SIGIR Forum Special Issue
“SIGIR 通讯”特辑

编辑：Donna Harman, Diane Kelly
作者：James Allan, Nicholas J. Belkin, Paul Bennett, Jamie Callan, Charles Clarke, Fernando Diaz, Susan Dumais, Nicola Ferro, Donna Harman, Djoerd Hiemstra, Ian Ruthven, Tetsuya Sakai, Mark D. Smucker, Justin Zobel.

译者：李祥圣 罗成 刘奕群 马少平

译者按

2017 年 8 月，第四十届国际计算机学会信息检索大会（ACM SIGIR）在日本东京召开，本次会议设立了一个特别的“经典论文奖”（Test of Time Award）环节，奖励那些在信息检索领域发展早期（1978 年到 2001 年之间）发表的经历了时间考验的优秀论文。与此同时，“SIGIR 通讯”杂志也出版了一期特刊，逐一回顾了这些论文的学术贡献，以及他们“历久弥珍”的价值。特刊邀请了一批在信息检索领域的资深专家，对每篇获奖的论文进行了简要的评述。我们将这些评述译成中文，希望可以与更多相关领域的研究人员一同回顾以 SIGIR 论文记录的信息检索学科的发展历程。

引言

ACM SIGIR 会议已经走过了 40 年的历程。本期 SIGIR 通讯特辑回顾了在早期 SIGIR 会议(1978-2001)中发表的那些”历久弥珍”的论文。这些论文共同见证了信息检索研究与实践的发展历史与演进过程，同时也反映了 ACM SIGIR 会议在信息检索学科发展中产生的巨大影响。

ACM SIGIR Test of Time Award 主要表彰那些对信息检索研究产生深远影响的会议论文。最初的方案是奖励在颁奖年之前 10 到 12 年发表的论文。例如，在 2014 年将表彰在 2002 至 2004 年之间发表的论文（译注：由于奖项是近年设立的，因此早期 SIGIR 会议的不少论文并没有被纳入评奖的范畴内）。为了更好地遴选那些在 1978 年到 2001 年之间发表的杰出论文，我们成立了一个由 Keith van Rijsbergen 教授（译注：ACM 会士, 英国皇家工程院院士，ACM SIGIR Salton 奖得主）领导的评选委员会。这个委员会还包括以下学者：Nicholas Belkin（译注：

ACM SIGIR Salton 奖得主，美国罗格斯大学教授）, Charlie Clarke（译注：ACM SIGIR 执行委员会前主席，FaceBook 公司研究员）, Susan Dumais（译注：ACM 会士, ACM SIGIR Salton 奖得主，微软公司杰出科学家）, Norbert Fuhr（译注：德国杜伊斯堡埃森大学教授，ACM SIGIR Salton 奖得主）, Donna Harman（译注：美国国家标准与技术研究院科学家）, Diane Kelly（译注：ACM SIGIR 执行委员会主席，田纳西大学教授）, Stephen Robertson（译注：ACM SIGIR Salton 奖得主，英国剑桥哥顿学院院士）, Stefan Rueger（译注：英国开放大学教授）, Ian Ruthven（译注：英国思克莱德大学教授）, Tetsuya Sakai（译注：日本早稻田大学教授）, Mark Sanderson（译注：澳大利亚皇家墨尔本大学教授）, Ryen White（译注：微软公司 Cortana 团队技术研发负责人）和翟成祥（译注：美国伊利诺伊大学香槟分校教授 ）。

评选委员会首先基于引用数等标准建立了一个候选论文集合，同时也向所有委员们征求论文提名，对候选论文集合进行补充。在上世纪 80 年代比较活跃的一批资深的信息检索学者还成立了一个特别委员会，审核在 1978 年到 1989 年之间发表的优秀论文。最终，候选论文集合中的每一篇论文都交由一个三位委员组成的小组来进行评分。得分最高的 30 篇论文获得了“历久弥珍奖”，即经历了时间考验的优秀论文（Test of Time Award）。

为了纪念 1978 年到 2001 年 ACM SIGIR Test of Time Award 的获奖论文，我们邀请了一批信息检索领域的资深学者对其中的每一篇论文进行简评。每篇论文的简评首先**概括总结了论文的主要贡献**，也阐述了**这篇论文为何对今天的信息检索研究仍然具有价值**。除了少数版权不在 ACM 的论文，本期特刊包括了所有获奖论文的翻印版（译注：并不包含在本翻译稿内），读者亦可通过 ACM Digital Library 里的原始会议论文集访问这些论文。

作为评审委员会的成员，我们在阅读这些历久弥珍的论文时受益匪浅。不少论文的写作风格与目前 SIGIR 会议论文风格并不完全相同。他们的写作简洁而朴实，又有着高度的创新性、严谨性和开放性。我们盼望每位读者都可以通过阅读这些论文去仔细地回顾信息检索学科的发展历程，并从中受益。我们也鼓励所有信息检索领域的同仁们随我们一同回顾这些经典论文的内容，并且考虑**如何使这些工作中蕴含的宝贵财富贡献于我们当前的研究问题**。

Test of Time 获奖论文, 1978-2001

- Stephen E. Robertson, C. J. (Keith) van Rijsbergen, and Martin. F. Porter. 1980. **Probabilistic models of indexing and searching**. In Proceedings of the 3rd annual ACM conference on Research and development in information retrieval (SIGIR '80). Butterworth & Co., Kent, UK, 35-56. ACM: http://dl.acm.org/citation.cfm?id=636673

评论人：Djoerd Hiemstra（译注：荷兰特温特大学教授）

概要：这篇论文主要关注和比较了利用概率检索模型估计查询词词项权重的方法。作者比较的方法包括：文档频度倒数（IDF）方法，二值独立模型（译注：Binary Independence，即由 Robertson 和 Sparck Jones 提出的词项权重估计模型，又称 RSJ 模型）的几种不同实现方法，以及二维泊松模型（译注：该模型假设每个查询词的权重受到两个分布的影响，即与其最相关的一部分文档集合中的频度分布，以及其他文档组成的文档集合中的频度分布）的几种不同实现方法。本文还基于二维泊松模型提出了一种新的词项权重估计方法。实验结果显示，使用作者所提出的二维泊松模型估计词项频度会带来一定的效果提升，但是比起其他更简单的估计方法并未有明显的性能优势。

贡献：这篇文章的主要理论贡献是：提出了一个实际可行的基于二维泊松模型的查询词权重估计方法。二维泊松模型是在本文发表之前 2-3 年由 Abraham Bookstein, Don Swanson 和 Stephen Harter 等人提出的。本文另一方面的贡献在于，作者在两个不同的数据集合上开展了近 50 组不同条件下的比较实验。是否使用相关性反馈信息、如何对训练集和测试集进行划分等因素都被纳入实验比较的范畴。

历久弥珍的价值：本文说明了正确地发表负面结果的重要性。文中对二维泊松模型进行了充分的理论分析和实验验证，进而对该模型的有效性做出了结论性的判断：二维泊松模型并不会带来搜索质量的改善。作者指出，估计二维泊松模型的众多参数是十分困难的，相应的参数估计问题应当在未来的工作中被更多重视。作者指出的这一未来工作方向并不是信息检索研究中无关紧要的注脚，与此相反，自 1980 年以后，Robertson 教授在如何使二维泊松模型更加有效应用于信息检索问题的方向上投入了大量的精力，并最终催生了 BM25 词权重估计模型。关于 BM25 模型的更多信息，包括 Robertson 与 Stephen Walker 在 SIGIR 1994 中的相关论文，我们将在后续介绍 SIGIR 1994 Test of Time Award 时再行细致讨论。

- Stephen E. Robertson, M. E. (Bill) Maron, and William S. Cooper. 1982. The unified probabilistic model for IR*. In Proceedings of the 5th annual ACM conference on Research and development in information retrieval (SIGIR'82). Springer-Verlag New York, Inc., New York, NY, USA, 108-117. DOI:https://doi.org/10.1007/BFb0036342ACM: http://dl.acm.org/citation.cfm?id=636723

评论人：Djoerd Hiemstra（译注：荷兰特温特大学教授）

概要：这篇论文提出了一种新的信息检索模型。这个模型统一了两种既有的模型：一种是由 Maron 和 Kuhns 在 1960 年提出的概率索引模型 Probabilistic indexing model [17]，记为模型 1；以及由 Robertson 和 Sparck-Jones 在 1976 年提出的概率检索模型 Probabilistic retrieval model [22] 记为模型 2。统一模型 1 和模型 2 的新检索模型，即模型 3，能够在判断文档相关性的过程中更好地利用两类交互数据：每一个独立用户对于多篇文档的交互数据，和多位用户对于同一篇文档的交互数据。

贡献：模型 1 和模型 2 及其各自的变体已经在包含信息检索在内的诸多领域取得了广泛的成功。如今它们最为人们熟知的名字分别是统计语言模型 Statistical language models 和朴素贝叶斯分类器 naive Bayes classifiers。有趣的是，在作者提出这些理论的时候，事实上很难收集到大规模的用户行为数据，直到近些年商用搜索引擎开始记录用户的查询和点击后，这类用户与文档的交互数据才比较丰富。 迄今为止，“从理论层面建立统一的检索模型“仍然没有得到很好的解决。从这个角度来说，这篇 35 年前的论文于今天的信息检索研究仍有借鉴意义。

历久弥珍的价值：在这篇文章发表的时代，SIGIR 还是一个不太正式的会议。这篇文章是那个时代的一个美好的符号。这篇文章通过对前人工作的回顾提供了研究问题的清晰的上下文背景。该文的第一作者开玩笑的说，他应该对文中的主要不足负责，因为他与他的共同作者相隔 6000 英里。论文希望读者来批判性地讨论这种统一的模型可能存在的问题。以如今严格的审核标准来看，他提出的一些思路和假设很可能因为缺少依据而不成立，但后来的事实证明，这些想法也是推动学术前进不可或缺的部分。这篇文章的一个不太正式的拓展版后来发表在 Information Technology [21]上。

- Jeffrey Katzer, Judith Tessier, William Frakes, and Padima Das-Gupta. 1983. A study of the overlap among document representations. In Proceedings of the 6th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '83). ACM, New York, NY, USA, 106-114. DOI: https://doi.org/10.1145/511793.511809

评论人： Nicholas J. Belkin（ACM SIGIR Salton 奖得主，Rutgers 大学教授）
概要：这篇论文主要调研了使用不同文档表示（译注：这里的“文档表示”主要关注两类文档表示方法，一种是文档中的原文，即 free-text；另外一种是索引人挑选的描述文档的短语，即 descriptor）对于检索结果的影响。作者首先通过 PsychInfo 中的 45 名参与者收集了 52 个采用自然语言描述的搜索任务。4 位参与后续实验的被试，在实验系统中尝试完成上述的搜索任务。实验系统对不同的搜索任务隐式选取了不同的文档表示方式。实验结果表明，采用不同的文档表示方法进行检索时，不管采用严格或相对宽泛准则去度量相关性，检索得到的文档集合在精度、召回等指标上的表现都是比较接近的。但同时也发现对于同一检索任务，采用不同的文档表示方法得到的检索结果之间的重合度较低。

贡献：这篇论文是第一个聚焦**不同的文档表示方法对检索系统和检索结果影响**的研究。如果从精度、召回等指标上来看，采用不同的文档表示方法对检索系统的性能影响很小。但实际检索得到的文档列表的重叠度却很低。作者首次发现了**文档表示方法对于检索系统的直接影响**。这个发现是许多后续研究工作的一个基石，同时也间接地为后续的一些理论创新奠定了基础，例如多元表示方法（译注：Principle of Polyrepresentation，即从不同人的不同认知角度、和同一文档的不同功能属性对文档进行表示）等。对于搜索评价相关的研究而言，这篇文章也十分重要，它指出了对于理解和对比信息检索系统这项任务，**仅仅知道一些数字是远远不够的。**

历久弥珍的价值：这篇文章为我们树立了一个很好的典范，我们始终需要批判地去看待信息检索研究中已经存在的知识体系，**特别需要关注那些看似简单的结果背后的事实**。在这项工作中，作者一开始关注于理解“精度”和“召回”到底在度量什么，如何更好地对比系统的性能。通过对细节的分析，他们发现了文档表示方法的影响。长远来看，后者是意义更加重大的研究发现。这篇论文因此也成为实际意义远超研究者预估的一个很好的例子。最为重要的是直到今天它仍然是一篇值得阅读的基石性研究。

- Ellen M. Voorhees. 1985. The cluster hypothesis revisited. In Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR ’85). ACM, New York, NY, USA, 188-196. DOI: https://doi.org/10.1145/253495.253524

评论人：Donna Harman （美国国家标准与技术研究院）
概要：这篇论文提出了一套新的检验聚类假设（cluster hypothesis） 是否正确的方法（译注：这里的聚类假设指 20 世纪 70 年代由 Jardine 和 van Rijsbergen 等人提出的“文档之间的关系暗含了文档的相关性内容”，即首先对文档进行聚类，在聚类得到的簇上检索可以更加高效地获得好的结果）。作者在四个数据集（MED, CACM, CISI 和 INSPEC）上采用聚类的方式进行了一系列的检索和测试。具体地，作者采用了单链聚类的方法（译注：Single-link clustering，即采用两个簇之间最近的元素距离来度量簇间距离）并尝试在簇上采用了两种不同的搜索策略：一种尝试去检索所有的簇，直到获得某一特定数量的搜索结果；另外一种尝试仅检索每个簇中

排在最前位置的文档。作者将这两种搜索与简单的顺序搜索进行了比较，发现即便在一些数据集上聚类假设成立，聚类搜索的效果相对于顺序搜索也较差。

贡献：在 20 世纪八十年代，计算机的运算速度还比较慢。出于对效率的追求，研究者进行了大量关于聚类搜索的研究，一些研究的结论是在聚类得到的簇上进行搜索会更加高效。这篇论文在几个差异较大的数据集上研究了这个问题，他们得到的结论是在聚类的过程中，**把相关的文档聚在一起和把不相关的文档排除在外同样重要**。在实际应用中，这往往和数据集与每个查询的性质紧密相关。这就解释了为什么通常意义下顺序搜索会获得更好的结果。在这篇文章的影响下，后续的研究工作逐渐远离了聚类搜索这一话题。

历久弥珍的价值：这篇论文对聚类假设进行了深入的研究，不仅关注如何评价聚类假设，同时分析了聚类假设和搜索质量之间的联系。文章在四个各不相同的数据集上对这些因素进行了深入的分析，这也是一个早期的关注聚类如何影响搜索质量的研究工作。

- C. J. (Keith) van Rijsbergen. 1986. (invited paper) A new theoretical framework for information retrieval. In Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '86). ACM, New York, NY, USA, 194-200. DOI: https://doi.org/10.1145/253168.253208

评价人：Nicola Ferro（意大利帕多瓦大学副教授，欧洲信息检索评测组织 CLEF 协调人）

概要：这篇文章更新了信息检索的一些基础理论，克服了单纯依靠关键词的搜索方法，并**且在信息检索领域正式引入了语义的概念**。通过定义逻辑不确定性原则（Logical Uncertainty Principle）提出了一个全新的形式化的方法框架，为基于逻辑推理的信息检索模型的发展奠定了基础（译注：逻辑不确定性原则，作者给出的定义是，给定两个句子 x 和 y，在一个特定的数据集上，y->x 这一事实的不确定性可以用我们需要多少额外的信息来确认 y->x 来度量；作者这里并没有精确的给出逻辑符号"->"的定义，可以简单理解为两段文本之间的相关性）。

贡献：本文（以及发表在 Computer Journal [24]上的拓展版）提出了一个非常清晰的理论，最基本的检索操作具有的相应的逻辑含义。基于这一观察，作者说明了逻辑是在布尔检索（Boolean retrieval，即基于词项命中的匹配方法）以及协调检索（co-ordination retrieval，即采用条件概率来描述相关性的检索方法）的基础。这篇论文还探索了适合检索这一任务的逻辑解释，表明信息检索中固有的不确定性需要一种非经典逻辑来解释。本文的讨论围绕逻辑不确定原则的介绍展开，重点阐述了相关性（Relevance）这一概念本身的不确定性。基于逻辑不确定性原则，作者引入了最少修改（minimal revision）这个概念来评价信息检索问题中的逻辑含义。本文的贡献不仅限于文中提到的这些结论，**更在于它引领了信息检索领域之后长达 15 年的基于逻辑视角的研究浪潮** [8, 9]。

历久弥珍的价值：这篇文章是一篇大师级的杰作，它修正了一些信息检索中基本的方法和理论，引入了信息检索研究的一种新范式。直至今天，这篇论文仍然意义重大。它提醒我们对基本理论保持批判的态度，不断地追问自己是否有需要去**修正基本理论、超越基本理论**。如果我们认为信息检索领域仍缺少“完美模型”来解释用户行为和检索系统的表现，那就更加应该保持这样批判性思考的态度。Norbert Fuhr 也在另外一篇工作中表达过类似的观点 [10]。不止如此，由于逻辑表示方法在数据库相关研究中也处于重要的地位，本文也因此拓展了信息检索与数据库研究的关联关系，并使我们得以从一个更加统一的视角来审视信息获取问题。这种统一考虑结构化与非结构化数据处理的思路直至今天仍有重要的参考价值。与此同时，对于语义的关注也是当前信息检索研究的热点领域之一（如实体搜索、实体链接等），本文提出的不少概念对于这一领域的研究也可能发挥更重要的价值。

- Joel Fagan. 1987. Automatic phrase indexing for document retrieval. In Proceedings of the 10th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR ’87). ACM, New York, NY, USA, 91-101. DOI: https://doi.org/10.1145/42005.42016

评论人： Donna Harman（美国国家标准与技术研究院）

概要：这篇文章在五个不同的数据集上（CRAN, CACM, INSPEC, MED, and CISI），使用了多种不同的方式定义文档检索中“非句法短语”概念（译注：主要是通过调节词项个数、文档频率等参数来构建不同的短语抽取方法）。文中非常细致地分析了为什么基于非句法短语的检索方法相对于 TF*IDF 方法在大部分情况下无法显著改进检索效果的原因。
（译注：更加早期的信息检索中通常认为文档和查询可以简单地组织为一个词的集合，这带来的问题是有一些词过于宽泛，有一些词过于精确，事实上都不适合作为索引词。因此后来出现了基于**短语**的检索方法，一些较为简单的方法主要基于词与词的共现频率等因素，另外一类较为复杂的方法主要是基于句法分析。）

贡献：这篇论文在 5 个不同的数据集上，针对文档检索这一问题，对非句法短语的定义、应用做了大量细致的分析。这些分析比较了一些相关的工作，也进行了一系列的失败样例分析。实验的结论表明：用短语来索引文档的确存在一定的困难，这些困难直到今天仍然存在。在这篇文章的影响下，更多的研究者将研究重心投向了除短语索引之外的其他更加有效的搜索方法。

历久弥珍的价值：这篇论文除了对短语检索进行了深入讨论之外，还为我们在**设计和完成复杂实验**方面树立了一个很好的典型。本文向我们示范了如何分析结果，并且将实验分析以一种容易理解的方式展示出来。本文中使用的数据集都较小，因而可以细致分析其搜索过程，理解检索中面临的真实问题。这篇文章的一个拓展版在 1989 年 3 月的美国信息科学学会会刊（Journal of Association for Information Science）发表。

- Karen Sparck Jones. 1988. A look back and a look forward. In Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '88). ACM, New York, NY, USA, 13-29. DOI: https://doi.org/10.1145/62437.62438

评论人： Justin Zobel（墨尔本大学教授，计算与信息系统学院院长）

背景： 这篇文章完成于信息检索发展了 30 年左右的时间节点，彼时正是计算机磁盘成本快速下降，互联网井喷式发展和高性能个人电脑即将出现的前夜。这些因素将会在之后为信息检索这一领域带来翻天覆地的变化。

贡献：本文是对此前（译注：1988 年）25 年信息检索研究的一个深入的回顾，并对之后一段时间的信息检索研究趋势进行了展望 。这篇论文不太容易用简短的文字概括，它的话题涵盖了信息检索研究的方方面面，对如何开展相关研究提出了建议，也对信息检索学科最初十年的研究工作进行了总结和回顾。

历久弥珍的价值：这篇文章回顾了信息检索研究早期发展的历史，记录了一些对于今天仍旧有价值的研究方法的演进过程。论文针对如何开展信息检索研究提出了诸多有价值的建议，也通过不少反例论证了自己的观点。论文所提出的很多想法具有非常重要的价值，例如，为什么以及如何对于研究工作的前提假设提出质疑。文中指出，当我们有一个研究上的新想法时，往往我们还没有完备的数据、系统、技术和完整的评价方案。**我们应该开展扎实的、易于理解的研究，而不是追求快速发表论文这样的短期成就。**作者在原文中将这样的快速论文称为“一次性的草稿”（one-off feasibility sketches）。这其中最重要的是**对于实验显著性的解释，只有显著的实验结果，才能让猜想和假设落地**。如作者所言，“在信息检索中，看起来正确的结论远远不够。”（plausible arguments are not enough in IR）。这句话到今天读来仍然振聋发聩。

- Donna Harman. 1988. Towards interactive query expansion. In Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '88). ACM, New York, NY, USA, 321-331. DOI: https://doi.org/10.1145/62437.62469

评论人：Ian Ruthven（英国思克莱德大学教授）

概要：这篇文章研究了在检索到较少数量的相关文档甚或完全无法检索到相关文档的情况下，如何通过更换新的查询词项来提升原始查询搜索性能的方法。一系列实验证明：通过将相关性反馈视为一种交互过程，有可能显著地改进搜索的性能 。

贡献：这篇论文围绕如何为搜索者提供一个可以用于查询扩展的词项列表做出了诸多贡献。文中得出了一些清晰的结论：1. 较短的扩展词列表比较长的扩展词列表对用户更有价值；2. 用户

自行选择的扩展词项要比自动扩展的词项更为有效；3. 在查询修改的过程中，通过提供扩展词项列表的方式可以使用户更有效地与系统进行交互。

历久弥珍的价值：这是一篇在信息检索实验研究中具有经典意义的优秀论文。文中提出了一个清晰的研究问题，并且围绕相关的一系列的子问题开展了系统性的研究，得到了让人信服的结果。文中的实验建立在对当时最优方法全面而客观的理解之上，同时也回应了作者自己提出的研究问题。这篇论文开创了交互式信息检索研究（译注：interactive IR）的全新领域，即以交互的方式改进用户与搜索引擎之间的信息传递。

- George W. Furnas, Scott Deerwester, Susan T. Dumais, Thomas K. Landauer, Richard A. Harshman, Lynn A. Streeter, and Karen E. Lochbaum. 1988. Information retrieval using a singular value decomposition model of latent semantic structure. In Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '88). ACM, New York, NY, USA, 465-480. DOI: https://doi.org/10.1145/62437.62487

评价人： Fernando Diaz（纽约大学副教授）

概要：为了解决词汇失配等问题，作者在“词-文档”矩阵（译注：term-document matric）上应用了奇异值分解模型，将词和文档在相同的低维空间中加以表示。论文在两个典型的任务——标准文本检索（ad hoc text retrieval ）和专家发现（expert finding）中验证了模型效果。（译注：词汇失配，即 vocabulary mismatch，指用户查询中的词项并未在相关文档中出现。早期的文本检索中，索引通常由人工构建，出现这一现象的原因可能是索引者选取的索引词与用户查询中的词并不完全重合。在当今的网络检索中，由于内容作者和搜索用户对于同一信息实体的描述可能具有较大差异，词汇失配的现象依然存在。）

贡献：隐式语义分析（译注：Latent Semantic Analysis, LSA）提供了一个有良好理论基础的方法，可以将查询词项和文档表示在同一个低维空间中。尽管相关工作中已有大量针对文档和词项的聚类研究工作，与那些基于线性代数的方法相比，这篇论文的研究思路更加贴近后续一些重要的模型，例如 PLSA（probabilistic latent semantic analysis [13]），LDA（latent Dirichlet allocation [4]）和词的分布式的表示（distributed term representations [18]）等等。

历久弥珍的价值：本文除了贡献了一套崭新的学说之外，还给出了“一篇可靠的 SIGIR 论文”的范本：**首先是小规模环境下的探索实验，接下来是进一步深入的理论分析，最后则是详尽的实验比较**。这篇论文中所提出的概念及模型对后续的信息检索、自然语言处理、机器学习等领域的研究工作产生了重要的影响，这也提示我们，值得对作者后续一个时期的工作进行回顾，以便更好地了解 LSA 方法适用的环境，它的优势与局限。

- Richard K. Belew. 1989. Adaptive information retrieval: using a connectionist representation to retrieve and learn about documents. In Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '89). ACM, New York, NY, USA, 11-20. DOI: https://doi.org/10.1145/75334.75337

评价人：Mark D. Smucker（加拿大滑铁卢大学教授）

概要：本文在信息检索领域引入了连接式表达方法（connectionist representation，译注：即神经网络）来改进检索质量。这种连接方法可以在检索的过程中将词、文章、作者等因素全部联系起来。系统接收用户查询后，将与查询相关的一系列节点依照关联扩散的方式呈现给用户。用户浏览搜索结果后，可以对这些节点（译注：这里的节点可能代表一篇文档）的重要性给出反馈。来自用户的反馈不仅可以用于重新过滤结果，也用来修正网络的连接权重。

贡献：这篇论文主要的贡献是展现了信息检索系统如何从用户交互中学习词，文档，作者之间的关系。检索系统不仅可以学到哪些文档对于给定的查询是相关的，也可以从群体用户的重复行为中学到词根（stems of words）和其他关系。和其他相关性反馈（relevance feedback）的一个重要区别是，作者设计的系统预见了当前的互联网搜索是如何利用成千上万普通用户的交互行为提升系统性能的。论文所提出的适应性检索技术（adaptive IR techniques）也已被当前的不少实际应用系统（如专利检索系统）所应用。

历久弥珍的价值：这篇论文想法新颖，富含见解深刻的讨论。文中采用的用户交互方式将反馈整合进用户搜索、浏览的过程，可以提供更加易用的搜索体验，与今天广泛采用的 ten-blue-link 大不相同。这样的用户界面也可以帮助用户更好地理解搜索的机制。此外，与通常采用的文档粒度相关性反馈不同，本文提供了一种如何在更丰富的特征层面提供反馈的策略。

- Annelise Mark Pejtersen. 1989. A library system for information retrieval based on a cognitive task analysis and supported by an icon-based interface. In Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '89). ACM, New York, NY, USA, 40-47. DOI: https://doi.org/10.1145/75334.75340

评论人： Ian Ruthven（英国思克莱德大学教授）

概要：这篇论文关注人们在进行信息交互时的一些具体的行为策略，以及这些策略的使用方法和原因，基于这些策略，作者进一步地设计了用户界面，帮助用户以更加自然的方式获取信息。通过一系列具体的任务分析，本文提出了一个用于访问图书馆中小说的“隐喻式界面”（译注，隐喻式界面，Metaphorical interface，本文中的一个例子是用户界面中有一栋房子，象征图书馆，用户点击房子的门即可进入图书馆访问信息）。作者在公共图书馆系统中对本文提出的用户界面进行了长达六个月的评价。

贡献：这篇论文以今天的眼光来看仍然是十分独特的。它从一些真实信息获取场景中的认知研究出发，总结出了**在一个新颖的交互系统中需要遵从的设计准则**。作者提出的这个名为“书房”的系统，在设计和交互方式方面与此前的系统截然不同，并且进行了长达六个月的评价与验证。

历久弥珍的价值：这篇论文有三个有价值的贡献。首先，文章呈现了一种创新的研究思路：**通过研究真实搜索环境下的认知需求来引导用户交互设计**，帮助用户更加自然地进行交互。其次，本文证明通过这样“隐喻式“的设计，搜索系统可以变得与经典的系统明显不同，可以帮助用户以更加熟悉、更加符合认知规律的方式获取信息。再次，本研究是少数几个在真实环境下经过长时间评价和验证的系统之一，而作者这篇文章最初的想法也正是起源于这一环境。

- Howard Turtle and W. Bruce Croft. 1990. Inference networks for document retrieval. In Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR ’90). ACM, New York, NY, USA, 1-24. DOI: https://doi.org/10.1145/96749.98006

评价人： Jamie Callan（美国卡耐基梅隆大学教授）

概要： 这篇论文提出了用于文档检索的推理网络（Inference Network）方法，主要侧重于使用文档不同方面的表示，例如：文档内容，文档属性，文档之间的相关性，查询变体对应多种信息需求的表达，以及包含高级操作符的结构化查询等。此前的一些基于概率的检索模型事实上都可以统一到这个框架中来。

贡献：这篇论文将此前信息检索领域已经探索了一段时间的一些想法归纳成了一个独立完整的概率框架。这个框架为今天广泛使用的两个开源搜索框架，Inquery 和 Indri，提供了理论基础。这两个搜索框架的特点是他们可以很好地融合多种信息，例如文档内容、复杂文档结构、复杂查询结构等。这一概率框架还为全球最大的商用自然语言搜索引擎之一，即 West Publishing WIN 提供支持。即使在今天，也鲜有检索模型能像推理网络一样具有广泛的建模能力。

历久弥珍的价值：很少有关于检索模型的论文可以**对许多的研究领域产生广泛的影响**，像本文这样深入细致的研究就更少了。这篇论文涉及的大多数研究问题和研究内容到今天仍然具有意义。

- Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey. 1992. Scatter/Gather: a cluster-based approach to browsing large document collections. In Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '92). ACM, New York, NY, USA, 318-329. DOI: https://doi.org/10.1145/133160.133214

评论人：Susan Dumais （ACM Fellow, ACM SIGIR Salton 奖得主 ）

概要：这篇论文讲述了一种通过文档聚类来协助浏览大型文档数据集的技术，被命名为 Scatter/Gather。这项技术的基本原理如下：首先，整个文档集合会被“打散”（scatter）到不同类簇（cluster）中；在检索的过程中，包含相关信息的类簇会被“聚集”起来（gather），然后再次被打散到不同的类簇中，以此往复。为了支持这样不断迭代的操作， 作者提出了一个快速聚类的算法，和概括每个类簇内容的技术。

贡献：在信息检索中，聚类方法较早地用于改进文档检索的效率和效果。本文提出的“Scatter/Gather”这一框架将文档聚类用于支持大型数据集的交互式浏览 。为了实现实时的交互，他们提出了一种线性时间的聚类算法，Buckshot。具体地，Buckshot 首先将所有文档的一个随机子集进行聚类，并把这个结果作为初始的聚类中心。对于初始得到的这些类簇，他们采取了一个名为分馏法（Fractionization）的方法来确定类簇的中心，这一方法从时间上来说相对较慢，但是更加准确。

历久弥珍的价值：Scatter/Gather 的一个**重要的研究动机是提供从“浏览”到“搜索”的一致体验**，这到今天仍然是一个有趣且尚未解决的问题。这篇 1992 年的论文侧重于浏览场景，而随后发表在 SIGIR 1996 的一篇拓展研究关注如何在 Scatter/Gather 的过程中利用更精确的搜索来组织文档子集。

- David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers*. In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '94). Springer Verlag New York, Inc., New York, NY, USA, 3-12. DOI: https://doi.org/10.1007/ 978-1-4471-2099-5 1 ACM: http://dl.acm.org/citation.cfm?id=188495

评论人：Fernando Diaz （纽约大学副教授）

概要：作者在基于主动学习（Active Learning）的文本分类中引入了一种不确定采样方法（Uncertainty sampling）。这种技术可以应用于任何输出为“类别+对应概率”的分类模型中。对于迭代递增（incrementally）地接受相关性反馈的模型（译注：即针对一个训练序列，

逐个读取训练样本进行训练，逐步优化的模型），不确定采样方法可以提升模型的性能。实验证明不确定性采样方法即便使用更少的数据，性能也可以超过已有的基线方法（随机采样和基于相关性的采样）。在阅读这篇文章时，应该参考作者在随后一年的 SIGIR Forum 上发表的一个勘误材料，这份材料描述了实验中的一个错误，纠正了一些实验结果。

贡献：在这个工作之前，对于类别分布不均匀（译注：skewed class distribution）的文本分类问题，通常采用随机采样、基于相关性采样或基于启发式规则采样来收集训练数据。Lewis 和 Gale 提出了一种优雅的、直观的、简单的主动学习方法。尽管后来有很多本研究的后续工作，特别是在机器学习领域，这篇文章中提出的不确定性采样始终是一个需要用来比较的基线方法。此外，在我读过的论文中，**这篇文章是为数不多的“作者发现问题并发表了勘误材料”的论文之一**。

历久弥珍的价值：不确定性采样仍是主动学习里一个重要且基本的概念，或许它不是最有效的方法，但是它简明扼要地表达了主动学习的思想 。更重要的是，作者发现问题之后公开发表了勘误材料，**这样诚实的态度应该得到整个学术圈的赞同和鼓励**。有很多已发表的论文似乎都需要类似的附属说明或勘误。

- Stephen E. Robertson and Stephen Walker. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval*. In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR ’94). Springer-Verlag New York, Inc., New York, NY, USA, 232-241. DOI: https://doi.org/10.1007/978-1-4471-2099-5 24 ACM:http://dl.acm.org/citation.cfm?id=188561

评论人：Charles Clarke（加拿大滑铁卢大学教授）

概要：作为一篇经典的概率信息检索方法论文，这篇论文首先着重于对检索中权重函数理论进行分析，然后将这些方法应用到 BM 系（Best Match family）排序函数中。第一个思路是关于词的“相关性饱和度”，也就是说**词频的影响应该趋于一个渐进线最大化**（译注：即某一词项对于相关性的影响应该存在边际效益，不会因为词频的增长无限放大）。第二个思路是说对于文档长度归一化的方法实质上体现出在**“搜索视野”和“内容冗余”之间的权衡**（译注：如对长文档的惩罚小，则搜索视野大，很容易带来内容冗繁的问题；反之搜索视野较小，可以在一定程度上避免冗繁文档排序靠前的问题）。本文提出的排序函数在 TREC-2 数据集中进行了测试，结果表明这个方法相比于之前的概率排序模型有很大的改进。

结论：在这篇论文发表后，BM 系的排序函数获得了快速的发展，例如 BM25（1994 年 11 月发表于 TREC-3），以及随后的 BM25F（2004 年发表于 CIKM）。这些排序函数已经在多个数据集和论文中无数次地证明了其有效性。Lucene 以及大多数其他的开源搜索引擎也以 BM25 为关

键的排序函数。在大多数基于机器学习的排序系统中，无论在学术界还是工业界，BM25 通常都是极为重要的一类特征。

历久弥珍的价值：抛开 BM25 在实践中重要影响不谈，单纯地阅读这篇论文就是一种享受。作者通过对于“二维泊松模型”（译注：该模型假设每个查询词的权重受到两个分布的影响，即与其最相关的一部分文档集合中的频度分布，以及其他文档组成的文档集合中的频度分布）一步步改进，虽然没有获得直接的成功（译注：我们回顾的第一篇“历久弥珍”的论文，就是关于二维泊松模型），但却为词饱和模型（Term saturation model）的发明一点一点夯实了基础。这一过程证明了理论与实际应用价值之间存在着密不可分的联系。任何研究搜索的人，不论来自于学术界或工业界，都会经常遇到 BM25。如果不充分理解 BM25 背后的理论依据，BM25 中的一些细节看上去是武断的，或者说是一个经验性的结果。在读完这篇论文后，你就会更好地理解 BM25 的优雅和简洁。

- James P. Callan, Zhihong Lu, and W. Bruce Croft. 1995. Searching distributed collections with inference networks. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR ’95). ACM, New York, NY, USA, 21-28. DOI: https://doi.org/10.1145/215206.215328

评论人：James Allan （美国麻省大学教授）

概要：这篇论文提出了一个针对分布式信息检索问题的模型。作者做了大量的实验来证明模型的有效性。该模型的基础是信息检索中一个形式化的推理网络（Inference network）方法（该方法应用于 Inquery 检索系统中）。实验用到的数据多达 3Gb、100 多万篇文档。通过文档来源进行划分，实验数据被分成了 17 个子数据集。

贡献：这篇论文为数据表示、数据排序、数据选择、排序列表合并等问题提出了新的方法，评价结果坚实可靠。论文表明一个文档的相关性信任函数可以用来计算一个数据集里是否包含相关文档的置信程度。论文证明了在所有数据集上对文档的相关性分数做归一化（译注：即把不同的数据集上的文档相关性分数做一个全局的归一化）其实没有必要，利用包含某一文档的数据集的可信程度（译注：即该数据集包含相关文档的可能性） 对文档分数加权可以达到近似的效果，并且更加高效。作者还发现，当只用到一部分数据集时，搜索精度（Precision）只会很轻微地受到影响。最后，一系列的实验得出了几个重要的结论：第一，如果减少用于表示一个数据集合的词项数量，检索性能只会受到很小的影响（除了一些极端的参数设置外）；第二，减少每一个子集中的候选结果数量对于召回率影响很小，但是候选的检索文档精度降低了 50%。

历久弥珍的价值：尽管互联网的搜索引擎已经说明了，在数据资源十分充足的情况下，构建一个大型搜索索引是可行而且有效的，但相关信息不易聚集的情况仍然非常普遍。这篇论文针对这个相对宽泛的的问题提出了一个思路，并阐述了如何从不同的角度评价这个问题。另外，这

篇论文也提供了一个很好的示范：如何形式化地去定义和解决一个解决方案（文档检索），并将这个方案应用于其他问题（数据检索）。最后，文章展示了运用大量实验去验证一个模型、理解不同参数设置的影响的重要性。

- Jinxi Xu and W. Bruce Croft. 1996. Query expansion using local and global document analysis. In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '96). ACM, New York, NY, USA, 4-11. DOI: https://doi.org/10.1145/243199.243202

评论人: Donna Harman（美国国家标准与技术研究院）
概要：这篇论文在几个不同的数据集上深入考察了三种不同的查询扩展方法。第一种**是全局文档分析**，利用一个名词的集合和名词周围固定窗口大小的上下文，在整个数据集上生成一个概念库（一种自动化词库）。原始的查询会在这个概念库上进行匹配，然后选择合适的概念进行查询拓展。第二种方法是基于**局部反馈**（也就是常说的伪相关性反馈，pseudo relevance feedback），通过排序靠前的文档集合来生成查询拓展。第三种方法，**局部上下文分析**，是以上两种办法的一种结合，利用排序靠前的文档集合去库中匹配新的概念进行查询扩展。实验一共用到了 TREC-3，TREC-4，和 WEST 三个不同的数据集。

贡献：这篇论文除了引入了新的方法（方法三，局部上下文分析，local context analysis）外，还对这三种方法可能面对的若干问题做了详细描述。对于 TREC-3 以及 TREC-4 数据集，局部上下文分析的表现较好（23%的性能提升），但在 WEST 数据集上表现一般。全局文档分析方法（方法一）表现较差，基于局部反馈的方法（方法二）在 TREC 数据上表现不错，但对 WEST 数据表现一般。文章对不同数据集的差异进行了分析，同时发现局部上下文分析在相关文档较少时表现更加鲁棒。

历久弥珍的价值：这篇论文研究了应该选择哪种类型的概念对查询进行词项扩展、或词项权重的重新估计。这项工作完成于语言模型（language model）流行之前。在查询扩展中，哪些概念是有用的？不同的方法如何去挖掘这些概念？这些问题非常有趣。尽管两种局部分析方法（译注：一种直接基于排序靠前的文档，另外一种基于排序靠前的文档去概念库中匹配）表现都很不错，但这两种方法发现的概念是不同的。在信息检索中，这篇文章是对一个研究领域采用不同的数据集、进行系统研究的一个好的案例，包括如何去分析结果，让读者更好地理解问题的方方面面。

- Amit Singhal, Chris Buckley, and Mandar Mitra. 1996. Pivoted document length normalization. In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '96). ACM, New York, NY, USA, 21-29. DOI: https://doi.org/10.1145/243199.243206

评论人： Mark D. Smucker（加拿大滑铁卢大学教授）

概要：在检索方法设计中，处理不同文档的长度差异是一个重要的问题。在已知文档长度的条件下，中心文档归一化方法（Pivoted document normalization）可以利用文档的相关性（译注：即该文档是相关文档的概率），结合余弦相似度来选择更加合适的文档。
（译注：一个文档被检索到的概率与其归一化因子的大小是负相关的，该方法基于初始得到的检索结果，对归一化因子进行调整，使得更有可能是相关文档的结果因子降低，被检索到的概率增加。）

贡献：这篇文章一个重要的贡献是清楚地表明了**文档的性质会影响它们被检索到的概率**，这一影响与查询无关。这里，作者关注的文档特征是文档长度，类似的思路也可以用于其他的文档属性，例如考虑垃圾网页的特征对其进行过滤等。

历久弥珍的价值：这篇论文清楚地表明检索方法中存在一些固有的偏置（bias）。例如，相对于文档固有的相关概率，余弦相似度更有可能偏向选择短的文本，这个偏置可以利用中心文档归一化方法进行校正。作者发现了这一现象并且校正了文档长度带来的偏置后，再去回顾其他的检索方法，发现这些检索方法之所以没有关注这一现象，是因为它们的设计或者间接地解决了这一问题，或者根本无需考虑这一问题（例如面对文档长度差异很小的数据集合）。

- James Allan, Ron Papka, and Victor Lavrenko. 1998. On-line new event detection and tracking. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '98). ACM, New York, NY, USA, 37-45. DOI: https://doi.org/10.1145/290941.290954

评价人： Fernando Diaz（纽约大学副教授）

概要：本文提出了监测文字新闻信息流的方法。在传统的基于过滤的方法之上，作者定义了**发现**（即事件检测）和**组织**（即事件跟踪）这两个新的任务。这项研究还对当时比较前沿的话题发现和话题跟踪研究成果进行了总结。

贡献：在这篇论文之前，大多数关于文本流的研究主要建立在 Luhn 的信息过滤模型 [15]之上。这些研究主要采用与文本检索（Ad-hoc retrieval）非常相似的技术手段 [3]。这篇文章首先定义了主题检测与追踪这一任务，还指出这一针对信息流的研究任务区别于普通文本检索的诸多独特的方面。现在，许多用于流文本分析的评价方法（例如，基于时间的摘要，temporal summarization，实时概要 realtime summarization）都是从这一工作中得到启发。本文中提出的一些概念在以上许多的后续工作中被广泛采用。

历久弥珍的价值：越来越多的流数据（例如社交媒体数据）使得人们对话题发现与探索再度产生了兴趣。人们对这一任务的理解比文档检索（Ad-hoc retrieval）要浅得多。在如何构建和评价这样的信息流分析系统方面，本文是一个创新性很强的工作。

- Krishna Bharat and Monika R. Henzinger. 1998. Improved algorithms for topic distillation in a hyperlinked environment. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '98). ACM, New York, NY, USA, 104-111. DOI: https://doi.org/10.1145/ 290941.290972

评价人： Charles Clarke（加拿大滑铁卢大学教授）

概要：这篇工作发表时，利用 Web 链接结构分析来进行排序的研究才刚刚兴起，本文从三个重要方面拓展了以（Kleinberg，1998）为代表的早期工作：1）减少了单个节点的影响；2）通过文档-查询的相似性对节点进行权重估计；3）图剪枝。本文在真实互联网数据集上进行了实验，取得的检索效果在精度意义（译注：Precision@10）上取得了突破性的改进。
（译注：在当时的网络资源检索中，网络链接结构被认为是非常有效的信息，可以找到高质量的网络资源，本文的主要贡献是在单纯的网络结构分析中加入了内容分析，从而在精度上获得了大幅提升。）

贡献：虽然这篇文章的一些具体的方法已经被拓展延伸，或者被新的方法取代。这项研究对于后续工作的影响和启发很大，在谷歌学术上取得了超过 1000 次的引用。

历久弥珍的价值：在信息检索的导论课中，链接分析通常会介绍原始的 PageRank 算法。谷歌通过页面排序取得了巨大的成功，并且以 PageRank 算法作为搜索引擎的基石。学生们通常会忽视在同一时间出现的大量相关工作。对链接分析更深层次的理解需要回顾经典的一些论文，比如这篇工作，它们有可能引领我们进入那些还未被深入探索的研究方向。

- Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '98). ACM, New York, NY, USA, 335-336. DOI: https://doi.org/10.1145/290941.291025

评价人：Jamie Callan（卡耐基梅隆大学教授）

概要：这篇论文定义了一个新的概念，“最大边际相关性”（Maximum Marginal Relevance, MMR），可以通过文档重排序，在不牺牲文档的相关性的前提下减少内容冗余带来的负面影响。文章阐述了如何用 MMR 改善搜索引擎的排序，以及改进单个文档与多篇文档列表的摘要抽取。这篇两页的海报论文（poster）**说明了短论文在这个领域也能产生巨大的影响**。

贡献：据说 MMR 的灵感来自于一个让人沮丧的现象，即当时商业搜索引擎返回很长的文档列表，但是不同结果的内容几乎相同 。MMR 是一个简单的改进搜索结果质量以及文档概要提取的技术，同时启发了后续大量关于结果多样化的研究。许多年来，这一直是相关领域研究者选择的一个基线模型。

历久弥珍的价值：这篇论文短而简洁，开门见山。两个看似十分不同的问题（排序与摘要抽取）被证明是都是排序的问题，都可以从基于多样化的文档重排序中受益。

- Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '98). ACM, New York, NY, USA, 275-281. DOI: https://doi.org/10.1145/290941.291008

评价人：Djoerd Hiemstra（译注：荷兰特温特大学教授）

概要：受到语音识别中采用的模型启发，Ponte 与 Croft 提出了一个文档的语言生成模型，并且根据每个模型估计了给定文档，生成相应查询的概率。这个概率可以被用来进行搜索结果的排序。这篇论文在两个 TREC 测试集上进行实验，说明了语言模型的方法可以远远地好于标准的 TF-IDF 方法。

贡献：语言模型对信息检索的贡献不能过于夸大。Ponte 与 Croft 在 1998 年的论文通过定义查询的生成模型（译注：即根据文档生成查询的概率，可以用于估计相关性），简单的词频（译注：Term Frequency）估计器以及概率平滑的必要性，在信息检索中形式化地建立了语言模型。同年，两个研究小组（BBN 与 Twente/TNO，他们各自独立研究出了相似的模型）在

TREC 上用实验证明了语言模型的有效性。到 2001 年，SIGIR 组织了两次关于语言模型的研讨，2003 年**信息检索领域的主要研究路线被称之为“信息检索与语言模型中的挑战”** [2]。语言模型对一些实际的应用具有杰出的贡献，其中包括 Lemur，Lucene 以及 Terrier 等。

历久弥珍的价值：这篇论文通过介绍语言模型，开拓了一种信息检索研究的新思路。语言模型是**基本不需要假设**的概率索引方法。例如，他们不需要参数分布，不用假设文档在某个预先定义的类别里，甚至也不对文档相关性做显式的建模。这与传统的概率索引模型为代表的研究（例如二维泊松模型，BM25 公式等）产生了明显的区别，开辟了信息检索研究的新方向。

- Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR ’99). ACM, New York, NY, USA, 50-57. DOI: https://doi.org/10.1145/312624.312649

评价人： Susan Dumais（ACM Fellow, ACM SIGIR Salton 奖得主）

概要：这篇论文提出了概率隐式语义索引（Probabilistic Latent Semantic indexing (PLSI)）。这是隐式语义索引（Latent Semantic Indexing，LSI）的一个变体。PLSI 是基于数据的共现信息，构建一个隐式的类别（译注：这个类别是统计意义上的，并不是通过显示的语法规则可以明确定义的）。当在信息检索中将 PLSI 运用于“词-文档”矩阵时，PLSI 相对于向量空间的词匹配以及 LSI 模型，在四个数据集（MED, CRAN, CACM, CIS）上都可以有效改进检索精度。

贡献：与 LSI 方法相比，PLSI 提供了低维度的、具有概率意义的一种形式化表示方法，同时也是一个文档生成模型。这篇论文也对最大似然估计的泛化做出了贡献，为混合模型提出了一种缓和 EM 算法（tempered EM）。实验结果表明，PLSI 方法相对于 LSI 和词匹配有显著的效果改进。

历久弥珍的价值：本文对于采用概率模型进行降维的优点进行了概括。PLSI 已经广泛地用于文档检索，其中包括应用于语言模型和协同过滤。另外，PLSI 与非负矩阵分解也有着紧密的联系，是启发 LDA（Latent Dirichlet Allocation）的研究之一。

- Adam Berger and John Lafferty. 1999. Information retrieval as statistical translation. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR ’99). ACM, New York, NY, USA, 222-229. DOI: https://doi.org/10.1145/312624.312681

评价人： James Allan（美国麻省大学教授）

概要：本文基于统计机器翻译为信息检索提出了一个新的思路，其中查询被看作是一个想象中的相关文档的“翻译”。这篇论文基于 IBMs 模型 1（Brown et al, Computational Linguistics 1990，译注：一种机器翻译方法）构建了检索模型。IBMs 模型 1 是利用单个词（unigram）的统计翻译模型。作者通过建立“查询-文档”对来决定翻译正确的概率。这个方法的实验结果与 TF-IDF 相比，可以带来稳定的性能改进。

贡献：这篇论文成功地将单语言的信息检索问题类比为一个统计翻译过程。文中证明这样的方法是有效的。作者提出了一种**直接合成训练数据**的方法，不需要另外进行相关性的标注，在本文的训练中，这样直接构造的训练数据是有效的。基于翻译的模型受到 Ponte 与 Croft （SIGIR 1998）的语言模型的启发，从思路上与语言模型（language model）是相似的。另外，翻译模型自然地考虑同义词和多义词的情况，从而更好地弥合了所谓的词汇鸿沟（vocabulary gap）。在这个意义上，它也是在语言建模框架下进行查询扩展的一个很好的实例。

历久弥珍的价值：本文是连接计算机科学中两个领域的一个很好的例子。作者从统计机器翻译模型中推导出一个形式化的检索模型，进而实现模型、构建所需的训练数据并对检索效果进行评价。事实上在单一语言中使用翻译模型的想法给人的第一印象是违反直觉的，但是仔细想想这个模型事实上形式简洁，非常容易理解。这使得它能够很顺利的应用于信息检索任务中。建议读者仔细思考这种方法与各种查询扩展技术之间的关系。

- Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, and John Riedl. 1999. An algorithmic framework for performing collaborative filtering. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR ’99). ACM, New York, NY, USA, 230-237. DOI: https://doi.org/10.1145/312624.312682

评价人：Paul Bennett（微软公司高级研究员）

概要：本文针对几种基于近邻（neighborhood，推荐系统中认为从属性、行为上较为相似的实体）的推荐方法进行了经验性的实证比较，并研究了不同的设计对于推荐系统实际表现的影响。在已知一些“用户-商品”打分的情况下，作者给出了一个基于近邻推荐的框架：（1）基于与当前用户的相似性，估计其他用户的行为对于当前推荐的重要程度（译注：如果用户 A 的行为与当前使用推荐系统的用户行为越相似，他们越有可能兴趣相近，A 的行为数据对于当前的推荐过程越重要）；（2）针对当前的用户或商品，选定一个用户子集用于推荐（即邻居集合）；（3）使用相似用户评分的加权组合进行推荐（可能需归一化）。作者随后研究了相似性度量的选择、基于差异的权重估计，显著性的权重估计和多种组合其他用户评分的方法。对于相似性度量函数，作者发现 Pearson 系数和 Spearman 系数的性能相当，均优于其他选项。作者指出，许多相似性度量方法没有考虑用户参与相关性计算的数据量，作者提出采用数据量来估计相关性的显著性权重（译注：以两个用户的相关性而言，如果这两个用户共同点评过的商品越多，则说明这个相关关系越可靠）。虽然这里所说的显著性权重估计方法相对简单（通常对

于共同评分的商品数量低于 50 时，作者认为显著性较低），作者提出的“显著性加权”这个概念在基于邻居的推荐方法中却很重要。在相似性的计算中，作者还研究了“差异加权”，那些评分方差较大的商品权重也相应地较大（译注：这个方法也是较为直观的，因为评分的分化越明显，越说明这个商品可能暗含了用户在某个方面的立场）。例如，几乎所有人都喜欢《泰坦尼克号》(Titanic)，而对《西雅图夜未眠》（Sleepless in Seattle）的评分却有很大差异，因为它们将那些喜欢动作电影的人和喜欢浪漫电影的人区分开来。作者调查方差加权的方法没有实际作用，但是调研那些关键商品的努力仍是有效的 [23]，因为它们可用于快速地对人群进行分类，或在冷启动情景下快速对用户进行建模。最后，当进行预测时，用户的评分分布可能存在固有的差异。因此，当最终生成预测时，采用 z-score 更为准确。如果一些用户是“倾向于给低分”的用户，采用 z-score 可以使得他们的评分与那些“倾向于给高分”的用户更具可比性 。

贡献：本文是最早的基于大规模实际历史数据进行推荐系统的实证分析工作之一。文中采用了电影预测网站的数据（1K 用户，100K 评分）。 结果表明对于显著性的权重估计比估计相似度的函数选择更加重要。尽管本文发现两种对用户评分进行归一化的方法 （平均数+差异和 z-score）表现相当，但本文首先指出在构建推荐系统时需要处理不同用户评分的差异，这是一个很大的贡献。

历久弥珍的价值：这是一篇关于基于邻域协同过滤推荐的创新度很高的论文。 Ning 等人[19]最近对基于近邻的推荐中的若干方法进行了调研。他们认为，尽管现在一些基于模型的方法已经在预测精度上做得比基于近邻的方法更好，基于近邻的方法仍有重要意义，因为它们可以更好地提供令人惊喜的推荐，并且可以综合利用局部相似性。基于近邻的方法不仅计算简单、直观，且性能稳定。特别是在一些基于交互信息或基于会话的推荐系统中，效率格外重要。在现代基于近邻的推荐方法中，相似性的概念可以是隐式的，例如在图方法中，就有基于用户-商品的二部图进行随机游走的方法。作者在本文中证明了显著性加权的重要性，这对于基于近邻的推荐方法[19]以及整个信息检索领域来说是十分重要的。此外，这是第一项通过 z-score 来归一化个体评分的研究工作。虽然本文中 z-score 相对于“平均数-差值”方法没有显著的优势，在后续工作中发现其实二者之间还是存在一定的差异 [12]。有趣的是，在文本的未来工作展望中，作者还提到，基于奇异值分解（SVD）的推荐方法可以启发基于矩阵分解的推荐方法。最后，本文是最早考虑推荐系统覆盖率的工作之一，后来这被重新定义为推荐系统多样性的概念，即衡量在所有的商品中，展现给用户的商品的比例 [1]。

- Chris Buckley and Ellen M. Voorhees. 2000. Evaluating evaluation measure stability. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR ’00). ACM, New York, NY, USA, 33-40. DOI: https://doi.org/10.1145/345508.345543

评价人：Justin Zobel（墨尔本大学教授 ）

背景：1992 年，TREC 一开始就使用了比当时研究者用的最大数据还要大两个量级的数据集合，这对于当时的信息检索研究产生了立竿见影的影响。但在某些方面，整个领域的研究人员还没有为这次飞跃做好准备。许多实验室面临技术上的困难，而且许多关于实验设计的知识和理论都是基于非常小的数据集得出的。开发 TREC 平台的团队根据提供的当时最前沿的信息做出了一系列决策，其中有几项至今仍然有影响。但是在 TREC 最初的几年中，有些决策只能依赖人们的直觉，因为当时的 TREC 组织者也没有足够的依据说明那些决策是不是最优的，或者是在当时条件下最好的选择。这篇文章发表于 2000 年，即使 TREC 开展了 8 年后，这篇文章中提到的一些处理方法仍然可以认为是当时的“拇指法则”（译注：即一种经验性的、可以广泛应用于许多情况，但又不够准确的法则）。

概要和贡献：在这篇论文中，Buckley 与 Voorhees 表明，TREC 测试集（不同检索算法的检索结果、以及对查询文档对的相关性标注）可用于对评价指标的可靠性进行十分敏感的判断。这篇文章中一些具体的结论已经（合情合理地）被更新的工作所取代，但是本文帮助建立了一套如何进行系统性能评价的方法，使得“性能评价”成为一个独立的研究方向。

历久弥珍的价值：像许多经典的文献一样，本文讨论的问题建立在坚实的观察和分析之上。作者并没有简单地解决一个现有的问题，而是试图了解现有方法中可能存在的缺陷，并且提出一套方法来检查这些缺陷是否确实存在。

- Kalervo J¨arvelin and Jaana Kek¨al¨ainen. 2000. IR evaluation methods for retrieving highly relevant documents. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR ’00). ACM, New York, NY, USA, 41-48. DOI: https://doi.org/10.1145/345508.345545

评论人：Tetsuya Sakai（日本早稻田大学教授）

概要：作者提出了考虑多级相关性的信息检索评估方法。他们的第一个提议是针对每个相关性等级绘制精度-召回率（precision-recall）曲线。第二个提议是计算累积增益（cumulative gain，CG）和考虑衰减的累积增益（discounted cumulative gain，DCG），并可视化地呈现了搜索系统的（D）CG 曲线与理想排序列表的曲线之间的差异。本文还针对一些结构化的查询（Structural queries）进行了案例研究（Case study）。

贡献：这篇 2000 年的 SIGIR 论文提出了累积收益（译注：Cumulated Gain，CG）和考虑衰减的累积收益（Discounted Cumulated Gain，DCG）；在 2002 年，他们的 ACM TOIS 论文上提出了归一化的累积收益（nCG）和归一化的衰减累积收益（nDCG）。虽然 nCG 与 Pollock 在 1968 年提出的标准化滑动率基本相同（即理想的排名列表的概念，Pollock 称之为主列表，他将其定义为“知识库中的所有文档”，提出采用“基于主值进行降序排序”），nDCG 的新思想带来了巨大的转变。IR 研究人员开始基于多级相关性标注来评估和优化系统，取代了传统的的两级相关性度量方法，如召回率和精确度。事实上，SIGIR 2000 年的论文中的 DCG 提议是非常及时的：根据 Hawking 和 Craswell（“TREC Book”，第 5 章，2005 年第 205 页）等人的记

录，在 2000 年 4 月举行的 Infortortics 搜索引擎会议上，“TRECers ”和一群搜索引擎代表聚集在一起，后者强调了多层次的相关性评估的重要性。几年后，nDCG [7]成为网页搜索评估以及许多其他 IR 评估任务中的不可或缺的评价指标。

历久弥珍的价值：虽然信息检索评估中倾向于“忘记用户”（或至少是尝试消除用户因素），并专注于一个数字形式的指标，如 nDCG，但本文强调了站在用户的视角理解性能评估的重要性。本文中的图 3 非常经典，值得我们不断回顾，我们看到 DCG 远比“精度-召回”图更加直观（译注：精度召回图即顺序地考察搜索结果列表，在每一个位置计算精度和召回，在二维平面中绘制散点并连接成线）。在后者中，精度的位置是依赖于对应的召回的位置的，而召回没有什么直接的意义。最后，文中有这样一个完美的说法：“这是一个统计学上的一致且显著的差异，但用户能注意到吗？”（It is a consistent and statistically significant difference but are the users able to notice it?”）。作者明确地区分了统计学意义和实际应用的意义，并强调后者才是真正重要的。

- John Lafferty and Chengxiang Zhai. 2001. Document language models, query models, and risk minimization for information retrieval. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR ’01). ACM, New York, NY, USA, 111-119. DOI: https://doi.org/10.1145/383952.383970

评价人：Nicola Ferro（意大利帕多瓦大学副教授，欧洲信息检索评测组织 CLEF 协调人）

概要：本文对信息检索中的语言模型（language model）进行了改进，将语言模型组织为一个非常优雅的理论框架——风险最小化框架（risk minimization framework）。我们可以从风险最小化框架中推演出传统的概率模型和其他的语言模型。此外，基于这一框架，本文介绍了一种查询扩展的方法，并解决了影响此前查询扩展模型性能的训练问题。

贡献：本文对语言模型的形成作出了重大贡献。具体地，它介绍了一个基于贝叶斯决策理论的风险最小化框架，在当时已有的一些检索模型基础上提供了一个统一的理论框架（包含向量空间模型，经典概率模型，和其他语言模型）。这个新框架的优点远不止此。实际上，它首次引入了**查询语言模型**，与**文档语言模型**相对应，并提出了采用 KL 距离（Kullback-Leibler divergence）来度量查询语言模型和文档语言模型之间的距离，作为文档排序的一个依据。这个新框架可以更加自然地建模查询扩展。最后，本文利用 Markov 链估计模型参数，减少了对大量训练数据的依赖。

历久弥珍的价值：在语言模型演变的历程中，本文是一篇基石性论文。它也是一个非常有价值的例子，展现了如何以简明而可靠的方式扩展现有的框架。此外，文中的方法已经被应用于许多其他任务中，例如实体搜索，多语言信息检索，或 RDF 图上的排序。即使在今天，我们也

可以进一步探索如何利用马尔可夫链来估计模型参数，例如将其他类型的用户信息（例如用户与 IR 系统的交互）整合到语言模型中 。

- Victor Lavrenko and W. Bruce Croft. 2001. Relevance based language models. In Proceedings of the 24th annual international ACM SIGIR conference val (SIGIR '01). ACM, New York, NY, USA, 120-127. DOI: https://doi.org/10.1145/383952.383972

评价人： Fernando Diaz（纽约大学副教授）

概要：作者为语言模型（language model）提出两种伪相关反馈方法，并证明其在检索与话题跟踪任务中具有很好的效果。

贡献：本文的第一点贡献是将相关性从理论上纳入语言模型。此外，本文提出的基于相关性的方法还有很好的应用效果。文中的第一种方法，RM1，是一种对排序靠前的文档的向量进行加权组合的方法，可以很容易地用于大多数的检索系统。如果采用相关性模型（译注：即伪相关反馈的信息）更新原始的查询模型，将会极大地提升检索的性能。此外，与相对保守的查询扩展方法相比（即从几篇文档中找几个词加入原始查询），Lavrenko 和 Croft 主张大范围的查询扩展，即从许许多多的文档中拿出上千个词加入原始查询。

历久弥珍的价值：相关性模型，特别是文中的 RM3 模型，在许多检索任务中仍然保持着最好的效果。本文以及其在 A Generative Theory of Relevance [14]的拓展部分详细研究了如何进行有效的自动查询扩展。我遇到很多使用相关性模型的论文，里面的超参数选择都不太恰当，说明很多研究者对该算法并没有充分理解。

- Chengxiang Zhai and John Lafferty. 2001. A study of smoothing methods for language models applied to Ad Hoc information retrieval. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01). ACM, New York, NY, USA, 334-342. DOI: https://doi.org/10.1145/383952.384019

评阅人：Jamie Callan（美国麻省大学教授）

概要：本文对查询似然检索模型（query-likelihood retrieval model）中的平滑技术做了仔细研究。它具体研究了三种平滑方法（Jelinek-Mercer，Dirichlet，absolute discounting），分析了不同长度的查询以及不同平均文档长度的数据集带来的影响，系统地梳理了每种平滑方法的性能和贡献。

（译注： 在语言模型的估计中，数据通常非常稀疏，平滑技术可以减小数据稀疏性带来的不确定性。）

贡献：本文发表时，平滑技术在检索模型的作用还没有被充分了解。 Zhai 和 Lafferty 发现，平滑有两方面的重要作用：查询模型（query model）解释了每个查询词对查询的重要性，其作用类似于 IDF。估计（Estimation）方法改进了文档模型（document model）中的最大似然概率估计。作者发现查询建模对于长查询来说至关重要，但对于短查询作用不大。 Jelinek-Mercer 平滑方法和 Dirichlet 平滑方法分别可以改进查询模型与文档模型。文中讨论了每种平滑方法对参数值的灵敏程度，并给出了每个参数的建议取值范围。

历久弥珍的价值：这篇论文鼓励大家探索为什么平滑技术在查询似然模型中能发挥作用，以及应该如何改进平滑的方法。本文给出了一些缺省的平滑参数设置，被很多研究者广泛使用。它还提醒我们，我们大多数情况下都不经思考地直接使用缺省的参数设置，但这样往往不是最好的选择。

参考文献

[1] Gediminas Adomavicius and YoungOk Kwon. Improving aggregate recommendation diversity using ranking-based techniques. IEEE Trans. on Knowl. and Data Eng., 24(5):896–911, 2012.

[2] James Allan et al. Challenges in information retrieval and language modeling. In ACM SIGIR Forum, volume 37, pages 31–47. ACM, 2003.

[3] Nicholas J. Belkin and W. Bruce Croft. Information filtering and information retrieval: Two sides of the same coin? Commun. ACM, 35(12):29–38, 1992.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022, 2003.

[5] Abraham Bookstein and Don R. Swanson. Probabilistic models for automatic indexing. Journal of the Association for Information Science, 25(5):312–316, 1974.

[6] Chris Buckley. The SMART project at TREC, pages 301–320. MIT Press, 2005.

[7] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In Proceedings of the 22Nd International Conference on Machine Learning, ICML '05, pages 89–96, New York, NY, USA, 2005. ACM.

[8] F. Crestani, M. Lalmas, and C.J. van Rijsbergen. Information Retrieval: Uncertainty and Logics: Advanced Models for the Representation and Retrieval of Information. The Information Retrieval Series. Springer US, 2012.

[9] Fabio Crestani, Sandor Dominich, Mounia Lalmas, and Cornelis Joost van Rijsbergen. Mathematical, logical, and formal methods in information retrieval: An introduction to the special issue. J. Am. Soc. Inf. Sci. Technol., 54(4):281–284, 2003.

[10] Norbert Fuhr. Salton award lecture information retrieval as engineering science. SIGIR Forum, 46(2):19–28, 2012.

[11] Stephen P. Harter. A probabilistic approach to automatic keyword indexing. part i and ii. Journal of the Association for Information Science, 26(5):197–206, 280–289, 1975.

[12] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst., 22(1):5–53, 2004.

[13] Thomas Hofmann. Probabilistic latent semantic indexing. In Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.

[14] Victor Lavrenko. A Generative Theory of Relevance. PhD thesis, University of Massachusetts Amherst, 2004.

[15] H. P. Luhn. A business intelligence system. IBM J. Res. Dev., 2(4):314–319, 1958.

[16] Hao Ma, Irwin King, and Michael R. Lyu. Effective missing data prediction for collaborative filtering. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07, pages 39–46, New York, NY, USA, 2007. ACM.

[17] Melvin E. Maron and J. Lary Kuhns. On relevance, probabilistic indexing and information retrieval. Journal of the ACM (JACM), 7(3):216–244, 1960.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems, NIPS'13, pages 3111– 3119, USA, 2013. Curran Associates Inc.

[19] Xia Ning, Christian Desrosiers, and George Karypis. A Comprehensive Survey of Neighborhood-Based Recommendation Methods, pages 37–76. Springer US, Boston, MA, 2015.
[20] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98, pages 275–281, New York, NY, USA, 1998. ACM.
[21] Stephen E. Robertson, Melvin E. Maron, and William S. Cooper. Probability of relevance: a unification of two competing models for document retrieval. Information technology: research and development, 1(1):1–21, 1982.
[22] Stephen E. Robertson and Karen Sparck-Jones. Relevance weighting of search terms. Journal of the American Society for Information Science, 27(3):129–146, 1976.
[23] Mingxuan Sun, Fuxin Li, Joonseok Lee, Ke Zhou, Guy Lebanon, and Hongyuan Zha. Learning multiple-question decision trees for cold-start recommendation. In Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM '13, pages 445–454, New York, NY, USA, 2013. ACM.
[24] C. J. Van Rijsbergen. A non-classical logic for information retrieval. The Computer Journal, 29(6):481, 1986.