全文可在：http://dx.doi.org/10.1561/1500000035

语义匹配搜索

李杭
华为技术有限公司，香港
hangli.hl@huawei.com

许俊
华为技术有限公司，香港
nkxujun@gmail.com

now
the essence of knowledge
波士顿 — 代尔夫特

全文可在：http://dx.doi.org/10.1561/1500000035

基础和趋势® R在信息检索

出版、销售和发行： 现在出版商公司

邮政信箱 1024
马萨诸塞州汉诺威 02339
美国
电话 +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

北美以外：
现在出版商公司
邮政信箱 179
公元 2600 年代尔夫特
荷兰人
电话 +31-6-51115274

本出版物的首选引用是

H. Li 和 J. Xu。*搜索中的语义匹配*. 基础和趋势® R在信息检索，卷。7，没有。5，第 343–469 页，2013 年。

这个基础和趋势® R issue 用 L 排版A TEX 使用类文件设计
尼尔·帕里克 (Neal Parikh) 着。在无酸纸上印刷。

书号：978-1-60198-805-8

全文可在：http://dx.doi.org/10.1561/1500000035

基础和趋势® 在信息检索

2013 年第 7 卷第 5 期

编辑委员会

全文可在：http://dx.doi.org/10.1561/1500000035

编辑范围

主题

基础和趋势® in Information Retrieval 发布调查
和以下主题的教程文章：

- 红外的应用
- IR 架构
- 协同过滤和推荐系统
- 跨语言和多语言 IR
- 分布式 IR 和联合搜索
- IR 的评估问题和测试集
- IR 的形式模型和语言模型
- 移动平台上的 IR
- 结构化文档的索引和检索
- 信息分类与聚类
- 信息提取
- 信息过滤和路由
- 元搜索、排名聚合和数据融合
- IR 的自然语言处理
- IR 系统的性能问题，包括算法、数据结构、优化技术和可扩展性
- 问题解答
- 单个文档的汇总，多个文档和语料库
- 文本挖掘
- 主题检测和跟踪
- IR 中的可用性、交互性和可视化问题
- IR 的用户建模和用户研究
- 网络搜索

给图书馆员的信息

基础和趋势® in Information Retrieval，2013，第 7 卷，第 5 期。
ISSN 纸质版本 1554-0669。ISSN 在线版本 1554-0677。也可作为纸质和在线订阅相结合的形式提供。

基础和趋势® 在信息检索
卷。7, No. 5 (2013) 343–469
 DOI：
10.1561/1500000035

now
the essence of knowledge

搜索中的语义匹配

李杭
华为技术有限公司，香港
hangli.hl@huawei.com

许俊
华为技术有限公司，香港
nkxujun@gmail.com

内容

全文可在：http://dx.doi.org/10.1561/1500000035

抽象的

相关性是确保用户对搜索的满意度的最重要因素，搜索引擎的成功在很大程度上取决于其相关性性能。据观察，大多数相关的不满意案例是由于查询和文档之间的术语不匹配（例如，查询“ny times”与仅包含“New York Times”的文档匹配不佳），因为术语匹配，即，词袋方法，仍然是现代搜索引擎的主要机制。因此，可以毫不夸张地说，查询和文档之间的不匹配是搜索中最关键的挑战。理想情况下，如果查询和文档具有主题相关性，人们希望看到它们相互匹配。最近，研究人员付出了巨大的努力来解决这个问题。主要方法是进行语义匹配，即进行更多的查询和文档理解来表示它们的含义，并在丰富的查询和文档表示之间进行更好的匹配。随着大量日志数据的可用性和先进的机器学习技术，这变得更加可行，并且最近取得了重大进展。本综述系统详细地介绍了新开发的机器学习技术，用于搜索中的查询文档匹配（语义匹配），特别是网络搜索。它侧重于基本问题，以及查询文档匹配的形式方面、短语方面、词义方面、主题方面和结构方面的最新解决方案。解释的想法和解决方案可能会激励工业从业者将研究成果转化为产品。介绍的方法和进行的讨论也可能会激发学术研究人员寻找新的研究方向和方法。query 和 document 的匹配不仅限于搜索，在问答、在线广告、跨语言信息检索、机器翻译、推荐系统、链接预测、图像标注、药物设计等应用中都可以找到类似的问题，作为匹配来自两个不同空间的对象的一般任务。技术 query 和 document 的匹配不仅限于搜索，在问答、在线广告、跨语言信息检索、机器翻译、推荐系统、链接预测、图像标注、药物设计等应用中都可以找到类似的问题，作为匹配来自两个不同空间的对象的一般任务。技术 query 和 document 的匹配不仅限于搜索，在问答、在线广告、跨语言信息检索、机器翻译、推荐系统、链接预测、图像标注、药物设计等应用中都可以找到类似的问题，作为匹配来自两个不同空间的对象的一般任务。技术

2个

引入的方法可以概括为更通用的机器学习技术，在本调查中称为学习匹配。

H. Li 和 J. Xu。*搜索中的语义匹配*. 基础和趋势®
信息检索，卷。7，没有。5，第 343–469 页，2013 年。
DOI：10.1561/1500000035。

R在

全文可在：http://dx.doi.org/10.1561/1500000035

1个

介绍

1.1 查询文档不匹配

一个成功的搜索引擎必须擅长相关性、覆盖面、新鲜度、响应时间和用户界面。其中，相关性[156,171,157]是最重要的因素，也是本次调查的重点。

本次调查主要以一般网络搜索为例。然而，所讨论的问题不仅限于网络搜索；它们存在于所有其他搜索中，例如企业搜索、桌面搜索以及问答。

搜索仍然严重依赖词袋方法或基于术语的方法。也就是说，查询和文档表示为词袋（术语），文档根据文档术语进行索引，“相关”文档根据查询术语检索，查询和检索文档之间的相关性得分根据匹配计算query term 和 document term 之间的度数，最后检索到的文档根据相关性得分进行排序。这种简单的方法在实践中非常有效，它仍然构成了现代搜索系统的基础 [131、52、6]。

表 1.1：查询文档不匹配的示例。

询问	文档	学期匹配	语义的匹配
西雅图最好的酒店	西雅图最好的酒店	部分的	是的
泳池时间表	游泳池时间表	部分的	是的
自然对数变换	对数变换	部分的	是的
中国香港	中国香港	部分的	不
为什么窗户那么贵	为什么mac那么贵	部分的	不

然而，词袋方法也有局限性。它有时会遇到查询文档不匹配的缺点。对于商业网络搜索引擎报告的大多数不满意案例，用户抱怨他们找不到信息，而信息确实存在于系统中，原因是查询和文档不匹配。在其他研究中观察到类似的趋势（参见 [206, 207]）

术语级别的匹配度高并不一定意味着相关性高，反之亦然。例如，如果query是“ny times”，而文档只包含“New York Times”，那么query和文档在term level的匹配度很低，尽管它们是相关的。表 1.1 中给出了查询文档不匹配的更多示例。[1个]

Query document mismatch 发生在搜索者和作者使用不同的术语（表示）来描述同一个概念时，由于人类语言的本质，这种现象普遍存在，即相同的意思可以用不同的表达方式和相同的表达方式来表达可以代表不同的意思。根据 Furnas 等人的说法，平均有 80-90% 的时间，两个人会用不同的表示形式命名相同的概念 [67]。

[1个]中国孔是美国演员。

表 1.2：关于“太阳和地球之间的距离”的问题。

“多远”地球太阳“多远”太阳	地球到太阳的平均距离 太阳到地球的平均距离 地球到太阳的平均距离 地球到太阳的距离
地球到太阳的平均距离 地球到太阳的距离 太阳到地球的距离	
地球和太阳之间的距离 地球到太阳的距离 地球到太阳的距离 地球到太阳的距离 “多远” 太阳 地球	地球到太阳的距离 地球到太阳的距离 地球到太阳的距离 太阳到地球的距离 太阳到地球的距离 太阳到地球的距离 太阳到地球的距离 地球到多远向太阳
地球离太阳多远 地球离太阳多远 太阳到地球的距离	

表 1.2 显示了代表相同搜索需求“太阳与地球之间的距离”的示例查询，表 1.3 显示了代表相同搜索需求“Youtube”的示例查询，这些查询是从商业搜索引擎的搜索日志中收集的 [117]。理想情况下，我们希望看到搜索系统针对不同的查询变体返回相同的结果。然而，网络搜索引擎仍然不能有效地满足需求。这是不匹配问题的另一方面。

在网络搜索中，查询文档不匹配更容易发生在尾页和尾查询上。这是因为对于头页和头查询，通常附加的信息更多。一个首页可能在搜索日志中有大量的锚文本和相关查询，它们为页面提供了不同的表示。如果查询与任何表示匹配，则匹配度将很高。然而，尾页很少发生这种情况。因此，不匹配是搜索中长尾挑战的典型示例。

1.2 搜索中的语义匹配

不匹配的根本原因是搜索时没有进行语言分析。计算机理解语言很难，

表 1.3：关于“Youtube”的查询。

优酷	优管	你管
ytube	优酷	宇管
优图博	YouTuber	youtubecom
youtube om	YouTube 音乐视频	YouTube视频
YouTube	youtube com	youtube公司
youtub 网站	you tube 音乐录影带 you	你管
优酷	tube com yourtube you	你的管
YouTube	tub	你管视频剪辑中国的膳
YouTube视频	www you tube com	宿条件 youtube com
优酷网	www youtube com	www youtube公司
优酷	www 你管	www utube com
ww youtube com	www 优管	www 你管
视频	优酷网	优管
你管com	优酷	你管视频
u 管	我的管	图管
外管	我们的管子	图管

但是，如果不是不可能的话。一种超越词袋的更现实的方法，在本次调查中称为语义匹配，将进行更多的查询分析和文档分析，以表示查询的含义和具有更丰富表示的文档，然后与表示进行查询文档匹配. 分析可以包括术语归一化、词组分析、词义分析、主题分析和结构分析，可以从形式方面、词组方面、词义方面、主题方面和结构方面进行匹配，如图1.1所示. 直观上，如果query和aspect所代表的文档的含义相同，那么它们应该很好地匹配，从而被认为是相关的。在实践中，查询和文档的更多方面可以匹配，查询和文档越有可能相关。通过语义匹配，我们可以预期可以成功克服查询文档不匹配的挑战。

术语规范化，包括亚洲语言的分词、欧洲语言的复合、欧洲语言的拼写错误校正，通常应该在查询文档匹配之前进行。我们将术语规范化称为形式方面的匹配。查询文档在phrase方面的匹配是指两者应该在phrase层面进行匹配，而不是word层面。例如，如果query是“hot dog”，那么它应该被识别为一个词组并匹配文档中完全相同的词组，而不应该单独匹配文档中的单词“hot”和“dog”。词义方面的匹配是将query中的词组与文档中具有相同词义的词组进行匹配。例如，“ny”应该匹配“New York”。如果查询和文档具有相同的主题，那么它们应该在主题方面匹配。例如，如果查询是“microsoft office”而文档是关于 Microsoft Word、PowerPoint 和 Excel 的，那么两者在主题方面应该匹配。查询和文档也可以在结构方面进行匹配，其中结构是指语言结构。例如，查询“distance between sun and earth”与文档标题“how far is sun from earth”相匹配（请注意，这两个表达具有非常不同的语言结构）。

我们还可以考虑其他方面的查询文档匹配，例如语义类和命名实体。我们将在第 9 节的结论和开放问题中对此进行讨论。

语义匹配也是计算机科学其他领域中使用的一个术语，它代表了一个与本次调查不同的概念。给定两个类似图的结构，例如两个数据库模式，语义匹配被定义为一个运算符，它标识两个结构中在语义上相互对应的节点 [73]。

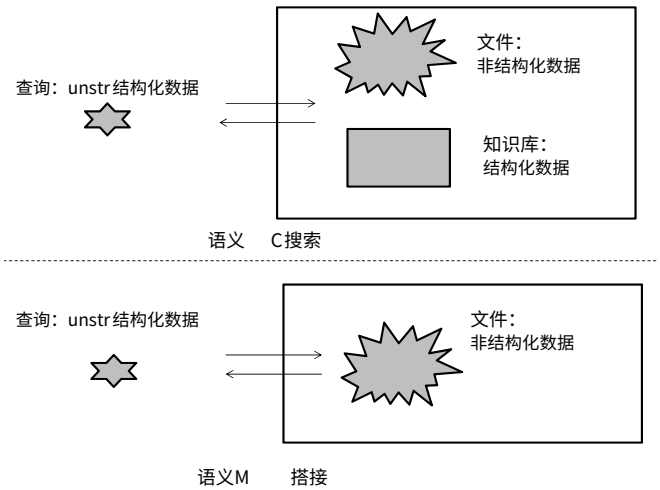
语义匹配也不同于所谓的语义搜索，不同的研究者对其有不同的定义。其中之一旨在通过使用来自语义网的信息（例如，[77]）来丰富传统搜索系统的搜索结果。例如，查询“yo-yo ma”的搜索结果在语义搜索中增加了大提琴手的图像、音乐会时间表、音乐专辑等。Bast 等人的语义搜索。要求用户制定一个

图 1.1：语义匹配：如果query和document在形式、词组、意义、主题、结构等方面所表达的含义相同，则认为它们相互匹配，被认为是相关的。

使用描述实体之间关系的运算符进行查询，结合从文档和本体中找到的信息，并返回给用户。支持特殊搜索需求，例如“寻找具有可食用叶子和原产于欧洲的植物”[11]。相比之下，我们这里关注的语义匹配是在搜索引擎内部进行的，用户不需要做任何与常规搜索不同的事情。

图 1.2 说明了语义匹配和语义搜索之间的区别。语义匹配涉及通过查询来搜索文档，其中文档和查询都是非结构化数据。语义搜索通常关注通过查询来搜索文档和知识库，其中文档和查询是非结构化数据，而知识库是结构化数据。

查询文档不匹配在信息检索 (IR) 的悠久历史中得到了研究。在传统的 IR 中，查询扩展、伪相关反馈和潜在语义索引 (LSI) 等方法得到了深入研究和广泛应用。如今，网络搜索收集了大量日志数据，并且开发了先进的机器学习技术。我们真的可以更有效地利用大数据和机器学习



数字　1.2：语义的　匹配与语义搜索。

解决挑战　查询的范围　文档不匹配，如 expl　艾恩于这项调查。

1.3 匹配　和兰金　G

在传统的IR中，搜索中排序和匹配的区别并不明确。给定一个查询，从索引中检索文档，并在查询和每个文档之间进行匹配。文档相对于查询的相关性表示为两者之间的匹配度，使用IR模型（匹配模型）如BM25或信息检索语言模型（LM4IR）计算。之后，根据匹配分数对文档进行排名（排序）。在这样的框架下，匹配分数和排名分数是等价的。2个

网络搜索发生了变化。文档（网页）的重要性被发现对相关性排名有用，重要性得分

2个我们注意到，在现代网络搜索中，不仅要考虑相关性，还要考虑新鲜度、多样性和其他因素。我们限制自己在本次调查中的相关性。

通过PageRank等模型计算的网页需要纳入排名机制。此外，还有许多指示查询与文档之间的相关性（匹配）程度的信号，并且可以计算代表这些信号的匹配分数。如何将匹配分数和重要性分数结合起来成为一个关键问题。一种简单的方法是线性组合分数并手动调整权重。还可以考虑使用训练数据自动构建排名模型的更复杂的机器学习技术。事实上，用于此目的的机器学习技术，称为学习排名，已经被深入研究并广泛应用于网络搜索[128、115]。因此，在网络搜索中，

如下所述，已经开发了用于学习查询和文档（通常是异构对象）之间匹配度的机器学习技术，在本文中称为学习匹配。学会搭配其实就是*特征学习*从机器学习的角度来看，学习排名。

1.4 其他任务中的语义匹配

信息检索和自然语言处理中的其他任务也依赖于*语义匹配*，例如释义和文本蕴含 [62、54]、问答 [21]、跨语言信息检索 (CLIR) [141、140]、在线广告 [31]、相似文档检测 [32、33] 和简短文本对话 [176、130]。表 1.4 总结了任务的特征。

例如，CLIR 是信息检索的一个子领域，涉及在检索另一种语言的文档时接收一种语言的查询的问题。任务中自然需要将查询或文档从一种语言翻译成另一种语言。两种语言的query和document的不匹配对CLIR和形式方面（复合、分词、拼写纠错）、意义方面（选择）的匹配提出了更大的挑战

的翻译）和主题方面也被尝试和验证是有帮助的 [141, 140]。

再例如，在线广告利用网络来传递营销信息并吸引消费者。它通常涉及在其网站上展示广告的出版商和提供广告的广告商。给定一些广告，需要找到合适的网站进行展示，即对发布者的内容和广告商的广告进行有效匹配。不匹配在这里也是不可避免的。已经提出了一些方法来解决形式方面、意义方面和主题方面的不匹配挑战 [31]。

短文本对话是最近提出的一个研究问题 [176, 130]。它由人与计算机之间的一轮对话组成，前者是人的消息，后者是计算机对消息的评论。短文本对话构成了自然语言对话的一个步骤，它也将问答归为特例。在基于检索的方法中，还需要考虑消息和评论之间的语义匹配，在该方法中，对大量消息和评论对进行索引，并根据消息检索、选择和返回最合适的评论。还提出了解决任务中的不匹配问题的方法 [176、130]。

1.5 搜索中语义匹配的机器学习

一个自然的问题是，是否可以使用机器学习技术来自动学习搜索中语义匹配的模型。这正是我们在本次调查中要解决的问题。

该任务可以形式化为匹配模型的学习$F(q, d)$ 或条件概率模型$P(r|q, d)$ 使用监督学习技术或学习条件概率模型$P(q|d)$ 使用无监督学习技术，其中q表示查询，d表示文档，并且r表示相关级别。请注意，这里的查询和文档被视为不同的（异构）对象。

表 1.4：需要语义匹配的任务的特征。任务中涉及两个自然语言文本（A 和 B）。

任务	文本类型	之间的关系 文本
搜索	A=查询， B=文件	关联
问答	A=问题， B=答案	回答问题—— 化
跨语言 IR	A=查询， B=文件 （不同语言）	关联
简短的文字对话	A=文字，B=文字	消息和通信 换货
相似文件检测在线广告	A=文字，B=文字 A=查询，B=广告。	相似 作为广告的相关性。
释义	A=句子， B=句子	等价
文本蕴涵	A=句子， B=句子	蕴涵

可以定义不同的模型，显式或隐式表示语义匹配，即在形式方面、短语方面、意义方面、主题方面、结构方面等不同方面进行匹配。由于query document mismatch是一个长尾现象，因此有必要假设没有单一的信号是足够的，并在不同方面构建匹配模型，并结合它们在相关性排序中的用途。

以下是一些经过充分研究的方法，包括通过查询重构匹配、与术语依赖模型匹配、与翻译模型匹配、与主题模型匹配以及与潜在空间模型匹配。本调查将详细解释这些方法。

通过查询重新制定匹配旨在重新制定查询，以便它可以与文档中的语义等价表达式更好地匹配。查询的重新表述包括拼写

纠错、分词、合并等。查询重组的主要问题包括原始查询的重写、原始查询和重组查询的搜索结果的混合、相似查询的挖掘以及查询扩展。

词袋方法的一个直接扩展是根据查询和文档中的多个词执行匹配。这正是术语依赖模型中描述的过程。可以用模型表示查询术语和文档术语之间的不同匹配关系，例如，术语在查询和文档中的共现。直觉上，如果几个术语在查询和文档中同时出现，那么它们可能代表相同的概念并表明更强的相关性。

查询与文档的一部分（例如标题）之间的匹配可以建模为释义或翻译，其中将一种语言表达转换为另一种语言表达。之前已经提出将匹配作为统计翻译任务，并且该方法最近在网络搜索中取得了重大进展，部分原因是大量的点击数据变得可用并且可以用作训练数据。

给定一组文档，主题建模技术可以帮助找到文档的主题，其中每个主题由多个单词表示。已经提出了概率和非概率模型。在搜索中，可以检测query的主题和文档的主题，并根据主题进行query和文档的匹配。

我们可以在两个不同的向量空间中表示查询和文档，基于点击数据中的查询文档关联将它们映射到一个较低维度的隐藏语义空间，并在潜在语义空间中进行查询和文档之间的匹配。这是与潜在空间模型匹配的方法的基本思想。许多传统的 IR 模型，如向量空间模型 (VSM)、BM25 和 LM4IR 都可以解释为潜在空间模型的特例，因此潜在空间模型对于 IR 来说是非常基础的。

两个异构对象之间的匹配不限于搜索。它存在于许多其他应用中，包括释义和文本蕴含、问答、在线广告、跨语言信息检索、相似文档检测、短文本对话、机器翻译、推荐系统（协同过滤）、链接预测、图像注释和药物设计. 将在不同应用程序中开发的技术概括为更通用的机器学习方法是必要且重要的，以便更深入和更广泛地研究这些技术。在本次调查中，我们将其称为学习匹配。

1.6 关于本次调查

本次调查侧重于基本问题，以及搜索中查询文档匹配的最新解决方案。解释的想法和解决方案可能会激励工业从业者将研究成果转化为产品。介绍的方法和进行的讨论也可能会激发学术研究人员寻找新的研究方向和方法。

第 2 节给出了用于搜索中查询文档匹配的机器学习公式，并展示了它在网络搜索中的实现。第 3-7 节描述了用于查询文档匹配的五组学习技术，即通过查询重构匹配、与术语依赖模型匹配、与翻译模型匹配、与主题模型匹配以及与潜在空间模型匹配。第 8 节描述了技术的概括、学习匹配，并介绍了协作过滤和释义和文本蕴含的方法。第 9 节总结调查并讨论未解决的问题。第2-8节是独立的，因此读者可以根据自己的兴趣和需要选择章节阅读。

本次调查更侧重于机器学习和语义匹配。一些调查论文或书籍涵盖了一些相关主题，例如 LM4IR [204]、查询扩展 [40]、搜索和浏览日志

挖掘 [163、94]，以及 IR [135] 上以特征为中心的视图。也鼓励读者参考这些材料。

我们假设读者对机器学习和信息检索有一定的了解。那些想了解更多关于这些领域的基础知识的人应该参考相关的书籍和论文。本次调查涉及的机器学习技术包括统计语言模型[204]、统计机器翻译[99]、学习排序[128、115、116]、图形模型[24]、主题模型[25]、矩阵分解[103]、内核方法[158]、稀疏方法[3个]和深度学习[4个]. 关于信息检索的基本技术的解释可以在 IR [131, 52, 6] 的教科书中找到。

[3个]可以在 www.di.ens.fr/fbach/ 找到巴赫的稀疏方法教程。

[4个]有关深度学习的教程可以在 www.deeplearning.net/tutorial/ 找到。

参考

[1] Jacob Abernethy、Francis Bach、Theodoros Evgeniou 和 Jean-Philippe Vert。协同过滤的新方法：具有频谱正则化的算子估计。*J.马赫。学习。水库。*, 10:803–826, 2009 年 6 月。

[2] Deepak Agarwal 和 Bee-Chung Chen。基于回归的潜在因子模型。在*第 15 届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集*, KDD '09，第 19–28 页，美国纽约州纽约市，2009 年。ACM。

[3] Farooq Ahmad 和 Grzegorz Kondrak。 学习拼写错误来自搜索查询日志的模型。在*人类语言技术与自然语言处理经验方法会议论文集*，HLT '05，第 955–962 页，美国宾夕法尼亚州斯特劳兹堡，2005 年。计算语言学协会。

[4] Jaime Arguello、Jonathan L. Elsas、Jamie Callan 和 Jaime G. Carbonell。用于博客推荐的文档表示和查询扩展模型。在*网络日志和社交媒体国际会议*, 第 10-18 页。美国人工智能出版社，2008 年。

[5] Ricardo Baeza-Yates、Carlos Hurtado 和 Marcelo Mendoza。用于提高网页排名的查询聚类。在 JesÍšs Favela、Ernestina Menasalvas 和 Edgar ChÍćvez 的编辑中，*网络智能的进步*, 第 3034 卷*计算机科学讲义*, 第 164-175 页。施普林格柏林海德堡出版社，2004 年。

[6] Ricardo A. Baeza-Yates 和 Berthier A. Ribeiro-Neto。*现代的信息检索——搜索背后的概念和技术*. Pearson Education Ltd.，英格兰哈洛，第 2 版，2011 年。

[7] Bing Bai、Jason Weston、David Grangier、Ronan Collobert、Kunihiko Sadamasa、Yanjun Qi、Olivier Chapelle 和 Kilian Weinberger。监督语义索引。在*第 18 届 ACM 信息与知识管理会议论文集*, CIKM '09, pages 187–196, New York, NY, USA, 2009. ACM。

[8] Bing Bai、Jason Weston、David Grangier、Ronan Collobert、Kunihiko Sadamasa、Yanjun Qi、Olivier Chapelle 和 Kilian Weinberger。学习使用（很多）单词特征进行排名。*信息。退货。*, 13(3):291–314, 2010 年 6 月。

[9] Suhrid Balakrishnan 和 Sumit Chopra。协同排名。在 *第五届 ACM 国际网络搜索与数据挖掘会议论文集*, WSDM '12，第 143–152 页，美国纽约州纽约市，2012 年。ACM。

[10] Niranjan Balasubramanian、Giridhar Kumaran 和 Vitor R. Carvalho。探索减少长时间的 Web 查询。在*第 33 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '10, pages 571–578, New York, NY, USA, 2010. ACM。

[11] Hannah Bast、Florian Bäurle、Björn Buchhold 和 Elmar Haussmann。Broccoli：触手可及的语义全文搜索。*CoRR*, 2012.

[12] 道格比弗曼和亚当伯杰。搜索引擎查询日志的聚集聚类。在*第六届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集*, KDD '00, pages 407–416, New York, NY, USA, 2000. ACM。

[13] Steven M. Beitzel、Eric C. Jensen、Ophir Frieder、David Grossman、David D. Lewis、Abdur Chowdhury 和 Aleksandr Kolcz。使用标记和未标记训练数据的自动 Web 查询分类。在 *第 28 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '05, pages 581–582, New York, NY, USA, 2005. ACM。

[14] Michael Bendersky 和 W. Bruce Croft。分析大规模搜索日志中的长查询。在*2009 年网络搜索点击数据研讨会论文集*, WSCD '09，第 8-14 页，美国纽约州纽约市，2009 年。ACM。

[15] Michael Bendersky 和 W. Bruce Croft。使用查询超图对信息检索中的高阶术语依赖性建模。
在*第 35 届国际 ACM SIGIR 信息检索研究与发展会议论文集*，SIGIR '12，第 941–950 页，美国纽约州纽约市，2012 年。ACM。

[16] Michael Bendersky、W. Bruce Croft 和 David A. Smith。搜索查询的联合注释。在*计算语言学协会第 49 届年会论文集：人类语言技术*, ACL '11, 第 102–111 页。计算机语言学协会，2011 年。

[17] Michael Bendersky、Donald Metzler 和 W. Bruce Croft。使用加权依赖模型学习概念重要性。在*第三届 ACM Web 搜索与数据挖掘国际会议论文集*, WSDM '10，第 31–40 页，美国纽约州纽约市，2010 年。ACM。

[18] Michael Bendersky、Donald Metzler 和 W. Bruce Croft。具有多个信息源的有效查询公式。在*第五届 ACM 国际网络搜索与数据挖掘会议论文集*, WSDM '12，第 443–452 页，美国纽约州纽约市，2012 年。ACM。

[19] Yoshua Bengio、Holger Schwenk、让-塞巴斯蒂安·塞纳卡尔、弗雷德里克·莫林和让-吕克·高万。神经概率语言模型。
在黎明。福尔摩斯和 LakhmiC。耆那教的编辑们，*机器学习的创新*, 第 194 卷*模糊与软计算研究*, 第 137-186 页。施普林格柏林海德堡出版社，2006 年。

[20] Paul N. Bennett、Krysta Svore 和 Susan T. Dumais。分类增强排名。在*第 19 届万维网国际会议论文集*, WWW '10, pages 111–120, New York, NY, USA, 2010. ACM。

[21] Adam Berger、Rich Caruana、David Cohn、Dayne Freitag 和 Vibhu Mittal。 弥合词汇鸿沟：回答问题的统计方法*第 23 届国际 ACM SIGIR 论发现。 文集*
信息检索研究与发展会议, SIGIR '00, pages 192–199, New York, NY, USA, 2000. ACM。

[22] 亚当·伯杰和约翰·拉弗蒂。作为统计翻译的信息检索。在*第 22 届国际 ACM SI-GIR 信息检索研究与发展会议论文集*, SIGIR '99, pages 222–229, New York, NY, USA, 1999. ACM。

[23] Shane Bergsma 和 Qin Iris Wang. 学习名词短语查询切分。在*2007 年自然语言处理和计算自然语言学习实证方法联合会议论文集*, EMNLP-CoNLL '07, pages 819–826, Prague, Czech Republic, June 2007. 计算语言学协会。

[24] 克里斯托弗・M・毕晓普。*模式识别与机器学习（信息科学与统计学）*. Springer-Verlag New York, Inc.，锡考克斯，新泽西州，美国，2006 年。

[25] 戴维・布莱。概率主题模型。*公社。美国计算机学会*, 55(4):77–84, 2012 年 4 月。

[26] David M. Blei、Andrew Y. Ng 和 Michael I. Jordan。潜在狄利克雷分布。*J.马赫。学习。水库。*, 3:993–1022, 2003 年 3 月。

[27] Paolo Boldi、Francesco Bonchi、Carlos Castillo、Debora Donato、Aristides Gionis 和 Sebastiano Vigna。查询流图：模型和应用程序。在*第 17 届 ACM 信息和知识管理会议论文集*, CIKM '08, pages 609–618, New York, NY, USA, 2008. ACM。

[28] Thorsten Brants、Francine Chen 和 Ioannis Tsochantaridis。具有概率潜在语义分析的基于主题的文档分割。在*第十一届信息与知识管理国际会议论文集*, CIKM '02, pages 211–218, New York, NY, USA, 2002. ACM。

[29] 埃里克・布里尔和罗伯特・摩尔。用于噪声信道拼写校正的改进错误模型。在*计算语言学会第38届年会论文集*，ACL '00，第 286–293 页，美国宾夕法尼亚州斯特劳兹堡，2000 年。计算语言学协会。

[30] Andrei Broder、Peter Ciccolo、Evgeniy Gabrilovich、Vanja Josifovski、Donald Metzler、Lance Riedel 和 Jeffrey Yuan。在线扩展赞助搜索的稀有查询。在*第 18 届万维网国际会议论文集*, WWW '09, pages 511–520, New York, NY, USA, 2009. ACM。

[31] Andrei Broder、Marcus Fontoura、Vanja Josifovski 和 Lance Riedel。上下文广告的语义方法。在*第 30 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '07, pages 559–566, New York, NY, USA, 2007. ACM。

[32] 安德烈·布罗德。关于文件的相似性和包含性。在*序列的压缩和复杂性会议记录 1997*, SEQUENCES '97, pages 21–, Washington, DC, USA, 1997. IEEE Computer Society。

[33] 安德烈·布罗德。识别和过滤近似重复的文档。在*第 11 届组合模式匹配年会论文集*, COM '00，第 1-10 页，伦敦，英国，英国，2000 年。施普林格出版社。

[34] Peter F. Brown、Vincent J. Della Pietra、Stephen A. Della Pietra 和 Robert L. Mercer。统计机器翻译的数学：参数估计。*电脑。语言学家。*，19(2):263–311, 1993 年 6 月。

[35] 范布，李航，朱晓燕．字符串重写内核。在*计算语言学协会第 50 届年会论文集：长篇论文 - 第 1 卷*，ACL '12，第 449–458 页，美国宾夕法尼亚州斯特劳兹堡，2012 年。计算语言学协会。

[36] 范布，李航，朱晓燕．字符串重写内核介绍。在*第二十三届人工智能国际联合会议论文集*, IJCAI'13, 第 2982–2986 页。美国人工智能出版社，2013 年。

[37] Robin D Burke、Kristian J Hammond、Vladimir Kulyukin、Steven L Lytinen、Noriko Tomuro 和 Scott Schoenberg。从常见问题文件中回答问题：使用常见问题解答系统的经验。*人工智能杂志*, 18(2):57, 1997.

[38] 曹桂红、聂建云、高建峰和史蒂芬·罗伯逊。为伪相关反馈选择好的扩展项。在 *第 31 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '08, pages 243–250, New York, NY, USA, 2008. ACM。

[39] 曹欢欢, 姜大新, 裴健, 何奇, 廖振, 陈恩宏, 李航. 通过挖掘点击率和会话数据的上下文感知查询建议。在*第 14 届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集*, KDD '08, pages 875–883, New York, NY, USA, 2008. ACM。

[40] 克劳迪奥·卡皮内托和乔瓦尼·罗马诺。信息检索中自动查询扩展的综述。*ACM 计算机。生存。*, 44(1):1:1–1:50，2012 年 1 月。

[41] Lara D. Catledge 和 James E. Pitkow。表征万维网中的浏览策略。*电脑。网络。ISDN 系统。*，27(6):1065–1073，1995 年 4 月。

[42] 摩西・查里卡尔。来自舍入算法的相似性估计技术。在*第三届 ACM 计算理论年会论文集*, STOC '02，第 380–388 页，美国纽约州纽约市，2002 年。ACM。

[43] 清晨, 穆丽, 明周. 使用网络搜索结果改进查询拼写更正。在*2007 年自然语言处理和计算自然语言学习实证方法联合会议论文集*, EMNLP-CoNLL '07，第 181-189 页。美国中文网，2007 年。

[44] 陈天琪, 李航, 杨强0001, 余勇．使用梯度提升的一般函数矩阵分解。在*ICML '13：第 30 届机器学习国际会议论文集*, 第 28 卷*JMLR 会议记录*, 第 436–444 页，2013 年。

[45] 陈天琪, 赵征, 陆秋霞, 张维南, 于勇. 基于特征的矩阵分解。*CoRR*, abs/1109.2271, 2011.

[46] 蒋大卫．基于分层短语的翻译。*电脑。语言学家。*, 33(2):201–228, 2007 年 6 月。

[47] Freddy YY Choi、Peter Wiemer-Hastings 和 Johanna Moore。文本分割的潜在语义分析。在 Lillian Lee 和 Donna Harman 的编辑中，*2001 年自然语言处理经验方法会议论文集*，EMNLP '01，第 109–117 页，2001 年。

[48] CW 克莱弗登。*索引语言比较实验测试中相关性评估变化的影响*. 克兰菲尔德图书馆报告。克兰菲尔德学院 技术，1970 年。

[49]迈克尔・柯林斯和奈杰尔・达菲。用于解析和标记的新排名算法：离散结构上的内核和投票感知器。
在*计算语言学会第40届年会论文集*，ACL '02，第 263–270 页，美国宾夕法尼亚州斯特劳兹堡，2002 年。计算语言学协会。

[50] 尼克・克拉斯韦尔和马丁・苏默。点击图上的随机游走。在*第 30 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '07, pages 239–246, New York, NY, USA, 2007. ACM。

[51] W. Bruce Croft, Michael Bendersky, Hang Li, and Gu Xu. 查询表示和理解研讨会。*SIGIR论坛*, 44(2):48–53, 2011 年 1 月。

[52] W. Bruce Croft、Donald Metzler 和 Trevor Strohman。 *搜索引擎：实践中的信息检索*. Addison-Wesley 出版公司，美国，第 1 版，2009 年。

[53] Silviu Cucerzan 和 Eric Brill。拼写更正是一个利用网络用户集体知识的迭代过程。在*2004 年自然语言处理经验方法会议论文集*, EMNLP '04，第 293-300 页。美国中文网，2004 年。

[54] Ido Dagan、Oren Glickman 和 Bernardo Magnini。帕斯卡识别文本蕴涵挑战。在*第一届机器学习挑战国际会议论文集：评估预测不确定性视觉对象分类和识别文本蕴含*, MLCW'05, 第 177–190 页。施普林格出版社，柏林，海德堡，2006 年。

[55] Dipanjan Das 和 Noah A. Smith。将识别解释为概率准同步识别。在*ACL 第 47 届年会和 AFNLP 第 4 届自然语言处理国际联席会议论文集：第 1 卷 - 第 1 卷*，ACL '09，第 468–476 页，美国宾夕法尼亚州斯特劳兹堡，2009 年。计算语言学协会。

[56] Ali Dasdan、Paolo D'Alberto、Santanu Kolay 和 Chris Drome。使用搜索引擎查询接口自动检索相似内容。在*第 18 届 ACM 信息与知识管理会议论文集*, CIKM '09, pages 701–710, New York, NY, USA, 2009. ACM。

[57] Fabio De Bona、Stefan Riezler、Keith Hall、Massimiliano Ciaramita、Amaç Herdağdelen 和 Maria Holmqvist。从用户点击日志中学习查询相似性的密集模型。在*人类语言技术：计算语言学协会北美分会 2010 年年会*, HLT '10, pages 474–482, Stroudsburg, PA, USA, 2010. 计算语言学协会。

[58] Scott Deerwester、Susan T. Dumais、George W. Furnas、Thomas K. Landauer 和 Richard Harshman。通过潜在语义分析进行索引。 *美国信息科学学会杂志*, 41(6):391–407, 1990。

[59] 费尔南多・迪亚兹。正则化临时检索分数。在*第14届ACM信息与知识管理国际会议论文集*, CIKM '05，第 672–679 页，美国纽约州纽约市，2005 年。ACM。

[60] 费尔南多・迪亚兹和唐纳德・梅茨勒。使用大型外部语料库改进相关模型的估计。在*第 29 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '06, pages 154–161, New York, NY, USA, 2006. ACM。

[61] 丁浩，泷川一学，真冢博，朱山峰。预测药物与目标相互作用的基于相似性的机器学习方法：简要回顾。*生物信息学简报*，第 bbt056 页，2013 年。

[62] Bill Dolan、Chris Quirk 和 Chris Brockett。大型释义语料库的无监督构建：利用大规模并行新闻源。
在*第20届计算语言学国际会议论文集*, COLING '04，美国宾夕法尼亚州斯特劳兹堡，2004 年。计算语言学协会。

[63] Huizhong Duan 和 Bo-June (Paul) Hsu. 查询完成的在线拼写更正。在*第 20 届万维网国际会议论文集*, WWW '11, pages 117–126, New York, NY, USA, 2011. ACM。

[64] Susan T Dumais、Todd A Letsche、Michael L Littman 和 Thomas K Landauer。使用潜在语义索引的自动跨语言检索。在*AAAI 春季跨语言文本和语音检索研讨会*，第 15 卷，第 21 页，1997 年。

[65] Ofer Egozi、Shaul Markovitch 和 Evgeniy Gabrilovich。使用显式语义分析的基于概念的信息检索。*ACM 跨。信息。系统。*, 29(2):8:1–8:34，2011 年 4 月。

[66] Edward A. Fox 和 Joeseph A. Shaw。多个搜索的组合。在*第二届文本检索会议（TREC-2）*, 卷 500-215 的*NIST 特别出版物*, 第 243-252 页。美国国家标准技术研究院，1994 年。

[67] GW Furnas、TK Landauer、LM Gomez 和 ST Dumais。人机交流中的词汇问题。*公社。美国计算机学会*, 30(11):964–971, 1987 年 11 月。

[68] 高建峰, 何晓东, 聂建云. 用于网络搜索的基于点击的翻译模型：从单词模型到短语模型。
在*第19届ACM信息与知识管理国际会议论文集*, CIKM '10, pages 1139–1148, New York, NY, USA, 2010. ACM。

[69] 高建峰、聂建云．使用搜索日志进行查询扩展的基于概念的翻译模型。在*第 21 届 ACM 信息与知识管理国际会议论文集*, CIKM '12, pages 1:1–1:10, New York, NY, USA, 2012. ACM。

[70] 高建峰，聂建云，吴广元，曹桂红。用于信息检索的依赖语言模型。在*第 27 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '04, pages 170–177, New York, NY, USA, 2004. ACM。

[71] 高建峰、克里斯蒂娜·托塔诺娃和叶文头。用于网络搜索的基于点击率的潜在语义模型。在*第 34 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '11, pages 675–684, New York, NY, USA, 2011. ACM。

[72] 高建峰，袁伟，李晓，邓科峰，聂建云．平滑网络搜索排名的点击数据。在*第 32 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '09, pages 355–362, New York, NY, USA, 2009. ACM。

[73] Fausto Giunchiglia、Pavel Shvaiko 和 Mikalai Yatskevich。S-match：语义匹配的算法和实现。在*语义网论文集：研究与应用，第一届欧洲语义网研讨会*，ESWS '04，第 61-75 页。施普林格，2004 年。

[74] 安德鲁·R·戈尔丁和丹·罗斯。上下文　基于winnow的方法 *马赫。学*相关的拼写更正。1999 年 2 月。　　*习。*, 34(1-3):107–130,

[75] Jagadeesh Gorla、Stephen Robertson、Jun Wang 和 Tamas Jambor。信息匹配理论。*CoRR*, abs/1205.5569, 2012.

[76] 大卫·格兰吉尔和萨米·本吉奥。一种基于内核的判别性方法，用于从文本查询中对图像进行排名。*IEEE 跨。模式肛门。马赫。智能。*，30(8):1371–1384, 2008 年 8 月。

[77] R. Guha、Rob McCool 和 Eric Miller。语义搜索。在*第 12 届万维网国际会议论文集*, WWW '03, pages 700–709, New York, NY, USA, 2003. ACM。

[78] 郭家峰，顾旭，程雪琪，李航．查询中的命名实体识别。在*第 32 届国际 ACM SI-GIR 信息检索研究与发展会议论文集*, SIGIR '09, pages 267–274, New York, NY, USA, 2009. ACM。

[79] 郭家峰, 顾旭, 李航, 程雪琪. 用于查询细化的统一判别模型。在*第 31 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '08, pages 379–386, New York, NY, USA, 2008. ACM。

[80] 马蒂亚斯・哈根、马丁・波塔斯特、安娜・拜尔和本诺・斯坦因。实现最佳查询细分：毫无疑问。在*第 21 届 ACM 信息与知识管理国际会议论文集*, CIKM '12, pages 1015–1024, New York, NY, USA, 2012. ACM。

[81] 马蒂亚斯・哈根、马丁・波塔斯特、本诺・斯坦和克里斯托夫・布劳蒂加姆。重新访问查询分段。在*第 20 届万维网国际会议论文集*, WWW '11, pages 97–106, New York, NY, USA, 2011. ACM。

[82] 咏叹调 Haghighi 和露西 Vanderwende。探索多文档摘要的内容模型。在*人类语言技术论文集：计算语言学协会北美分会 2009 年年会*，NAACL '09，第 362–370 页，美国宾夕法尼亚州斯特劳兹堡，2009 年。计算语言学协会。

[83] David R. Hardoon 和 John Shawe-taylor。基 不同级别的 Kcca In*在第* 于内容的图像检索的精度。 *三国际* *基于内容的多媒体索引研讨会，IRISA*, 2003.

[84] David R. Hardoon、Sandor R. Szedmak 和 John R. Shawe-taylor。典型相关分析：学习方法应用概述。*神经计算。*, 16(12):2639–2664, 2004 年 12 月。

[85] Michael Heilman 和 Noah A. Smith。用于识别文本蕴涵、释义和问题答案的树编辑模型。在*人类语言技术：计算语言学协会北美分会 2010 年年会*, HLT '10, pages 1011–1019, Stroudsburg, PA, USA, 2010. 计算语言学协会。

[86] Ralf Herbrich、Thore Graepel 和 Klaus Obermayer。序数回归的支持向量学习。在*在国际人工神经网络会议论文集中*, 第 97–102 页，1999 年。

[87] Dustin Hillard、Stefan Schroedl、Eren Manavoglu、Hema Raghavan 和 Chirs Leggetter。提高赞助搜索中的广告相关性。在 *第三届 ACM Web 搜索与数据挖掘国际会议论文集*, WSDM '10，第 361–370 页，美国纽约州纽约市，2010 年。ACM。

[88] 托马斯·霍夫曼。概率潜在语义索引。在*第 22 届国际 ACM SIGIR 信息检索研究与发展会议论文集*，SIGIR '99，第 50-57 页，美国纽约州纽约市，1999 年。ACM。

[89] 黄建康、李峰建和 Yen-Jen Oyang。基于查询会话日志中的上下文信息的交互式网络搜索中的相关术语建议。*J. Am。社会。信息。科学。技术。*，54(7):638–649, 2003 年 5 月。

[90] 黄博森、何晓东、高建峰、邓力、Alex Acero 和 Larry Heck。使用点击数据学习网络搜索的深度结构化语义模型。在*第 22 届 ACM 国际信息会议论文集& 知识管理*, CIKM '13, pages 2333–2338, New York, NY, USA, 2013. ACM。

[91] 塞缪尔·休斯顿、J. 肖恩·卡尔佩珀和 W. 布鲁斯·克罗夫特。用于排序检索的索引词序列。*ACM 跨。信息。系统。*, 32(1):3:1– 3:26，2014 年 1 月。

[92] Aminul Islam 和 Diana Inkpen。使用 google web it 3-grams 进行真实单词拼写校正。在*2009 年自然语言处理经验方法会议论文集：第 3 卷 - 第 3 卷*, EMNLP '09, pages 1241–1249, Stroudsburg, PA, USA, 2009. 计算语言学协会。

[93] Kalervo Järvelin 和 Jaana Kekäläinen。基于累积增益的 ir 技术评估。*ACM 跨。信息。系统。*, 20(4):422–446, 2002 年 10 月。

[94] 姜大新, 裴健, 李航. 挖掘网络搜索的搜索和浏览日志：一项调查。*ACM 跨。智能。系统。技术。*, 4(4):57:1– 57:37，2013 年 10 月。

[95] 荣进, Alex G. Hauptmann, 和 Cheng Xiang Zhai. 用于信息检索的标题语言模型。在*第 25 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '02, pages 42–48, New York, NY, USA, 2002. ACM。

[96] Rosie Jones、Benjamin Rey、Omid Madani 和 Wiley Greiner。生成查询替换。在*第 15 届万维网国际会议论文集*, WWW '06, pages 387–396, New York, NY, USA, 2006. ACM。

[97] In-Ho Kang 和 GilChang Kim。Web 文档检索的查询类型分类。在*第 26 届国际 ACM SIGIR 信息检索研究与发展会议论文集*，SIGIR '03，第 64–71 页，美国纽约州纽约市，2003 年。ACM。

[98] Maryam Karimzadehgan 和 ChengXiang Zhai. 基于互信息的统计翻译模型估计，用于临时信息检索。在*第 33 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '10, pages 323–330, New York, NY, USA, 2010. ACM。

[99] 菲利普・科恩。*统计机器翻译*. 剑桥大学出版社，美国纽约州纽约市，第 1 版，2010 年。

[100] Philipp Koehn、Franz Josef Och 和 Daniel Marcu。基于统计短语的翻译。在*计算语言学协会北美分会 2003 年人类语言技术会议论文集 - 第 1 卷*，NAACL '03，第 48-54 页，美国宾夕法尼亚州斯特劳兹堡，2003 年。计算语言学协会。

[101]耶胡达科伦。因式分解与邻域相遇：一个多方面的协同过滤模型。在*第 14 届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集*, KDD '08, pages 426–434, New York, NY, USA, 2008. ACM。

[102]耶胡达科伦。邻居因素：可扩展且准确的协同过滤。*ACM 跨。知识 发现。数据*, 4(1):1:1–1:24，2010 年 1 月。

[103] 耶胡达・科伦、罗伯特・贝尔和克里斯・沃林斯基。推荐系统的矩阵分解技术。*电脑*, 42(8):30–37, 2009 年 8 月。

[104] Alexander Kotov 和 ChengXiang Zhai. Taping into knowledge base for concept feedback：利用 conceptnet 改进困难查询的搜索结果。在*第五届 ACM 国际网络搜索与数据挖掘会议论文集*, WSDM '12，第 403–412 页，美国纽约州纽约市，2012 年。ACM。

[105] Ralf Krestel、Peter Fankhauser 和 Wolfgang Nejdl。用于标签推荐的潜在狄利克雷分布。在*第三届 ACM 推荐系统会议论文集*, RecSys '09，第 61–68 页，美国纽约州纽约市，2009 年。ACM。

[106] Oren Kurland 和 Lillian Lee。语料库结构、语言模型和临时信息检索。在*第 27 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '04, pages 194–201, New York, NY, USA, 2004. ACM。

[107] TK Landauer、D. Laham 和 M. Derr。从段落到图表：信息可视化的潜在语义分析。*美国国家科学院院刊*, 101（增刊 1）：5214–5219，2004 年 4 月。

[108] Hao Lang、Donald Metzler、Bin Wang 和 Jin-Tao Li。使用分层马尔可夫随机场改进潜在概念扩展。在 *第19届ACM信息与知识管理国际会议论文集*, CIKM '10，第 249–258 页，美国纽约州纽约市，2010 年。ACM。

[109] Victor Lavrenko 和 W. Bruce Croft。基于相关性的语言模型。在*第 24 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '01, pages 120–127, New York, NY, USA, 2001. ACM。

[110] 马修・莱斯。用于支持详细查询的改进马尔可夫随机场模型。在*第 32 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '09, pages 476–483, New York, NY, USA, 2009. ACM。

[111] Daniel D. Lee 和 H. Sebastian Seung。非负矩阵分解算法。在 TK Leen、TG Dietterich 和 V. Tresp 的编辑中，*神经信息处理系统的进展 13*, 第 556–562 页。麻省理工学院出版社，2001 年。

[112]李俊和。多证据组合分析。在*第 20 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '97, pages 267–276, New York, NY, USA, 1997. ACM。

[113] Uichin Lee、Zhenyu Liu 和 Junghoo Cho。自动识别网络搜索中的用户目标。在*第 14 届万维网国际会议论文集*, WWW '05, pages 391–400, New York, NY, USA, 2005. ACM。

[114] Gina-Anne Levow、Douglas W. Oard 和 Philip Resnik。基于词典的跨语言信息检索技术。*信息。过程。管理。*, 41(3):523–547, 2005 年 5 月。

[115] 李航．学习排名信息检索和自然语言处理。
人类语言技术综合讲座,
4(1):1–113, 2011。

[116] 李航．学习排序的简短介绍。*IEICE 信息与系统汇刊*, 94-D(10):1854–1862, 2011。

[117] Hang Li, Gu Xu, W. Bruce Croft, Michael Bendersky, Ziqi Wang, and Evelyne Viegas。Qru-1：用于促进查询表示和理解研究的公共数据集。在*网络搜索点击数据研讨会论文集*, WSCD '12, 2012。

[118] 穆丽，杨章，朱木华，周明．探索基于分布相似性的查询拼写校正模型。

在*第21届国际计算语言学会议暨第44届计算语言学学会年会论文集*,
ACL-44，第 1025–1032 页，美国宾夕法尼亚州斯特劳兹堡，2006 年。计算语言学协会。

[119] Xiao Li, Ye-Yi Wang, 和 Alex Acero。从正则化点击图中学习查询意图。在*第 31 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '08, pages 339–346, New York, NY, USA, 2008. ACM。

[120] 李亚恩，段慧中，翟承祥．具有用于查询拼写校正的判别训练的广义隐马尔可夫模型。
在*第 35 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '12, pages 611–620, New York, NY, USA, 2012. ACM。

[121] 李亚恩，徐博君，翟承祥，王宽三。使用点击率进行信息检索的无监督查询分段。在*第 34 届国际 ACM SIGIR 信息检索研究与发展会议论文集*,
SIGIR '11, pages 285–294, New York, NY, USA, 2011. ACM。

[122] Yinghao Li, Wing Pong Robert Luk, Kei Shiu Edward Ho, and Fu Lai Korris Chung. 使用维基百科作为外部语料库改进弱的临时查询。在*第 30 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '07, pages 797–798, New York, NY, USA, 2007. ACM。

[123] 廖振，蒋大新，陈恩宏，裴建，曹欢欢，李航。从大规模搜索日志中挖掘概念序列以提供上下文感知查询建议。*ACM 跨。智能。系统。技术。*，3(1):17:1–17:40，2011 年 10 月。

[124] 大卫·利本-诺厄尔和乔恩·克莱因伯格。社交网络的链接预测问题。*J. Am。社会。信息。科学。技术。*, 58(7):1019–1031, 2007 年 5 月。

[125] 德康林和帕特里克潘特尔。发现问题回答的推理规则。*纳特。郎。工程。*，7(4):343–360, 2001 年 12 月。

[126] 肯尼斯 C 利特科夫斯基。使用语义关系三元组进行问答。在*在第 8 届文本检索会议 (TREC-8) 的记录中*, 第 349–356 页，1999 年。

[127] H. Liu 和 P. Singh。概念网— 一个实用的常识推理工具包。*BT技术杂志*, 22(4):211–226, 2004 年 10 月。

[128] 刘铁雁．学习排名信息检索。*成立。趋势信息 退货。*, 3(3):225–331, 2009 年 3 月。

[129] Yumao Lu, Fuchun Peng, Gilad Mishne, Xing Wei, and Benoit Dumoulin. 通过语义特征提高网络搜索相关性。在 *2009 年自然语言处理经验方法会议论文集：第 2 卷 - 第 2 卷*, EMNLP '09, pages 648–657, Stroudsburg, PA, USA, 2009. 计算语言学协会。

[130]路正东，李航．用于匹配短文本的深度架构。在 CJC Burges、L. Bottou、M. Welling、Z. Ghahramani 和
KQ Weinberger，编辑，*神经信息处理系统的进展 26*, 第 1367–1375 页。柯伦联合公司，2013 年。

[131] Christopher D. Manning、Prabhakar Raghavan 和 Hinrich Schütze。 *信息检索导论*. 剑桥大学出版社，纽约，纽约，美国，2008 年。

[132] K. Tamsin Maxwell 和 W. Bruce Croft。使用主题相关文本的紧凑查询词选择。在*第 36 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '13, pages 583–592, New York, NY, USA, 2013. ACM。

[133] Qiaozhu Mei、Dengyong Zhou 和 Kenneth Church。使用命中时间查询建议。在*第 17 届 ACM 信息和知识管理会议论文集*, CIKM '08, pages 469–478, New York, NY, USA, 2008. ACM。

[134] Aditya Krishna Menon 和 Charles Elkan。通过矩阵分解进行链接预测。在*2011 年欧洲数据库机器学习和知识发现会议记录 - 第二卷*, ECML PKDD'11, pages 437–452, Berlin, Heidelberg, 2011. Springer-Verlag。

[135] 唐纳德・梅茨勒。*信息检索的以特征为中心的观点*. 施普林格出版社，2012 年版，2011 年。

[136] 唐纳德・梅茨勒和 W. 布鲁斯・克罗夫特。术语依赖的马尔可夫随机场模型。在*第 28 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '05, pages 472–479, New York, NY, USA, 2005. ACM。

[137] 唐纳德・梅茨勒和 W. 布鲁斯・克罗夫特。使用马尔可夫随机场的潜在概念扩展。在*第 30 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '07, pages 311–318, New York, NY, USA, 2007. ACM。

[138] 亚历山德罗・莫斯奇蒂。用于依赖和组成句法树的高效卷积核。在*第 17 届欧洲机器学习会议论文集*, ECML'06, pages 318–329, Berlin, Heidelberg, 2006. Springer-Verlag。

[139] 亚历山德罗・莫斯奇蒂和法比奥・马西莫・赞佐托。用于从文本中进行关系学习的快速有效的内核。在*第 24 届国际机器学习会议论文集*, ICML '07，第 649–656 页，美国纽约州纽约市，2007 年。ACM。

[140] 聂建云．跨语言信息检索。*人类语言技术综合讲座*, 3(1):1–125, 2010。

[141] Douglas W Oard 和 Anne R Diekema。跨语言信息检索。*信息科学年度回顾 (ARIST)*, 33, 1998.

[142] Franz Josef Och 和 Hermann Ney。用于统计机器翻译的判别训练和最大熵模型。在*计算语言学会第40届年会论文集*，ACL '02，第 295–302 页，美国宾夕法尼亚州斯特劳兹堡，2002 年。计算语言学协会。

[143] Lawrence Page、Sergey Brin、Rajeev Motwani 和 Terry Winograd。pagerank 引文排名：为网络带来秩序。技术报告 1999-66，斯坦福信息实验室，1999 年 11 月。之前的编号 = SIDL-WP-1999-0120。

[144] 迪帕帕拉尼佩。从隐式用户反馈和文档结构中学习文档关于性。在*第 18 届 ACM 信息与知识管理会议论文集*, CIKM '09，第 365–374 页，美国纽约州纽约市，2009 年。ACM。

[145] Jae Hyun Park、W. Bruce Croft 和 David A. Smith。用于信息检索的准同步依赖模型。在*第 20 届 ACM 信息与知识管理国际会议论文集*, CIKM '11，第 17–26 页，美国纽约州纽约市，2011 年。ACM。

[146] 彭富春, Nawaaz Ahmed, Xin Li, and Yumao Lu. 用于 Web 搜索的上下文敏感词干提取。在*第 30 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '07, pages 639–646, New York, NY, USA, 2007. ACM。

[147] Yonggang Qiu 和 Hans-Peter Frei。基于概念的查询扩展。在*第 16 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '93, pages 160–169, New York, NY, USA, 1993. ACM。

[148] 史蒂芬·伦德尔。分解机。在*数据挖掘 (ICDM)，2010 年 IEEE 第 10 届国际会议*, 第 995–1000 页。IEEE，IEEE 计算机协会，2010 年。

[149] 史蒂芬·伦德尔。带有 libfm 的分解机。*ACM 跨。智能。系统。技术。*，3(3):57:1–57:22，2012 年 5 月。

[150] Stefan Riezler 和 Yi Liu. 使用单语统计机器翻译进行查询重写。*电脑。语言学家。*, 36(3):569–582, 2010 年 9 月。

[151] Eric Sven Ristad 和 Peter N. Yianilos。学习字符串编辑距离。 *IEEE 跨。模式肛门。马赫。智能。*, 20(5):522–532, 1998 年 5 月。

[152] 东南罗伯逊。IR 中的概率排序原则。*文献杂志*, 33(4):294–304, 1977。

[153] Stephen E Robertson、Steve Walker、Susan Jones、Micheline M Hancock-Beaulieu 和 Mike Gatford。霍加皮在 trec-3。*NIST 特别出版物 SP*, 第 109–109 页，1995 年。

[154] 罗曼·罗西帕尔和妮可·克雷默。偏最小二乘法的概述和最新进展。在 *2005 年子空间、潜在结构和特征选择国际会议论文集*, SLSFS'05, pages 34–51, Berlin, Heidelberg, 2006. Springer-Verlag.

[155] G. Salton、A. Wong 和 CS Yang。用于自动索引的向量空间模型。*公社。美国计算机学会*, 18(11):613–620, 1975 年 11 月。

[156] Tefko Saracevic。相关性：对信息科学概念的思考的回顾和框架。*美国信息科学学会杂志*, 26(6):321–343, 1975。

[157] 特夫科·萨拉切维奇。相关性：对文献的回顾和对信息科学概念的思考框架。第三部分：相关行为和影响。*J. Am。社会。信息。科学。技术。*，58(13):2126–2144, 2007 年 11 月。

[158] Bernhard Scholkopf 和 A lexander J. Smola。*使用内核学习：支持向量机、正则化、优化等*. 麻省理工学院出版社，美国马萨诸塞州剑桥市，2001 年。

[159] 威廉·罗布森·施瓦茨、阿尼鲁达·肯巴维、大卫·哈伍德和拉里·戴维斯。使用偏最小二乘法分析进行人体检测。在 *IEEE 第 12 届计算机视觉国际会议*, 第 24-31 页。IEEE，2009 年。

[160] Daniel Sheldon、Milad Shokouhi、Martin Szummer 和 Nick Craswell。Lambdamerge：合并查询重构的结果。在 *第四届 ACM 国际网络搜索与数据挖掘会议论文集*, WSDM '11，第 795–804 页，美国纽约州纽约市，2011 年。ACM。

[161] Dou Shen, Rong Pan, Jian-Tao Sun, Jeffrey Junfeng Pan, Kangheng Wu, Jie Yin, and Qiang Yang. Web 查询分类的查询扩充。*ACM 跨。信息。系统。*, 24(3):320–352, 2006 年 7 月。

[162] 史立新、聂建云．根据其实用程序使用各种术语依赖项。在*第19届ACM信息与知识管理国际会议论文集*, CIKM '10, pages 1493–1496, New York, NY, USA, 2010. ACM。

[163] 法布里齐奥·西尔维斯特里。挖掘查询日志：将搜索使用数据转化为知识。*成立。趋势信息 退货。*, 4(1-2):1-174, 2010 年 1 月。

[164] 阿米特·辛哈尔和费尔南多·佩雷拉。用于语音检索的文档扩展。在*第 22 届年度国际 ACM SIGIR 信息检索研究与发展会议论文集*，SIGIR '99，第 34–41 页，美国纽约州纽约市，1999 年。ACM。

[165] Richard Socher、Eric H. Huang、Jeffrey Pennin、Christopher D Manning 和 Andrew Y. Ng。用于释义检测的动态池化和展开递归自动编码器。在 J. Shawe-Taylor, RS Zemel, PL Bartlett、F. Pereira 和 KQ Weinberger，编辑，*神经信息处理系统的进展 24*, 第 801-809 页。柯伦联合公司，2011 年。

[166] Ruihua Song、Michael J. Taylor、Ji-Rong Wen、Hsiao-Wuen Hon 和 Yong Yu。从不同的角度查看术语接近度。在 *IR 研究论文集，第 30 届欧洲信息检索进展会议*, ECIR'08，第 346-357 页。施普林格出版社，柏林，海德堡，2008 年。

[167] Krysta M. Svore、Pallika H. Kanani 和 Nazan Khan。术语的跨度有多好？：利用接近度来改进网络检索。在*第 33 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '10, pages 154–161, New York, NY, USA, 2010. ACM。

[168] 谭斌和彭富春。使用生成语言模型和维基百科的无监督查询分割。在*第 17 届万维网国际会议论文集*, WWW '08, pages 347– 356, New York, NY, USA, 2008. ACM。

[169]陶涛，王宣慧，梅巧竹，翟承祥．具有文档扩展的语言模型信息检索。在*计算语言学协会北美分会人类语言技术会议主要会议论文集*，HLT-NAACL '06，第 407–414 页，美国宾夕法尼亚州斯特劳兹堡，2006 年。计算语言学协会。

[170] 涛涛、翟承祥．信息检索中邻近度度量的探索。在*第 30 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '07, pages 295–302, New York, NY, USA, 2007. ACM。

[171] Anastasios Tombros、Ian Ruthven 和 Joemon M. Jose。用户如何评估网页以查找信息。*J. Am。社会。信息。科学。技术。*, 56(4):327–344, 2005 年 2 月。

[172] 克里斯蒂娜・图塔诺娃 (Kristina Toutanova) 和罗伯特・摩尔 (Robert C. Moore)。用于改进拼写校正的发音建模。在*计算语言学会第40届年会论文集*，ACL '02，第 144–151 页，美国宾夕法尼亚州斯特劳兹堡，2002 年。计算语言学协会。

[173] 彼得・D・特尼。在网络上挖掘同义词：Pmi-ir 与 lsa on toefl。在*第 12 届欧洲机器学习会议论文集*, EMCL '01, pages 491–502, London, UK, UK, 2001. Springer-Verlag。

[174] 艾伦・M・沃里斯。使用词汇语义关系进行查询扩展。在 *第 17 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '94, pages 61–69, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[175] 王晨，毕克平，胡云华，李航，曹桂红。提取以搜索为中心的关键 n-gram，用于网络搜索中的相关性排名。在*第五届 ACM Web 搜索与数据挖掘国际会议论文集*, WSDM '12，第 343–352 页，美国纽约州纽约市，2012 年。ACM。

[176] 王浩，路正东，李航，陈恩红。用于研究短文本对话的数据集。在*2013 年自然语言处理实证方法会议论文集*，EMNLP '13，第 935–945 页。美国中文网，2013 年。

[177] 王建强和道格拉斯・W・奥德。跨语言信息检索的意义匹配。*信息。过程。管理。*, 48(4):631–653, 2012 年 7 月。

[178] 王凯、明兆彦、蔡达生。在基于社区的问答服务中寻找相似问题的句法树匹配方法。在*第 32 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '09, pages 187–194, New York, NY, USA, 2009. ACM。

[179] 王权，曹征，徐军，李航。用于可扩展主题建模的组矩阵分解。在*第 35 届国际 ACM SIGIR 信息检索研究与发展会议论文集*，SIGIR '12，第 375–384 页，美国纽约州纽约市，2012 年。ACM。

[180] 王全、徐军、李航和尼克·克拉斯韦尔。正则化潜在语义索引。在*第 34 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '11, pages 685–694, New York, NY, USA, 2011. ACM。

[181] Quan Wang, Jun Xu, Hang Li, and Nick Craswell. 正则化潜在语义索引：大规模主题建模的新方法。*ACM 跨。信息。系统。*, 31(1):5:1–5:44，2013 年 1 月。

[182]王宣晖、翟承祥．从搜索日志中挖掘术语关联模式以进行有效的查询重构。在*第 17 届 ACM 信息和知识管理会议论文集*, CIKM '08, pages 479–488, New York, NY, USA, 2008. ACM。

[183] 王紫琪，顾旭，李航，张明。一种快速准确的近似字符串搜索方法。在*计算语言学协会第 49 届年会论文集：人类语言技术 - 第 1 卷*，HLT '11，第 52–61 页，美国宾夕法尼亚州斯特劳兹堡，2011 年。计算语言学协会。

[184]邢伟和W. Bruce Croft。用于临时检索的基于 Lda 的文档模型。在*第 29 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '06, pages 178–185, New York, NY, USA, 2006. ACM。

[185] 文继荣、聂建云、张宏江。聚类搜索引擎的用户查询。在*第十届万维网国际会议论文集*, WWW '01, pages 162–168, New York, NY, USA, 2001. ACM。

[186] 吴浩成，胡云华，李航，陈恩红。网络搜索中相关性排名的查询细分。*CoRR*, abs/1312.0182, 2013.

[187] 魏武，李航，徐军．从带有元数据的点击二分图中学习查询和文档相似性。在 *第六届 ACM 国际网络搜索与数据挖掘会议论文集*, WSDM '13，第 687–696 页，美国纽约州纽约市，2013 年。ACM。

[188] 吴伟，路正东，李航．用于匹配查询和文档的学习双线性模型。*J.马赫。学习。水库。*, 14(1):2519–2548, 2013 年 1 月。

[189] Wei Wu, Jun Xu, Hang Li, and Satoshi Oyama. 使用内核方法学习用于搜索的稳健相关模型。*J.马赫。学习。水库。*, 12:1429–1458, 2011 年 7 月。

[190] 顾旭，杨双红，李航．使用弱监督的潜在狄利克雷分配从点击数据中挖掘命名实体。
在*第 15 届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集*, KDD '09, pages 1365–1374, New York, NY, USA, 2009. ACM。

[191]徐经芳、顾旭．学习罕见查询的相似度函数。在 *第四届 ACM Web 搜索与数据挖掘国际会议论文集*, WSDM '11，第 615–624 页，美国纽约州纽约市，2011 年。ACM。

[192] 徐金熙和 W. Bruce Croft。使用本地和全局文档分析进行查询扩展。在*第 19 届国际 ACM SIGIR 信息检索研究与发展会议论文集*，SIGIR '96，第 4-11 页，美国纽约州纽约市，1996 年。ACM。

[193] 徐军，李航，钟超良．使用内核的相关性排名。在 Pu-Jen Cheng、Min-Yen Kan、Wai Lam 和 Preslav Nakov 的编辑中，*信息检索技术*, 第 6458 卷 *计算机科学讲义*, 第 1-12 页。施普林格柏林海德堡出版社，2010 年。

[194] 徐军，魏武，李航，顾旭。解决术语不匹配的内核方法。在*万维网上第 20 届国际会议论文集*, WWW '11, pages 153–154, New York, NY, USA, 2011. ACM。

[195] 徐伟，刘新，龚一红．基于非负矩阵分解的文档聚类。在*第 26 届国际 ACM SIGIR 信息检索研究与发展会议论文集*, SIGIR '03, pages 267–273, New York, NY, USA, 2003. ACM。

[196] Gui-Rong Xue, Hua-Jun Zeng, Zheng Chen, Yong Yu, Wei-Ying Ma, WenSi Xi, and WeiGuo Fan. 使用网络点击数据优化网络搜索。在*第十三届 ACM 信息与知识管理国际会议论文集*, CIKM '04, pages 118–126, New York, NY, USA, 2004. ACM。

[197] 薛小兵，于涛，蒋大新，李航．从搜索日志数据中自动挖掘问题重构模式。在*计算语言学协会第 50 届年会论文集：短文 - 第 2 卷*，ACL '12，第 187–192 页，美国宾夕法尼亚州斯特劳兹堡，2012 年。计算语言学协会。

[198] Kenji Yamada 和 Kevin Knight。基于句法的统计翻译模型。在*计算语言学会第39届年会论文集*，ACL '01，第 523–530 页，美国宾夕法尼亚州斯特劳兹堡，2001 年。计算语言学协会。

[199] Yin Yang、Nilesh Bansal、Wisam Dakka、Panagiotis Ipeirotis、Nick Koudas 和 Dimitris Papadias。按单据查询。在*第二届 ACM 国际网络搜索与数据挖掘会议论文集*, WSDM '09，第 34–43 页，美国纽约州纽约市，2009 年。ACM。

[200] 邢毅和詹姆斯艾伦。利用主题模型进行信息检索的比较研究。在*第 31 届欧洲信息检索研究进展欧洲会议论文集*, ECIR '09，第 29–41 页，柏林，海德堡，2009 年。施普林格出版社。

[201] 邢毅和詹姆斯・艾伦。发现缺少的用于网络搜索的点击查询语言信息。在*第 20 届 ACM 信息与知识管理国际会议论文集*, CIKM '11, pages 153–162, New York, NY, USA, 2011. ACM。

[202] Wen-tau Yih、Kristina Toutanova、John C. Platt 和 Christopher Meek。学习文本相似性度量的判别预测。
在*第十五届计算自然语言学习会议论文集*，CoNLL '11，第 247–256 页，美国宾夕法尼亚州斯特劳兹堡，2011 年。计算语言学协会。

[203] Fabio massimo Zanzotto、Marco Pennacchiotti 和 Alessandro Moschitti。文本蕴涵识别的机器学习方法。
纳特。郎。工程。, 15(4):551–582, 2009 年 10 月。

[204]城乡斋．用于信息检索的统计语言模型批判性评论。*成立。趋势信息 退货。*, 2(3):137–213, 2008 年 3 月。

[205] 张贤、于浩、朱晓燕、李明和戴维・切里顿。从问题到答案的信息距离。在*第 13 届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集*, KDD '07, pages 874–883, New York, NY, USA, 2007. ACM。

[206] Le Zhao 和 Jamie Callan。术语必要性预测。在*第19届ACM信息与知识管理国际会议论文集*, CIKM '10，第 259–268 页，美国纽约州纽约市，2010 年。ACM。

[207] Le Zhao 和 Jamie Callan。选 In 的自动术语不匹配诊断*第 35 届国际会议*
择性查询扩展。 *论文集*
ACM SIGIR 信息检索研究与发展会议, SIGIR '12, pages 515–524, New York, NY, USA, 2012. ACM。