近年来国外信息检索的相关性研究进展

庞弘燊　徐文贤

摘　要　国外的相关性研究至今已有上百年的历史,出现了两个主要的研究流派,即面向系统的相关性研究和面向用户的相关性研究,相关性是动态的、多维的、可认知的和可测度的等观点,已成为学术界的共识。1997 年之后的相关性研究在基础研究和实证研究方面都有很大进展。信息检索的相关性研究已经深入到信息检索的各个领域,如模型、算法、聚类、查询扩展与精化、相关性判断等。随着相关性研究的不断深入发展,信息检索系统将会越来越贴近用户的信息需求。图 1。表 2。参考文献 20。

关键词　信息检索　相关性　国外

分类号　G252

ABSTRACT　This paper reviews the history of the relevance research for information retrieval abroad which appears two orientations, as the relevance research on system and the relevance research on users. It has become the consensus of the academic community that relevance is dynamic, multidimensional, perceptible and measurable. There is a great development of basic and practical research on relevance after 1997. The relevance research has penetrated into all areas of information retrieval, such as models, algorithms, clustering, query expansion, refinement and the relevance judgments. With the development of relevance research, information retrieval system will meet users' needs closely. 1 fig. 2 tabs. 20 refs.

KEY WORDS　Information retrieval.　Relevance.　Abroad.

CLASS NUMBER　G252

1　信息检索相关性研究的历史

信息检索相关性的研究经历了比较长的时间,可以划分为不同阶段,各个阶段的研究重点也有不同。Stefano Mizzaro 1997 年在他的分析中认为,相关性研究从起始到 20 世纪 90 年代中期,大体上可以划分为 3 个阶段:①1958 年之前;②1959—1976 年;③1977 年至 20 世纪 90 年代中期。其研究内容可以归纳为 7 个方面,即方法论基础、相关性类型、由用户所采纳的优于主题性的判断标准、描述相关性判断的模型、相关性动态属性、文档类型和相关性判断标准[1]。

第一阶段(1958 年以前):相关性的历史可能起源于 100 年前,并且首先出现在图书馆,图书馆用户很早就已经关注到寻找相关信息的问题了。这个阶段具有隐性的特点,在相关性概念的背后有很多研究,但都比较肤浅,几乎没有人能明确地阐述这一主题。这一阶段结束于 1958 年,以该年举行的科学信息国际会议(ICSI)为标志,在此次会议上人们开始明确地认识到相关性的概念。

第二阶段(1959—1976 年):这一阶段掀起了相关性研究的第一次高峰,出现了产生重大历史影响的大型实证研究以及一系列理论研究。结束的标志是 Saracevic 在 1975 年和 1976 年发表的相关性研究阶段性综述,综述对前人提出的相关性概念进行了归类。该阶段的实证研究主要包括 1955 年由 Kent 等提出的用查全率和查准率等指标进行的检索系统评估、Cranfield 测试(由 1957 年的 Cranfield I 和 1962 年的 Cranfield II 两个项目组成)、Cuadra 与 Katter(1967 年)和 Rees 与 Schultz(1967 年)两个小组分别展开的测试。这些实证研究都是从多个方

面对相关性进行研究的,而不仅仅局限于Mizzaro所述7个方面中的某一个。并且,这些实证研究的文章和Saracevic的文章都被下一阶段的文章所大量引用。

第三阶段(1977年至20世纪90年代中期):自20世纪90年代开始,相关性研究迎来了第二次高峰。该阶段研究的特点是从面向系统的观点逐步转移到面向用户方面,更多的是从用户认知的角度开展研究。影响比较大的有1994年的*Journal of the American Society for Information Science*(JASIS)相关性专辑,包括Schamber、Park、Barry和Wang等人在内的4篇博士论文,以及Schamber、Froehlich、Saracevic、Mizzaro、Borlund等人发表的综述与述评。

综观国外针对相关性的研究,从20世纪30年代算起,至今已有70多年的历史,其间出现过两个主要的研究流派,分别是面向系统的和面向用户的相关性研究。研究高峰分别集中于20世纪60年代至70年代前期和80年代中后期至今两个阶段。相关性是动态的、多维的、认知的和可测度的等观点,已经成为学术界的共识。为了更好地组织和梳理异彩纷呈的相关性研究,Saracevic、Mizzaro以及Borlund等人分别建立了各自的模型,以利于从整体上对相关性研究加以认识。在国外对相关性研究的文章当中,笔者发现实证研究是采用的最基本手段。

2　相关性研究

2.1　近年来国外相关性研究概况

通过对外文全文数据库(ScienceDirect、WileyInterScience、Springer LINK)的检索,在1997—2007年这一时间段检索出70多篇有关相关性研究的外文文献,分析论文发表的时间分布发现,相关性研究在2003年前后以及2007年呈现出研究的高峰期(见图1)。对文章的关键词进行分析,在70篇文章当中总共含有354个关键词(其中有1篇文章没有关键词标引),其中关于相关反馈、查询扩展、相关性排名、检索效率等方面的研究是比较多的(见表1)。

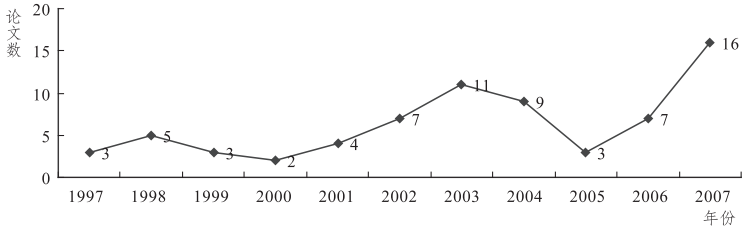

图1　国外相关性研究论文发表的时间分布

表1　国外相关性研究论文中词频数较高的关键词词表

关键词	词频
信息检索(Information retrieval)	45
相关性(Relevance)	27
相关反馈(Relevance feedback)	24
查询扩展(Query expansion)	7
相关性排名(relevance ranking)	6
检索效率(retrieval effectiveness)	6
文档检索(document retrieval)	5
遗传算法(Genetic algorithm)	5
评价(Evaluation)	4
信息需求(information needs)	4
查询公式(query formulation)	4

2.2　相关性的基础性研究

相关性(Relevance)一词在英语词典中解释为"the relation of something to the matter at hand"。从词义理解,就是指事物之间彼此的联系。近年来,关于相关性的基本定义、相关性的属性类别、相关性模型等基础性的研究都有进展,但大多都是根据以往的研究展开的。

Stefano Mizzaro在1997年认为,相关性存在很多种类型,而不仅只有一种。普遍认同的相关性是由两个实体的联系所组成,其中一类实体就是文档、文档替代物和信息等,另一类实体则是指问题、信息需求、检索请求、查询提问等。

相关性可以看作是这两类实体之间的组合联系,从每类实体当中各选取一个元素组合而成,比如文档与检索请求的相关性,信息与用户信息需求之间的相关性等[1]。

Stefano Mizzaro 在随后一年对研究作出了改进[2],认为相关性的诸多种类都可以被正确地归类到一个由他定义的四维空间当中,该四维空间的各维度内容分别是信息资源、用户信息需求、时间和构件(Component),这些维度既包括了系统性相关的因素,也包括用户性相关的因素,这样的归类有助于理解相关性的类别和相关性判断。然后他在此基础上提出了一个形式化的四维理论框架。该理论框架不仅具有很好的形式化表示,而且总结和吸收了很多有关相关性研究的成果,有助于人们更好地理解信息检索的本质,对进一步改进信息检索系统的设计与评价工作具有重要的理论指导意义。

Hongseok Park 在 1997 年为了找出相关性的维度和特性,对 24 名研究生的相关性判断进行了观测[3]。通过实验和数据收集发现:①相关性是多层面的;②存在有两种相关性维度——主要的和次要的;③相关性维度包含三个方面:问题、使用和价值;④问题维度方面相对其他两个方面来说是主要的。这些研究结果的影响体现在:①发现了四个重要的相关性维度的特性;②重要的相关性维度必须适用于系统中相关性的测量;③词库中词条的关系必须依据重要的相关性维度来设置;④重要的相关性维度对于有效地评价信息检索是很有用的;⑤由于相关性判断具有易变性,以上要点的提出使得在研究中能更有效地观测到用户的相关性判断。

Erica Cosijna 和 Peter Ingwersen 认为,各个学科领域已经开始尝试研究相关性概念,并提出了许多理论框架,然而,相关性也是一个很主观的概念,是很难界定的[4]。他们对 Saracevic 在 1996 年文章中提出的相关性属性和相关性类型之间的关系作了分析,并改进了其中的不足之处,以新的描述(表 2)来表现相关性属性和相关性类型之间的关系,从而更深入地揭示了相关性这一概念。

表 2 修正后相关性属性与类型之间关系的列表

相关性属性	相关性类型				
	系统/算法	主题	认知/有关	情境/有用性	社会认知
关系	查询表达式 = > 信息客体(基于特征)	用户查询表达式 = > 信息客体	知识状态/认知信息需求 = > 信息客体	已经感知到的情境、任务 = > 信息客体	所了解的社会领域、任务、文化背景 = > 信息客体
意图	①系统依赖的 ②隐藏在算法中的意图	①用户/评估者的期望 ②隐藏在查询表达式中的意图	高度个性化的、主观的、且与信息需求相联系的	高度个性化的、主观的、且与工作任务相联系的	高度个性化的、主观的、或者情绪化的
背景	调整搜索引擎性能(例如 TREC)	所有主观相关性类型的定义都是和用户或者评估者的背景相关的(用户/评估者的背景)			
判别方式	权重或者排序函数	语义层次的主题解释	对认识进行解释、选择以及过滤的过程	用户根据其需要利用信息客体的能力	在对所处环境有意义的背景下用户(组织)利用信息客体的能力
交互	自动的相关反馈或者查询修正	依赖于查询内容的相关性判断	依赖于查询的内容、特征、形式以及表述的相关性判断	与环境之间的交互	在所处环境当中的交互
情绪相关性:依赖于时间的推移					

Pia Borlund 认为,相关性所具有的多层面性主要是通过用户引用各种相关性标准、判断相关性检索信息对象而描述出来的[5]。他在文章中参考了许多相关性概念而展现其多层面性,如相关性的分类、类型、程度、标准和层次等方面。在概述了多种相关性概念并考虑到动态相关性问题的基础上,他提出了关于相关性的框架,表明人们对相关性认识已经达到了一个较高的水平。他还认为相关性是基于认知的角度,并且其中的相关性判断是通过信息检索这一互动过程而慢慢改进的。从认知的意义上说,信息作为一种精神上的构思,相当于对检索到信息对象的相关性进行评估,而与任何主观想象的相关性无关。

Tefko Saracevic 认为,从直观上说,大家所普遍理解的相关性都是围绕着关系的相关性展开的,都涉及到明确地提及或者暗含"关于"这个词的,可以说相关性是一个完全由用户定义的概念,这使得它既具有优势,同时也存有一定的不足。而从直观理解之外来认识相关性,它可以被视为一种事物各个组成部分之间关系的特性,也可以被视为是衡量关系之间联系的强度。在信息科学领域,相关性就相当于关系和测量[6]。

对于相关性的认识已经越来越深入,国外不仅仅是从相关性的概念去单纯地研究,更多的是深入到其属性来作分门别类的研究,还有的是构建出模型来描述相关性,并且大多研究都是结合实践,通过实验和调查等手段来进行分析。因而,笔者认为,近年来国外对相关性的基础性研究已经进入成熟的阶段,并且对相关性已经有一个相当完整的描述和比较一致的认识。

2.3　关于检索系统的相关性研究

在相关性的实证研究中,很多都是结合信息检索系统展开的,目的在于改进信息检索系统的性能,提高用户检索的效率以及检索出文档的相关程度。可以说这一部分研究是构架在面向用户的相关性基本理论之上的,通过研究面向系统的相关性,使得信息检索系统的建立能更符合用户的信息需求。

2.3.1　检索系统模型的研究

检索系统中所采用的模型类型有布尔检索模型、向量空间检索模型、概率模型、逻辑模型等,通过改进检索系统中的模型构建,可以有效地从根本上改善系统的相关性。

David E. Losada 和 Alvaro Barreiro 在 2003 年曾提出一个新的方法[7],把词项的相似性与倒排文件频率放在一起,组成一个信息检索的逻辑模型。在这信息检索的逻辑模型中把文档和查询作为逻辑公式,并运用了一些推理形式所提供的逻辑来决定相关性。

Leah S. Larkey 和 Margaret E. Connell 在 2004 年撰文,认为目前跨语言检索的两种概率方法已被广泛使用,有 INQUERY 这种基于相关性概率模型,还有基于语言模型的[8]。他们通过比较传统的扩展技术(伪相关反馈)和相关性模型,发现一种新的信息检索办法适用于纳入语言模型的正式框架当中,同时发现相关性建模和伪相关反馈可以实现同等水平的检索,并且良好的翻译概率显现出一个微小却很显著的优势。通过比较两种查询扩展的方法,他们发现相关性建模和伪相关反馈性能是一样的。对于西班牙语,而不是阿拉伯文数据,相关性建模明显优于伪相关反馈。然而,相关性建模的处理速度却明显慢于伪相关反馈的处理速度,因此目前还不清楚相关性建模是否具有实用的优势。

Gerald Benoît 通过实验表明,如果向量模型的排名列表足以满足用户的需要,那么检索系统相关性排名和用户的最终检索要求,以及各用户的检索结果之间都不会产生太大的差异[9]。

2.3.2 算法的研究

检索系统中信息反馈的性能,更多的依赖于其所采用的算法。近年来,遗传算法已被应用到检索系统当中,并且已被证明使用遗传算法对检索系统的相关反馈性能有显著的改善。

Cristina López-Pujalte 等人通过第二次实验来测试遗传算法在相关反馈中的作用,实验表明,在检索过程中使用遗传算法的性能比普通

方法提高了7%,最高甚至可以达到27%[10]。

2.3.3 聚类研究

聚类的基本思想就是把相似的文献归为同一类目,通过对聚类的研究,可以更有效地把检索的文献集合划分成相关文档和不相关文档两类。

Niall Rooney 等人在文献[11]中提到,上下文文档聚类是一种新颖的方法,它利用了信息论的方法来聚类语义相关文件并绑定了一个隐含的概念或特定的主题。通过评估查询和集群主题的概率分布之间的相似性,对基于集群的检索是有帮助的。他们还通过基于查询精化的方法来衡量相关反馈机制,即通过使用少数文档来修正查询概率的分布,用以判断查询的相关性。实验证明,如果只提供一种相关性的判断,检索效率会得到33%的提高。

2.3.4 查询扩展和精化的研究

查询扩展和精化,就是在用户浏览的过程中,通过相关反馈逐步对查询进行扩大和精化,并且可以允许用户在必要时对查询进行修正。而在检索系统中,加入相关反馈机制的目的,是通过检索策略的调整来增强对相关文献的响应而抑制非相关文献。

Koji Eguchi 等人制订了一种自适应聚类的方法[12],利用聚类的观点来精化查询信息,用以确保用户所关注的信息能反映在检索的过程当中。他们认为查询精化同样适用于万维网中的信息检索,并且通过实验证明,通过查询精化能把相关文档资料与非相关文档更好地分离出来,而检索效率也能得到有效的提高。但是这类改进,并不取决于要处理文档的大小,而是在很大程度上依赖于查询的质量。提高检索效率的因素有两个:①利用查询精化的自适应聚类方法的性能;②在聚类中基于精化查询相似度的相关排名列表的改进程度。

Fernando Martínez-Santiago 等人通过实验分析了把伪相关反馈(PRF)技术应用于分布式信息检索(DIR)的性能。实验表明,如果把 PRF 技术应用到全局的级别上,系统相关反馈的性能要比普通的本地系统反馈高出很多[13]。

Miles Efron 试图通过研究找出潜在语义索引(LSI)和 Rocchio 相关反馈这两者何者最优。这两种技术都是经典向量空间信息检索模型的扩展,都是基于最小二乘法的,而其共同点是众所周知的,即阐明了相关反馈和 LSI 之间新的关系。他的观点表明,该方法的不同之处并不是简单通过查询扩展来对检索到的文档进行扩展。Rocchio 相关反馈是根据内容进行线性判别分析,而 LSI 则是对主要成份中最紧密关系的分析。这种区别使得这两种技术在不同的问题当中各有所长:相关反馈较 LSI 在文档分类中能提供一个较好的一维投影;另一方面,LSI 可以提供一个 k 维不相关的准则来代表文件并含有较少差错。因此,当出现词间关系离散程度高的情况下,LSI 的维度减少使他可以构造一个线性模型,就像回退到变量更少的状态中去[14]。

Olga Vechtomova 和 Murat Karamuftuoglu 通过实验表明,从用户选取的短摘要(snippet)中抽取词项来进行查询扩展,比从全文中抽取词项来进行查询扩展效果更好,查询出的文档具有更好的相关性[15]。

2.3.5 相关性判断的研究

Saracevic 是第一位提出把用户的相关性判断作为信息检索系统性能评价指标的人[16],而近年来的研究也表明相关性判断对信息检索系统的检索效率有显著的影响。

Spink、Greisdorf 和 Bateman 等人对相关性用于评价信息检索系统的作用进行了很多研究,并取得了一系列令人瞩目的进展,主要有:①他们扩展相关性判断分布区域(middle range)为3个:完全相关、部分相关和部分不相关[17];②相关性判断分布对用户就某一问题早期的检索起着重要作用[18];③在实验中采用了相关性判断的中值(medium measure)来评价信息检索系统的性能,作为查全率和查准率的补充[19]。以上研究第一次把长久以来将相关性用于信息检索系统评价的理论探讨运用到实践中。

Amanda Spink 和 Howard Greisdorf 后来还通过调查来研究在各学科领域中用户的相关性判断[20],其中包括怎样对这些领域进行归类、测量和评价。实验分4个等级来收集、测量和描述终端用户的相关性判断,实验的对象是21名终端

用户,经由他们搜索自己的信息问题,并在1059条检索项目中作出了相关性判断。调查结果显示:①相关性的重叠区域,对于通过信息检索精确率的有效性来衡量检索系统的效率是有影响的;②当用户作出相关性判决时,相关性的积极和消极层面都是很重要的;③当相关度非常高的时候,主题性更多的时候是用来否决检索到的条目的;④有用性更多的时候是用来判断条目是否具有很高的相关度;⑤性质的相关性判断分布表明了一种新的信息检索评价中值测量方法的效果。

通过以上关于检索系统的相关性研究,可以发现更多的相关性理论已经应用到了检索系统当中,并已通过实验等方式得到验证,因此可以认为,国外对相关性的研究已经与检索系统的研究结合到一起,并通过各个领域的研究来促进双方的共同发展。

3 结语

近年来,国外对信息检索中相关性这一范畴的研究,已经深入到信息检索的各个领域,可以说呈现了一个百花齐放的研究态势。鉴于此种情况,也可把1997年之后的研究划为相关性研究的第四个阶段。因为这一时期的研究,与1997年之前的三个研究阶段有较大差异,既有面向用户的相关性研究,也有面向系统的相关性研究;同时,研究范围得到较大的扩展,信息检索的模型、聚类、文摘、查询扩展和精化等各个角度,都有关于相关性的研究,而这个阶段也秉承了第三阶段的研究风格,在面向用户的研究方面继续深化。

从国外关于相关性的基础性研究来看,很多研究都是面向用户的,该类型的相关性是主观的,通过研究用户的信息需求,来提高信息检索相关性。而关于检索系统的相关性研究,则普遍提出一些新技术或是新方法,还有就是通过改善已有的技术和方法来提高信息检索系统的性能,以期能符合用户的信息需求,从系统观的角度来改善和提高信息检索系统的相关性。

随着相关性研究不断地深入和发展,可以预见,未来的信息检索系统将会越来越贴近用户的信息需求,功能会越来越完善,除了能更好地获取用户的需要外,还能以各种方式反馈给用户最有用的信息。

参考文献:

[1] Stefano Mizzaro. Relevance: The whole history [J]. Journal of the American Society for Information Science, 1997, 48(9): 810 - 832.

[2] Stefano Mizzaro. How many relevances in information retrieval [J]. Interacting with Computers, 1998, 10(3): 303 - 320.

[3] Hongseok Park. Relevance of science information: Origins and dimensions of relevance and their implications to information retrieval [J]. Information Processing & Management, 1997, 33(3): 339 - 352.

[4] Erica Cosijn, Peter Ingwersen. Dimensions of relevance [J]. Information Processing & Management, 2000, 36(4): 533 - 550.

[5] Pia Borlund. The concept of relevance in IR [J]. Journal of the American Society for Information Science and Technology, 2003, 54(10): 913 - 925.

[6] Tefko Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science, Part II: nature and manifestations of relevance [J]. Journal of the American Society for Information Science and Technology, 2007, 58(13): 1915 - 1933.

[7] David E. Losada, Alvaro Barreiro. Embedding term similarity and inverse document frequency into a logical model of information retrieval [J]. Journal of the American Society for Information Science and Technology, 2003, 54(4): 285 - 301.

[8] Leah S. Larkey, Margaret E. Connell. Structured queries, language modeling, and relevance modeling in cross-language information retrieval [J]. Information Processing & Management, 2005, 41(3): 457 - 473.

[9] Gerald Benoît. Properties-based retrieval and user decision states: User control and behavior modeling [J]. Journal of the American Society for Information Science and Technology, 2004, 55(6): 488

-497.

[10] Cristina López-Pujalte, Vicente P. Guerrero-Bote, Félix de Moya-Anegón. Genetic algorithms in relevance feedback: a second test and new contributions[J]. Information Processing & Management, 2003,39(5):669-687.

[11] Niall Rooney, David Patterson, Mykola Galushka, Vladimir Dobrynin. A relevance feedback mechanism for cluster-based retrieval[J]. Information Processing & Management, 2006, 42(5): 1176-1184.

[12] Koji Eguchi, Hidetaka Ito, Akira Kumamoto, Yakichi Kanata. Adaptive document clustering using incrementally expanded queries[J]. Systems and Computers in Japan,2001, 32(2):64-74.

[13] Fernando Martínez-Santiago, Miguel A. García-Cumbreras, L. Alfonso Ureña-Lòpez. Does pseudo-relevance feedback improve distributed information retrieval systems [J]. Information Processing & Management,2006,42(5):1151-1162.

[14] Miles Efron. Query expansion and dimensionality reduction: Notions of optimality in Rocchio relevance feedback and latent semantic indexing[J]. Information Processing & Management, 2007, 43(5):1294-1307.

[15] Olga Vechtomova, Murat Karamuftuoglu. Query expansion with terms selected using lexical cohesion analysis of documents[J]. Information Processing & Management, 2007,43(4):849-865.

[16] Saracevic,Tefko. Comparative effects of titles, abstract and fulltext in relevance judgements[J]. Proceedings of the 1969 annual meeting of the American society for information science,1969(6): 293-299.

[17] Amanda Spink, Howard Greisdorf, Judy Bateman. Form highly relevant to not relevant: Examining different regions of relevance[J]. Information Processing & Management,1998,34(5):599-622.

[18] Amanda Spink, Judy Bateman, Howard Greisdorf. Successive searching behavior during information seeking: an exploratory study[J]. Journal of information science,1999,25(6):439-449.

[19] Howard Greisdorf, Amanda Spink. A new way to evaluate IR systems performance-median measure [C]. Proceedings of NOM 2000, New York.

[20] Amanda Spink, Howard Greisdorf. Regions and levels: Measuring and mapping users' relevance judgements[J]. Journal of the American Society for Information Science and Technology,2001,52(2):161-173.

庞弘燊 中国科学院国家科学图书馆情报学博士研究生。通讯地址:北京。邮编100190。

徐文贤 华南师范大学图书馆副馆长,研究馆员,博士,华南师范大学经济与管理学院硕士生导师。通讯地址:广州。邮编510631。

(收稿日期:2008-10-14;修回日期:2008-11-14)