

检索相关性及其提高路径

Relevance in Information Retrieval and Its Improvement

赵 荣/ZHAO Rong, 宋慧敏/SONG Hui-min

北京师范大学管理学院,北京 100875

School of Management, Beijing Normal University, Beijing 100875, China

[摘要] 阐述并分析了信息检索相关性的内涵和影响相关性的因素,归纳了相关性研究的状况,指出当今信息环境下提高检索系统相关性的具体路径。

[关键词] 信息检索,相关性,检索系统**[中图分类号]** G252.7**[文献标识码]** A**[文章编号]** 1000-7857(2005)11-0063-03

Abstract: Concepts of relevance in information retrieval are analyzed in the paper. Researches abroad on relevance are simply reviewed and improvements in relevance of retrieval system are introduced.

Key Words: information retrieval, relevance, retrieval system**CLCNumber:** G252.7**Document Code:** A**Article ID:** 1000-7857(2005)11-0063-03

1 概述

1.1 相关性的定义及内涵

尽管到目前为止还没有一个公认的检索相关性的确切定义^[1],但为使问题的讨论得以进行,我们在此先把“相关性”描述为:用户从检索系统中检出的信息集合与用户需求相符程度的指标。该描述的量化表示即为相关度。事实上,检索相关问题远比定义复杂得多。人们在决策过程中产生了信息需求,其中一部分是对数据的需求,而更多的是对某个主题的文献(广义的,包含各种载体类型)的需求,而后者所对应的是一种相关性检索而非前者的确定性检索(要么是有,要么是无;要么是对,要么是错)。在相关性检索时,系统不直接回答用户的问题本身,它只提供与用户的问题相关的文献或文献的属性信息与来源供用户参考和取舍。这里的“相关”是一个模糊的判断标准,由最终用户来把握。同一检索过程产生的结果集中的同一文献相关与否、相关度高低,会因人(其相关领域的知识储备和知识结构、对检索课题的了解程度、检索目的、心理特征、兴趣和认知能力等)、因时间(课题进程的不同阶段、检索时间压力等)、因条件(信息资源的描述和组织状态、系统的导引功能、检索花费等)不同而有不同的判断。同一检索结果集中的各文献的相关度亦不同。比如输入关键词“二战”,检

索系统返回的结果中,文献 A 论述“二战”战役,文献 B 描写“二战”人物,文献 C 分析“二战”历史背景……显然各个文献与用户需求的相符程度不同,用户对检索结果中的各文献满意程度不同。可见相关性是一个多维的、认知的以及动态的复杂概念。

建立信息检索系统的目的就是使用户能够查找出所需要的信息,即检出相关信息,屏蔽无关信息。当代信息环境下,信息检索的核心是在系统文档集中为用户检出最相关的子文档集,并如用户所期望的那样,把检出文档按相关度降序排列。实际上相关问题涉及到检索系统的诸多环节,提高检索相关度的努力贯穿信息检索过程

的始终。如何提高检索相关度是信息检索实践和理论研究要解决的核心问题。

1.2 影响相关性的因素

当今的信息检索系统,无论是一般的机检系统还是因特网上的搜索引擎,其工作原理都是模拟手工检索过程,但系统中的各个模块越来越趋向于完全的自动化。系统要求把多维的、认知的以及动态的人的相关性判断过程转化为单一的、机械的、固定的、可计算、可模式化的步骤。系统结构及工作原理^[2]如图 1 所示。

检索系统的工作过程包含存储和检索两个互逆的过程。信息检索是建立在信息存储有序化的基础之上的。系统根据自身

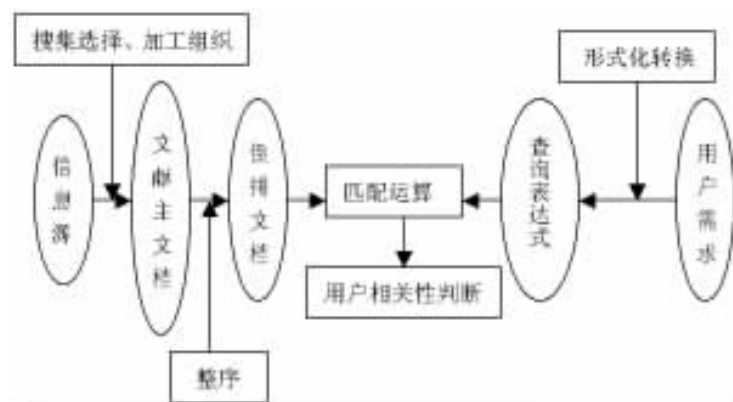


图 1 检索系统工作原理图

收稿日期: 2005-08-23

作者简介: 赵 荣,女,北京新街口外大街 19 号北京师范大学管理学院,副教授,研究方向为信息组织与检索;

E-mail: zhaorongbnu@eyou.com

特性、用户特征在一定专业范围内对信息来源进行连续性的采集、分析、选择,建立充足而适用的系统待检文献数据库;还要对文献进行标引(采用一定的标引语言和规则,对文献内容进行分析,从而赋予每篇文献以一定数量的内容标识和特征标识,如分类号、主题词、关键词等);描述(按一定标准和规范对文献的各种特征进行分析选择和记录);加工(内容精粹)及整序。将上面工作过程所产生的一条条记录作为一篇篇文献的“代表”,按一定顺序排列起来,形成主文档,再把记录中的属性值(即标引出的标识)抽出,按某种顺序重新排列,形成倒排文档,以支持按文献的某种特征标识快速检索。检索是存储的逆过程,但同样,用户需要先对自己的信息需求进行概念分析,形成内容标识,然后把把这些标识以及各标识之间的关系形式化地表示为查询表达式。检索时用户输入查询表达式,系统的检索代理软件自动将其与数据库倒排文档中的标识进行匹配运算,将命中的结果以一定的顺序显示给用户。用户对初步检索结果集进行相关性判断,如果满意则输出获取信息;如果不满意,可以修改或重构查询表达式,继续检索,直到得到满意结果为止。

由上述检索系统的工作过程可见检索的相关性判断有2个环节^[3]:①系统的相关性判断,即系统自动对相关度进行计算,输出检索结果;②用户的相关性判断,即用户在选择系统、拟定查询表达式及在系统命中的结果集中进行取舍时所做出的主观判断。系统谓之的“相关”并非待检文献与用户的需求相符,而是文档标识集合与查询表达式相匹配。这两种判断“天然”存在着距离,且距离不可能为零。我们所应做的具有实际意义的工作是使系统的相关性判断结果尽可能地逼近用户的相关性判断结果。

从系统工作原理分析,影响检索相关性的因素主要有以下几点。

1.2.1 系统文献数据库

文献数据库的规模、学科属性、收录范围等是能否检索到以及检索到多少相关文献的前提。

1.2.2 系统检索语言

也即是标引语言。在传统的检索工具和系统中,常常采用分类、标题、叙词等人工语言进行标引和检索。人工语言对同义词、近义词、属分词、相关词等进行了规范和控制,检索相关性较高。但在当今信息环境下,检索用户多为最终用户而非信息中介,因此要求标引和检索环节都使用直接性好的自然语言。但自然语言缺乏控制因而检索相关度较低。如何转换和控制自然

语言,利用智能技术在概念间建立联系,使模糊的查询表达能够逐渐明确,最终检索出准确而全面的相关信息,是提高检索相关性的关键。

1.2.3 索引方法

是全文索引还是部分索引、是否考虑超文本的不同标记所表示的不同含义、是否使用停用词表来过滤无用词汇、是否使用 Meta 标记信息、是否对图像替换信息和页面中的注解信息进行索引、是否支持截词检索技术、是否记录词位置信息以支持原文检索和限制检索、是否收集页面中的超链接等等,这些都会制约和影响系统相关度判别时的准确性。

1.2.4 检索匹配和检索结果排序算法

这一步是和上一个环节即索引方法密切相关的。它直接判定文档标识集合与查询表达式是否相关及相关度的大小。匹配算法历经单项模型、布尔模型到向量空间模型、概率模型、模糊集合模型再到概念模型、案例模型等。所采用的匹配策略从较为僵硬的二元匹配到具有灵活性的部分匹配及按相关度排序输出再到由人工智能技术支持的可进行同义扩展和相关概念联想,一步步地接近人对相关性的判断模式。在当今信息环境下,检索结果的排序算法对系统的相关性性能的影响尤为突出^[4]。因为首先网络资源的海量信息导致返回结果往往成千上万甚至几万几十万,用户不可能从中一一去判断、选择相关信息,一般只看排在前面的返回结果;其次,良莠不齐的网络信息质量与传统联机检索系统的信息供应商和数据库生产者所提供的专业、系统、经过筛选的信息源截然不同;再次,因特网的开放性导致搜索引擎检索系统也处于半开放状态。搜索引擎要保持其检索结果的准确性和公正性还要与 SEO 公司们进行操纵与反操纵的持续较量,不断改进其检索结果的排序算法。总之,在如此前所未有的复杂检索环境下,如何准确、客观、不受蒙蔽地组织检索结果,把与用户检索需求最相关的高质量信息资源纳入结果排序的前列是搜索引擎等信息检索系统所面临的严峻挑战。

1.2.5 用户需求的表达

如上所述,检索时用户的需求要形式化地表示出来,形成查询表达式,系统的相关性是指文档标识集合与查询表达式相匹配。但是实际上用户识别到的信息需求与实际真正的信息需求有差距,而用户表达出的信息需求即构造的查询表达式又与上述二者有区别。古希腊哲学家柏拉图有句名言:“人们要询问的既不是他知道的,又不是他不知道的。如果他知道,则没有必要再问;如果他不知道,则他也不知如何去

问。”用户意识到的信息需求往往处于模糊状态^[5],因此其构造的查询表达式常常是模糊的、范围失当的,不能准确表达其真正检索需求的。对于这样的用户需求表达,检索系统返回结果的相关性当然无从谈起了。过去在人际环境下,图书馆或情报所的专业检索人员会通过问题与用户交流,帮助他们消除不定因素,修改检索提问使其逐步与用户真正的检索需求相吻合。但在当今信息环境下,检索用户的构成是最为广泛的,大多是最终用户而非专业检索中介,这就更需要检索系统正确理解用户的需求表达,在人-机环境下能够与用户进行有效交流,使其模糊的信息需求趋于明确,确立正确的检索目标和范围,改善需求表达,从而提高检索的相关度。

2 提高检索相关度的路径

2.1 检索相关性研究现状

国外的相关性研究大致可分为3个阶段^[5],分别是1958年之前、1959到1976年以及1977年至今。第一阶段学者们的研究形成了面向系统的相关性的基本观点,即相关性主要是受系统的内在特征与操作影响的,主要包括分类表、主题词表、索引、词汇的语义以及语法描述、文献组织以及检索提问的分析与检索策略等。可见这一阶段的研究重点是集中在检索系统的存储过程中建库、标引、文献描述、索引和匹配等环节的。对检索提问的分析与检索策略的研究主要基于受控检索语言,并未基于最终用户的相关性判断特性,用户作为相关性判断的“决定因素”却未引起足够的重视。显然这一阶段的研究视角是不全面的。第二、三阶段的研究高峰分别集中于20世纪60年代至70年代前期以及20世纪80年代中后期至今。Mizzaro 对其进行了全面的综述,他把所有的相关性概念以及模型研究通过四维框架加以描述。

2.1.1 信息资源

包括文献、文献的替代品以及信息。

2.1.2 用户信息需求的描述

包括信息问题、信息需求、查询请求以及查询表达式。

2.1.3 时间

时间维非常典型地体现了相关性判断的动态性。

2.1.4 组件

信息资源和用户信息需求的描述以及二者结合中的所有元素都可以分解为主题(topic)、任务(task)和背景(context)等3个组件。后2个阶段虽然对影响相关性的各个因素都进行了研究,但视角发生了转变,形成了面向用户的相关性研究热点。学者们认为,相关性是论题的、面向用户的、多维

的、认知的以及动态的,并把认知方法和情境的方法引入研究中。用户的相关性判断虽然是一个十分复杂的决定于内部认知和外部情境因素的动态判断过程,但却是具有系统性与可测度性的。

我们认为,相关性研究中形成的面向系统的和面向用户的两个流派分别产生于不同的技术发展阶段和信息环境。前者的应用基础是联机检索系统为主流的历史阶段,那时在检索终端具体执行检索任务的多为专业检索中介或各领域的专业人士;而后的实践平台已发展到因特网时代,广泛的最终用户日益成为检索用户的主体。因此,面向用户的检索相关性研究和实践将是当今信息环境下的主流。

近年来一些实验检索系统和 Web 搜索引擎都从面向用户的视角探索和开发使用了许多提高检索相关性的技术、方法,极大地满足了用户的信息需求。下面就对其具体路径加以归纳。

2.2 提高检索相关度的路径

2.2.1 检索结果排序引入链接分析法

长期以来,检索系统检索结果的相关度排序主要采用词频和位置加权算法。但这种算法在自然检索语言环境下有失准确和客观,且较易为网站优化者们所操纵,他们可较为轻易地把其网页“优化”成“富含关键词网页”(Keyword Rich Pages),使其在检索结果中排在前面。这使检索结果的客观性和准确性受到侵害。链接分析法^[4]在关键词文本匹配的基础上,利用 Web 超链结构,对一个网页与其它网页的链接关系进行分析,以此确定该网页在检索结果中的排列顺序级别。这种方法创造性地应用了引文检索语言的原理。Page Rank 和 HITS 是 2 个基于这种思路的代表算法。Google 和百度正是根据这一技术思路开发并使用了独特的排序算法而分别成为英中文搜索引擎的佼佼者。

2.2.2 利用分类和聚类技术组织检索结果集

检索系统采用分类或聚类技术,自动把查询结果归入到不同的类别中。例如 Yahoo 的输出结果分别列于所从属的各级分类目录下;又如搜索引擎 Northern Light 独创的定制搜索文件夹 (Custom Search Folders) 技术^[4]把与检索提问机械匹配的结果自动组织成不同的类别,用户通过浏览类别,判断选择真正与自己的检索需求相关的检索结果。这种类目列表是多层的、动态的,随着用户一次次地点击、文件夹一层层地展开,用户一步步地筛选下去,逐步明确原本模糊的需求表达,最终得到与其检索需求最为相关的检索结果。

2.2.3 大众点击率作为排序参数

这种方法的典型代表是搜索引擎 Direct Hit。搜索引擎可追踪在每次特定检索过程中用户到底选中了哪些结果,据此最终把排名靠前但点击率不高的网页拉下来,而把最初排名靠后但点击率高、受用户欢迎的网页提到前面。这是因为他们认为多数人访问的网站就是最重要的网站,而网站的重要性则意味着某一方面内容的丰富和准确。这种追踪点击率,由网络大众集体确认网站重要性的方法具有客观性和公正性,在实际使用中有令人满意的效果。

2.2.4 提供信息导引功能

为应对当代信息环境下检索用户的“资源迷向”问题^[6],智能检索系统提供了信息导引功能,帮助用户确定自己需要的信息所在的领域,为用户提供详细、科学的信息分类知识。当用户输入关键词后,系统可推断关键词信息所在的领域,并将这些领域信息告知用户,让用户确定。如果系统不能推断关键词信息所在的领域,它就检查用户模型,看看是否有相关的领域信息,如果有,将这些领域信息告诉用户,让用户确认;如果没有,它要求用户提供领域信息,用户可以根据信息分类知识选择或手工输入领域信息。对于不同的用户,系统将学习用户提交的关键词和领域信息,并记录在其用户模型中;当用户下次提交查询任务时,自动让用户确认相关信息,避免用户重复操作。

2.2.5 基于内容的理解

智能检索系统需建立大规模的知识库,在此基础上对自然语言文本表述的文档主题和查询语句进行语法、语义分析,从而可在内容层面上真正理解并准确描述文档所论述的主题和理解用户的查询意图,即达到对文档内容和用户提问内容的理解,而非标引词与提问词的字面匹配。

特别是随着多媒体技术的发展,人们对图像视频、声音等多媒体信息的查询越来越多,基于内容理解而非文本匹配的检索系统更为人们迫切所需。IBM Almaden 研究中心研究开发的 QBIC 系统就是其中的典型代表,它允许使用例子图像、用户构建的草图和图画、选择的颜色和纹理模式、镜头和目标运动以及其它图形信息等对大型图像和视频数据库进行查询。

2.2.6 联想功能及歧义处理

一些检索系统利用分词词典、同义词典、同音词典来控制自然检索语言;通过主题词典、上下位词典、相关同级词典形成一个知识体系或概念网络,在知识层面或者概念层面上辅助查询;结合歧义知识描述库、全文索引、用户检索上下文分析以及用户相关性反馈等技术处理,高效、准确地提

供给用户最相关的信息。普林斯顿大学认知科学实验室开发的词典数据库系统 WordNet 即是基于英文的词汇语义网络系统^[6],它被成功地用于词义消歧、语言学自动处理、双语及多国语机器翻译等一系列语言工程,较好地改善了检索相关性。

2.2.7 相关反馈技术

传统检索系统不关心用户对查询结果的满意程度,即不关心用户对检索结果的相关性判断,因此若同一用户反复进行同一查询,其结果没有任何改进。相关反馈技术^[7]的开发,可使用户对系统返回的初步查询结果表明满意程度,系统根据用户的反馈对其查询意图、兴趣方向进行推理预测,进而改进检索结果,提高检索的相关性。英国格拉斯哥 (Glasgow) 大学的信息检索小组在 1999 年 11 月到 2001 年 10 月研究开发了一种相关反馈模型,该模型能够解释用户认为文献相关的原因以及用户如何进行相关判断。

3 结语

信息检索相关性问题的复杂性表现在它是由人来判断的,而影响相关性判断的因素又是多维的、认知的和动态的。当今信息环境下,普通人已成为检索用户的主体,这就产生了对面向用户的检索相关性技术的迫切需求。近年来各检索系统纷纷研究开发基于概念查询的智能检索技术,使得系统对文档内容的理解、用户查询计划、意图、兴趣方向的推理、预测更准确,从而使检索结果更切合用户需求或使用户做相关性判断时更容易。虽然深入开发和研究还面临着很多困难,但智能检索系统的实现仍然是提高检索相关性的解决方案,也是检索系统的发展方向。

参考文献 (References)

- [1] 王家钺.信息检索中“相关性”概念的研究[J].现代外语,2001,24(2): 182~191
- [2] 赖茂生,等.计算机情报检索[M].北京:北京大学出版社,1993
- [3] 符绍宏,赵荣,等.信息检索[M].北京:高等教育出版社,2004
- [4] 赵荣,等.搜索引擎检索结果的组织技术[J].情报学报,2004,23(1): 69~72
- [5] 成颖,孙建军.信息检索中的相关性研究[J].情报学报,2004,23(6): 689~696
- [6] 詹卫东.WordNet 简介.2003-06
http://ccl.pku.edu.cn/doubtfire/Course/Computational%20Linguistics/contents/In-tr2WordNet_zwd20030630.pdf
- [7] 信息检索的核心支撑技术.
<http://www.21cnbj.com/industrynews/searchengine2003/2003-08-11-212.html>

(责任编辑 王 芷)