势说明用户对正确的信息更苛求.因此,用户在深入信息细节之前(SBC),似乎对满意度和正确性有完全不同的认知.然而,通过分析两者SAC的趋势,可以发现认知非常相似(见图8):随着点击数量上升,SAC趋势都为下降.这反映出当用户深入细节之后(SAC),其对评估信息的质量往往具备一致的标准(即兼顾满意度和正确性),此时信息的满意度和正确性对于用户体验有一致的影响.

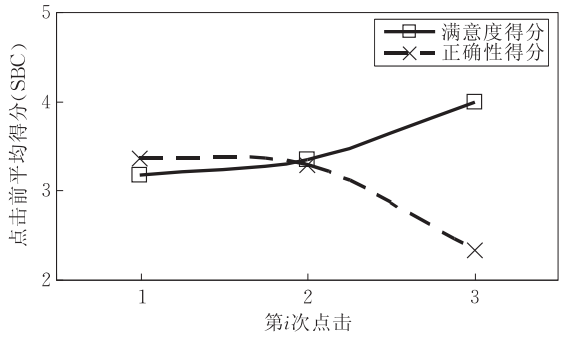

图7 满意度和正确性的SBC分布

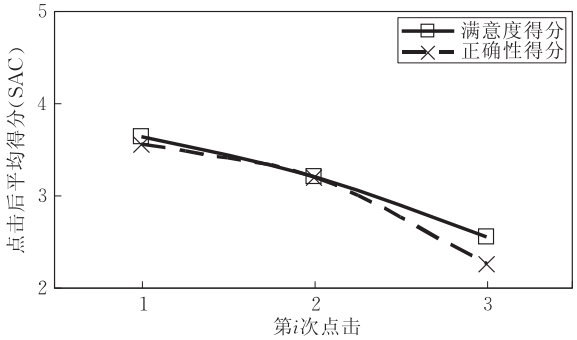

图8 满意度和正确性的SAC分布

综上,本文得出用户评估满意度和正确性的特性:

(1)点击一个检索结果时,用户的满意度标准低.当深入查看检索结果细节时,用户满意度标准高;

(2)用高标准的正确性来选择和评估检索结果.

4.2.5 用户情绪波动导致数据变化(Q6+Q14)

实验发现志愿者有情绪波动的现象.本文中,情绪波动是由正确性和满意度引起的情绪变化.本文检查了Q6(点击之前情绪 Mood-before-Click,MBC)和Q14(点击之后情绪 Mood-after-Click,MAC)的标注结果,并计算从第1个至第5个点击(即耐心层1级至5级)的情绪指标.此处,每个情绪指标是特定耐心层上,情绪选项被用户选取的次数.整个情绪波动趋势显示于图9(MBC)和图10(MAC)中.值得说明的是:每个曲线的下降趋势应被忽视(即高耐心层对应的低情绪指标不在观察范围内),原因在于更高耐心层的点击数量本身相当稀疏,相应地,高耐心层上每种情绪被选取的次数也都较低.由此,正确的观测点应为不同曲线在二维空间上的整体高度和高度的差异.通过对图9和图10的比较,本文发现两个现象:

(1)曲线A、B、C、D点击前和之后的差别不明显;

(2)曲线E在点击之前和点击之后的差别较大.

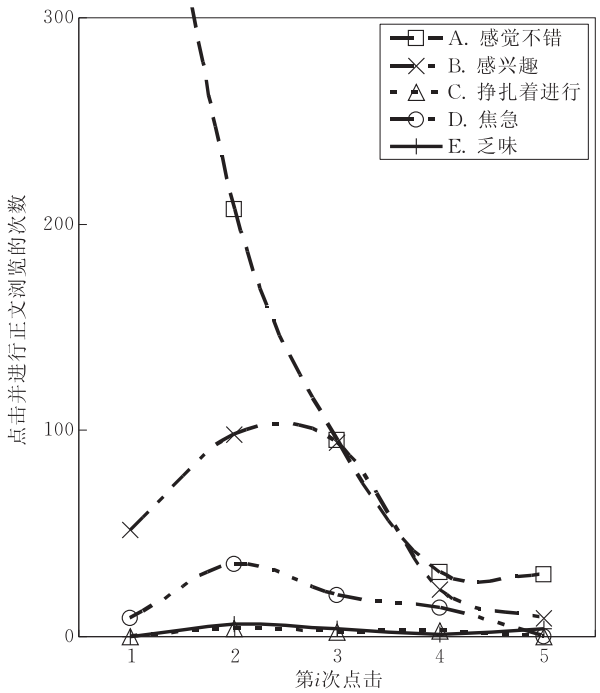

图9 用户点击之前的情绪波动(MBC)

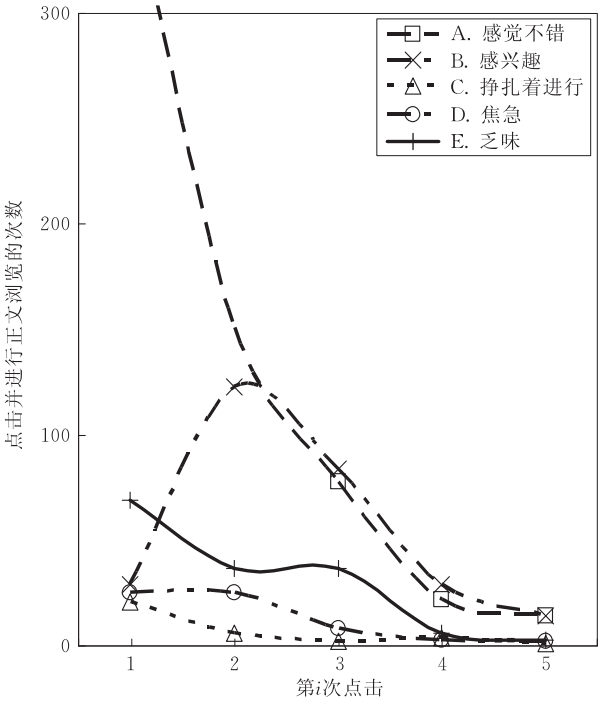

图10 用户点击之后的情绪波动(MAC)

上述现象说明用户的情绪波动通常发生在深入了解检索结果的细节之后，而不是在浏览结果列表的过程中.那么为什么检索结果的细节会引起用户明显的情绪波动呢?为了回答这一问题，本文专门提取了引起情绪波动的检索结果，并检查了不同点击对应的检索结果的错误概率(见表 8).从中可以发现几乎所有引起情绪波动的检索结果都为错误结果.因此，情绪波动应归咎于频繁发生的错误信息，尤其是明显的错误和噪音，例如广告、死链接、欺诈页面等(其它信息并不能动摇用户，因为初始时其内容的正确性对用户为未知因素).

表 8　引起用户情绪波动的错误检索结果发生概率

第 i 次点击	错误发生概率	第 i 次点击	错误发生概率	第 i 次点击	错误发生概率
1	1.000	2	0.9126	3	0.9219

如前所述，用户往往花费较长时间才能获取正确检索.因此，如果不正确的检索结果在这之前消极地影响了用户的情绪，用户可能没有足够的耐心检查接下来的检索结果，从而助长了用户错过正确检索结果的概率.因此，如果按照本文的假设，即正确的检索结果为最佳，那么，当检索系统进行排序时，出现在结果列表首端的不正确结果应予以屏蔽.

4.2.6　正确性识别(Q10+Q11)

根据量化的证据，本文得出了高排名正确检索结果的主要优势：

(1) 帮助用户快速认知准确的知识；

(2) 避免在观察正确知识的过程中迷惑用户；

(3) 减少消极情绪波动发生的概率.

因此，与满意的检索结果相比，正确检索结果不仅能够提供可靠的信息服务，而且提高了用户体验.

但是，如何鉴别出正确的检索结果呢?为了回答这一问题，本文检查了志愿者对权威性、来源可靠性、内容可靠性、合理性和认知一致性的观点(见表 2 中 Q10).本文分别计算了各个特性成功应用于确定正确性的概率(见表 9).分析正确检索结果和不正确检索结果的成功率的差别可以发现：权威度 A(见表 9)和认知一致性 E 是最为有效的特性.其在区分正确和错误信息时具有明显差异(表 9 显示两者都在判定正确和错误信息的成功率上具有显著差异).然而，由于难以在用户做出判定前直接获取其认知属性，自动计算认知的一致性比较困难.所以，权威度是识别正确性检索结果的一个可行特征.此外，通过检测认知的一致性(即用户同意检索结果的观点)发现识别正确结果的精确率很低，说明用户很少同意正确检索结果的意见.这进一步证明本文假设，当用户缺乏足够先验或相关知识时，通常不能准确认定正确性.

表 9　特征正确性识别的精确率

	精确率				
	A	B	C	D	E
正确信息	0.3439	0.3966	0.1304	0.0949	0.0343
错误信息	0.0796	0.3628	0.0531	0.0620	0.4425
差别	**0.2463**	**0.0338**	**0.0773**	**0.0329**	**0.4082**

注：A 高度权威性；B 来源可靠；C 内容可靠；D 看上去合理的；E 我有同样的想法.

此外，本文使用了相同的方法来检测特征的满意度识别效果.表 10 显示了成功率的分布情况.从中可以发现特征“提供用户所需的一切”为最佳.很明显，用户事实上最关心查询结果是否蕴含足够多的信息符合其查询偏好，但较少考虑准确性.

表 10　特征满意性识别的精确率

	精确率				
	A	B	C	D	E
满意	0.2992	0.3333	0.1197	0.0855	0.0086
不满意	0.1089	0.2657	0.2352	0.1237	0.0549
差别	**0.1903**	**0.0676**	**0.1155**	**0.0382**	**0.0463**

注：A 提供了我需要的一切；B 能够学到一些知识；C 有趣；D 网页漂亮；E 其它.

4.3　聚焦正确性的评估方法

根据上述分析，本文提出了一种全新的信息检索系统评测方法，主要用于检测正确信息的排名是否合理.由于正确的检索结果影响信息检索的性能，正确性评估方法应衡量 3 个特性：

(1) $P@n$：指出了开始的 n 个检索结果中出现正确结果的概率，可类比于相关结果评测时的正确率；

(2) $NDCG@n$：涉及正确程度和排名平衡点的标准化 $DCG@n$ 值，可类比于相关结果的 $NDCG$[12] 评测；

(3) 情绪波动代价系数 $MCost@n$：表明负面情绪波动在开始的 n 个检索结果中的代价.

在评测方法 2 中，传统的 DCG 综合考虑了相关度和排名：一个相关度较高的检索结果被排名很高，那么 DCG 值就会增加，否则将会减少.本文使用 DCG 测量正确度和排名的平衡点，即正确结果排名较高则 DCG 值较高，否则较小.

第 3 个特性 $MCost$，计算方法和 $NDCG$ 类似，但是关注于错误检索结果(即信息的本质并不正确，而非不相关)的满意度水平和排名的平衡点：一个满意度水平高的错误检索结果被排名很高，那么 $MCost$ 值就会增加，否则将减少.正如 4.2 节中的论述，如果一个错误检索结果令用户满意，它能很容易

误导用户且使用户点击更多相同或相似的错误结果.错误结果会引起用户情绪波动,并且这种波动会进一步消极影响随后的正确判定.因此,相比于错误且不受青睐的结果,这种错误将使用户满意的检索结果更为低劣.显然,错误结果前置,则 *MCost* 取值较高;错误结果后置,则 *MCost* 取值较低.由此,评测中对应较低 *MCost* 指标的系统性能较好,即系统在检索结果列表中向下排挤错误反馈的能力较强.情绪波动代价系数 *MCost@n* 的计算公式如下:

$$MCost@n=\frac{\sum_{i=1}^{n}\frac{2^{r(u_i)}-1}{\log(1+i)}}{Z_n} \quad (1)$$

其中,i 表示排序位置;u_i 表示排序为 i 的检索结果;$r(u_i)$ 表示 u_i 是否为本质错误的检索结果,如为错误则 $r(u_i)$ 为 1,否则为 0;Z_n 表示归一化系数,其值为当前 n 个排序位置都为本质错误的检索结果时,*MCost@n* 分子部分的计算数值,即全局错误的情绪代价.

本文重现了基于点击行为及分析的个性化检索系统(简写为 C-sys)[4].该检索系统的核心思想为:基于用户的点击行为,分析用户对系统反馈结果的粗糙满意度(即点击代表了用户对相应检索结果较为满意),借以利用机器自适应学习修正检索结果的相关度排序,实现面向拟合用户需求和意图的精准检索.Baseline 为标注平台使用的原始检索系统(即 Google 搜索引擎).实验以评测平台中志愿者选择的查询项以及相关反馈(即点击行为)为语料,利用系统 C-sys 实现检索结果重排序(封闭测试),并基于 *NDCG@10* 和 *MCost@10* 分别评测 Baseline 和 C-sys 满意性和正确性指标.

表 11 基于点击和视觉跟踪的重排序系统的性能
(主要对比 *NDCG* 与 *MCost* 的评价差异)

	C-sys	Baseline
NDCG@10	0.3343(优)	0.2926(劣)
MCost@10	0.9229(劣)	0.8976(优)

由于本文创建的标注平台仅仅能够保证每个查询项对应的首页检索结果中(即前 10 项检索结果中)仅有一项结果为本质正确(来自 Wikipedia 的标准结果网页),其余包括多于一项正确结果或没有正确结果的查询在构建标注语料时已被先期过滤(借以保证面向正确性和满意度标注及其分析的最低歧义性).因此,标注语料中的每个查询对应 19 个错误结果,*MCost* 的均值本源地较高.因此,实验的观测点应为同一评价方法内不同系统性能首先进行横向比较,然后将横向比较的差异在不同评价方法中进行纵向的一致性比较.比如,先横向比较系统 Baseline 和 C-sys 的 *NDCG*,记录差异;再横向比较两系统的 *MCost*,记录差异;最终对比两系统在 *NDCG* 上的差异与 *MCost* 上的差异是否一致.该一致性用于判定系统是否在满意度和正确性判定上具有一致的优势或劣势.此外,由于前 10 项检索结果中统一地仅有一项正确结果,所以 *P@10* 将始终为 0.1,不因系统变化而变化.因此这一评价标准在现有标注语料上不进行评测.

实验结果如表 11 所示,相比于 Baseline,系统 C-sys 获得了较高的 *NDCG@10* 值,但相对较低的 *Mcost@10* 指标(*Mcost* 值高表示负面情绪代价较高).由于系统 C-sys 是在原有 Google 检索系统(即 Baseline 系统)的排序结果上进行的排序优化,因此,实验结果说明系统 C-sys 在优化满意度结果排序的同时,误导了正确性结果的排序.同时也说明,虽然现有的 *NDCG* 测度在测量反馈结果的相关性和满意度方面具有一定优势,但难以检测错误检索结果对于用户体验的负面影响.实验结果也反映出:一味追求用户的满意度,并将其作为系统自适应学习与检索策略调整的标准,并不能真正在用户知识获取的正确性上给予正面支持.

5 总 结

目前,针对搜索引擎检索结果优劣性的评价基于一种公认的观点:用户满意的检索结果即为最佳结果.该观点将信息检索引入一种误区,即刻意追求用户满意度而忽略检索结果中知识本源的正确性.本文针对该问题提出一种全新观点:最优的检索结果应为"事实上正确"而非必然令用户满意的结果.为验证这一观点,本研究结合维基百科开发了基于 Google 搜索引擎的标注平台,利用各类标注结果给出了系列量化的依据,并在此基础上设计了一种涉及用户体验的全新评测标准.

实验结果表明传统搜索引擎在认定最佳检索结果时存在偏差,即查询结果虽令用户满意但其内容却蕴含或全部为错误的信息;同时也发现"正确信息并不能始终获得用户认可"的客观依据.在此基础上,新的评测标准揭示出当前面向个性化信息检索的排序算法在识别和推荐长期有效信息时的不足.

本文仅对这一问题进行了前瞻性的探索,未来工作将尝试对传统的 HITS 算法进行改进,提出新的基于正确性协作的 Correctness-based HITS 排序优化算法.

致　谢　苏州大学朱巧明教授和姚建民研究员对这一研究给予了长期支持与资助.仓玉同学对这一研究给予了有力协助.日本富士通公司汪菲女士对前期设想提出了宝贵意见.国家自然科学基金委和教育部以及江苏省和苏州市自然科学基金委对本文研究给予了长期支持.在此一并感谢!

参　考　文　献

[1] Allan J. Challenges in information retrieval and language modeling//Proceedings of the Workshop for Intelligent Information Retrieval. UMass, USA, 2003: 34-47

[2] Croft W B, Townsend S C, Larvrenko V. Relevance feedback and personalization: A language modeling perspective//Proceedings of the 2nd DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries. Dublin, Ireland, 2001: 49-54

[3] Robertson S E, Jones K S. Relevance weighting of search terms. Journal of the American Society for Information Science, 1976, 27(3): 129-146

[4] Xue G R, Zeng H J, Zheng C, Yong Y, Ma W Y, Xi W S, Fan W G. Optimizing web search using web click-through data//Proceedings of the 13th Conference of Information and Knowledge Management. Lisbon, Portugal, 2007: 118-126

[5] Laura A G, Thorsten J, Geri G. Eye-tracking analysis of user behavior in WWW search//Proceedings of the 27th Annual International Conference of Special Interest Group of Information Retrieval. Sheffield, UK, 2004: 478-479

[6] Agichtein E, Brill E, Dumais S. Improving web search ranking by incorporating user behavior information//Proceedings of the 29th Annual International Conference of Special Interest Group of Information Retrieval. Seattle, USA, 2006: 19-26

[7] Huang X J, Peng F C, An A J, Dale S. Dynamic web log seesion identification with statistical language models. Journal of the American Society for Information Science and Technology, 2004, 55(14): 1290-1303

[8] Giannis V, Epimenidis V, Paraskevi R, Euripides G P, Evangelos E M. Semantic similarity methods in wordNet and their application to information retrieval on the web//Proceedings of the 7th Annual ACM International Workshop on Web Information and Data. Bremen, Germany, 2005: 10-16

[9] Shen X, Tan B, Zhai C. Context-sensitive information retrieval using implicit feedback//Proceedings of the 28th Annual International Conference of Special Interest Group of Information Retrieval. Slvador, Brazil, 2005: 43-50

[10] Stokoe C M, Oakes M J, Tait J I. Word sense disambiguation in information retrieval revisited//Proceedings of the 26th Annual International Conference of Special Interest Group of Information Retrieval. Toronto, Canada, 2003: 159-166

[11] Li Y S, Shi S C, Zhang Y J. The application of user internet classification in personalized search engine. Journal of the China Society for Scientific and Technical Information, 2008, 27(4): 535-540

[12] Jarvelin K, Kekalainen J. IR evaluation methods for retrieving highly relevant documents//Proceedings of the 23th Annual Information ACM SIGIR Conference on Research and development in information retrieval. New York, USA, 2000: 41-48

HONG Yu, born in 1978, Ph. D., associate professor. His research interests focus on personal information retrieval, topic detection and tracking, information extraction and opinion mining.

KANG Yang-Yang, born in 1989, M. S. candidate. His research interest focuses on personal information retrieval.

YAO Jian-Min, born in 1971, Ph. D., professor. His research interest focuses on machine translation.

ZHU Qiao-Ming, born in 1963, Ph. D., professor. His research interest focuses on Chinese information processing.

ZHOU Guo-Dong, born in 1968, Ph. D., professor. His research interest focuses on natural language processing.

Background

Current universal search engines focuses on mining and recommending all possible relevant information to users. Further, the personal search engines are learning the real intention of users and recommending the information that extremely satisfies the personal requirements. But whether are the satisfactory search results factually factual? The answer should be no. In detail, it can be sure that users aim to explore the unknown when using the search engine, and thus they should have little or even no prior knowledge to support their judgments on the satisfaction with the search results. This may result in the probability that any result with a little relation to the query (even wrong one) can be determined to be satisfactory.

The doubts mentions above raise two questions: 1) whether are the satisfactory results always the helpful information for users to acquire knowledge? And whether can they consistently enhance the intelligence of information retrieval? 2) If the answer is no, whether does there exist other kind of information to compensate for the potential failings of users even or supersede the satisfactoriness? So, in this paper, we propose a hypothesis that the better search results should be factual information.

This research is supported by the National Natural Science Foundation of China (Nos. 61003152, 61272259, 60970056, 60970057, 90920004), the Special fund project of the Ministry of Education Doctoral Program (Nos. 2009321110006, 20103201110021), the Natural Science Foundation of Jiangsu Province (No. BK2011282), the Major Project of College Natural Science Foundation of Jiangsu Province (No. 11KIJ520003) and the Natural Science Foundation of Jiangsu Province, Suzhou City (No. SYG201030).