文章编号：1003-0077(2010)06-0075-10

搜索引擎查询推荐技术综述

李亚楠[1,2]，王 斌[1]，李锦涛[1]

（1. 中国科学院 计算技术研究所，北京 100190；
2. 中国科学院 研究生院，北京 100190）

摘 要：查询推荐技术，其用于找出与初始查询或关键词相关的其他查询或关键词，被广泛用于搜索引擎和广告检索系统中。作为当今搜索引擎的必备技术之一，查询推荐技术研究正受到越来越多的关注，近几年出现了很多验证查询推荐可用性及改进其算法的研究工作。为此，该文对查询推荐的发展过程、技术方法、评价体系等方面进行了归纳和总结，分析了查询推荐面临的挑战并讨论了现有解决方法及未来研究思路，希望能对相关研究人员有所帮助。

关键词：计算机应用；中文信息处理；综述；查询推荐；信息检索

中图分类号：TP391 **文献标识码**：A

A Survey of Query Suggestion in Search Engine

LI Yanan[1,2], WANG Bin[1], LI Jintao[1]

(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. Graduate University of Chinese Academy of Science, Beijing 100190, China)

Abstract: Query suggestion, i.e. generating related queries or keywords for an initial one, has been widely utilized in search engines and sponsored search systems. As one of the necessary techniques in search engines, query suggestion draws more and more attentions in NLP and IR community. In recent years, many studies have been done to validate the usefulness of query suggestion and to improve its effect. This paper introduces the state of the art in query suggestion, including its history, approaches and evaluation methods. The paper analyzes the challenges, discusses the possible solutions and suggests future works.

Key words: computer application; Chinese information processing; survey; query suggestion; information retrieval

1 引言

随着互联网的普及，搜索引擎已经成为人们获取信息的主要手段之一。目前搜索引擎采用的主要交互方式是：用户自主输入查询，检索系统根据输入的查询提供检索结果。但是，很多时候用户输入的查询并不能精确表达其搜索意图。一方面，用户输入的查询通常较短——平均只有两三个词[1-2]；另一方面，很多搜索引擎含有歧义[3]或意图模糊[4]；此外，很多时候，用户之所以要搜索就是因为对要检索话题知之甚少甚至毫无概念，这时候用户很难构造好查询。研究文献[5]表明，只有 25% 的查询能清晰表达用户的意图。

为了帮助用户更好地构造查询，搜索引擎普遍采用查询推荐技术，大家熟知的搜索结果页面中的“相关搜索”就是查询推荐的一个具体应用。查询推荐指发现或构造一组与原查询 Q 相关的查询{$Q1$, $Q2$,...}，这些相关查询可以通过修改原查询 Q 或整个替换 Q 来实现。例如，对查询“苹果 mp3”，可

收稿日期：2010-01-14 **定稿日期**：2010-08-30

基金项目：国家自然科学基金资助项目(60603094, 60873166)；北京市自然科学基金资助项目(4082030)；国家 973 计划资助项目(2007CB311103)；国家 863 计划资助项目(2006AA010105)；教育部科学技术研究重点资助项目(109028)

作者简介：李亚楠(1984—)，男，博士生，主要研究方向为信息检索；王斌(1972—)，联系作者，男，博士，副研究员，主要研究方向为信息检索；李锦涛(1962—)，男，博士，研究员，主要研究方向为多媒体技术；虚拟现实技术；普适计算技术等。

以通过修改查询词“mp3”来推荐查询“苹果 播放器”,也可以将整个查询替换为“ipod”。在不同文献中,针对这查询推荐的称呼可能会有所不同,例如,Query Suggestion[6], Term Suggestion[7], Query Recommendation[8], Query Substitution[9], Query Rewriting[10]。尽管在具体实现细节上,文献[11]认为上述方法间存在一些差异,但是从本文定义看,这些方法是相同的。

早在上世纪90年代,信息检索研究者就开展了一些查询推荐相关研究[12-15],一些早期搜索引擎(Altavista; Hotbot; Lycos)上也应用了初始的查询推荐技术。Rutgers大学的研究人员曾做了一系列关于检索系统的人机交互实验[13,16-17],这些研究显示[18],相对于完全透明、自动的查询扩展(Query Expansion)方式,用户更喜欢他们可以加以选择、控制的查询推荐方式。此外,很多研究人员又重新采用了可量化控制各种因素的用户评测实验[20-22],这表明:查询推荐技术在检索和浏览过程中的确能帮助用户进行更好的检索、节省搜索时间。如今,查询推荐技术已经成为了搜索引擎必备技术之一。

除了帮助搜索用户改进查询,查询推荐技术还可应用于广告检索系统中。搜索引擎公司通过广告检索系统将用户查询与广告关联起来。当用户搜索时,广告检索系统找出与其查询相关的广告,然后在搜索结果页面展示这些广告。然而,广告商通常只能提供很少一部分与其广告相关的查询。这使得广告只能展示给部分相关用户,因此影响广告商和搜索引擎的收益。广告检索系统通常采用查询推荐技术,根据广告商提供的广告相关查询找出更多相关查询或关键词,进而向用户推送更多相关广告以获取更多利润。例如,Google AdWords① 和百度推广②中的关键词推荐工具就属于此类应用。此外,查询推荐技术还应用于查询拼写检查、问答系统[23]、个性化搜索[24]等领域。

由于有着巨大的应用需求,查询推荐成为近几年的研究热点。从技术上看,查询推荐可以看作一个以搜索引擎查询或查询关键词为检索对象的信息检索问题。然而,不同于文章或网页,查询的自身特点使查询推荐面临诸多挑战:

- 首先,不同于文章或网页,查询通常只包含两三个词,缺乏充分的文本内容,传统信息检索模型不适合直接对其处理;
- 其次,很多查询信息稀疏。查询日志中多数查询出现次数很少,因此对这些查询处理时,可利用的相关属性信息很有限;
- 最后,用户查询复杂多样。查询日志中往往包含几千万甚至上亿不同的查询,即便同一查询,在不同用户心中也可能表示不同意图。此外,查询受时间、突发事件等因素影响。例如,情人节前后与查询“礼物”相关的查询是“玫瑰”、“巧克力”等,而在万圣节前后相关查询应该是“面具”等。

针对上述问题,学术界提出了多种方法用于查询推荐。例如,利用查询相关文档扩充查询以解决查询短的问题[25],挖掘查询间的间接隐含联系解决信息稀疏问题[50],根据搜索Session判定用户查询意图[7]等等。查询推荐作为自然语言处理的一种具体应用,很多技术方法都可以被其所采用。其方法主要有:基于特征匹配的传统信息检索模型[8,25],关联规则等传统数据挖掘方法[26],分类聚类等机器学习方法[9,27],基于查询统计特征分析的方法[28],图模型上的随机游走方法[10,29],以及一些利用经验规则融合各类方法的研究[7]。

查询推荐方法根据所依赖的数据不同可分为两类:基于文档的方法和基于日志的方法。1)基于文档的方法主要通过处理包含查询或查询词的文档来分析查询,从查询相关文档或人工编辑语料(如词典)中找出与输入查询相关的词或短语,然后利用这些相关词或短语构建推荐查询。2)基于日志的方法依靠分析搜索引擎查询日志寻找过去出现过的相似查询,然后向用户给予推荐。这两种方法各有利弊,基于日志的方法难以处理出现频率小的稀疏查询,基于文档的方法虽能处理稀疏查询,但是查找相关文档本身也是一个难题。近几年来,有研究提出将文档信息和日志信息结合起来进行查询推荐。

本文将系统阐述查询推荐技术的相关概念、理论方法及评价体系。后续内容组织如下:第2节总结了目前各类查询推荐技术方法;第3节讲述了查询推荐的评价体系;最后,第4节对本文进行了总结。

2 技术方法

搜索引擎查询推荐要解决的是一个应用问题,其可以看作是以查询或查询词作为处理对象的信息检索。这个检索问题的输入是用户初始查询,输出是一系列与初始查询相关的其他查询。然而,如前

① https://adwords.google.cn/

② http://e.baidu.com/pro/

面所述，搜索引擎查询不同于传统文本检索对象，其文本内容很少、信息稀疏，传统信息检索方法并不完全适用。与此同时，搜索引擎查询有其他一些特征可以利用。一方面，查询通常有一两个词或短语构成，而这些查询词和短语间的语义关系会体现在包含它们的文档或词典中；另一方面，查询日志中记录了查询的搜索者、时间、点击 URL 等属性，相关查询的这些属性也会表现一定的关联。查询推荐作为一个应用问题，需要考虑如何利用这些特征信息找出各查询间背后的联系，而不同的特征和数据又有一些不同的性质，需要相对应的合适方法对其处理。因此，查询推荐技术包括各种针对不同数据采用不同模型的方法。

本文根据查询推荐所依赖的数据和特征不同，对各种技术方法分别给予简要介绍。总的来说，根据所依赖的数据不同，查询推荐技术可分为两类：基于文档的方法和基于查询日志的方法。下面分别介绍这两类方法：

2.1　基于文档的方法

基于文档的方法主要通过处理包含查询的相关文档来找出与查询相关的词或短语，然后用这些词或短语构成要推荐的查询。基于文档的方法主要分为三类：全局文档集分析，局部文档集分析和分析人工编辑语料（如词典、维基百科等）。全局文档集分析利用所有文档分析文档中词与词的关系，找出与查询词关系紧密的其他词，进而构造推荐查询。局部文档集分析只分析部分文档来找出查询相关词，通常基于相关文档分析处理。随着信息技术和互联网的发展，现在有很多编辑良好的描述词与词之间关系的数据，例如 WordNet，HowNet，ODP 资源等。利用这些资源可以发现词与词间的语义联系，构造相关查询。

2.1.1　全局文档集分析

全局文档集分析利用所有文档分析文档中词与词的关系，找出与查询词关系紧密的其他词，进而构造推荐查询。最直观的想法，可以根据每个词 t_i $(1\leqslant i\leqslant n)$ 在各个文档 d_j $(1\leqslant j\leqslant m)$ 中出现的频率 w_{ij} 构造一个 $n\times m$ 矩阵 W，那么每个词 t_i 就可以用一个向量 $\vec{w}_i=(w_{i1},w_{i2},\cdots)$ 表示，这样词 t_i 和 t_j 间的相似度 $S(t_i,t_j)$ 转化为衡量向量 $\vec{w}_i$ 和 $\vec{w}_j$ 的相似度。最简单的方法可以通过 $\vec{w}_i$ 和 $\vec{w}_j$ 的内积计算，即 $S(t_i,t_j)=\sum_k(w_{ik}\times w_{jk})$。进一步地，可以采用 tf-idf 等信息检索思想定义权重 w_{ij}，将向量内积替换为更精确的相似度算法从而提高结果准确度[30]。由于文档数目 m 往往远远大于词数目 n，因此表示词的向量往往非常稀疏，这不利于相似度的计算。隐形语义索引（LSI）[31] 可以用于解决该问题，其对矩阵 W 进行 SVD 分解，重新表征标引项，降低维度。但是矩阵分解的计算复杂度非常高，对于真实大规模数据是难以承受的。

2.1.2　局部文档集分析

局部文档集分析只分析部分文档来找出查询相关词，通常基于相关文档分析处理。直觉上，与查询相关的词或短语将有更大可能出现在相关文档中，只分析相关文档就可以找出查询相关词。然而相关文档难以获得，常用的方法是假设检索返回的排名靠前文档是相关的。由于这些文档并非真正的相关文档，因此也常被称为伪相关文档。

有很多方法利用伪相关文档检索查询相关词。首先，伪相关文档中出现的高频非停用词可以作为查询相关词[32]，由于伪相关文档中都包含查询词，这种方法其实就是找出在查询词出现的条件下出现概率最高的那些词语作为相关词。另一个知名方法是 Xu 和 Croft[15] 提出的 LCA（Local Context Analysis）。LCA 计算伪相关文档中每个词与整个查询而非某个查询词的关系紧密程度，因此拥有更高的准确度。但是 LCA 效率较低，查询推荐需要实时处理，LCA 计算复杂度偏高。

上述方法几乎可以处理各种查询，即便对于罕见查询，只要能返回搜索结果，也同样可以处理。但是这些方法的前提假设是能找到查询相关文档，能否找出查询相关文档本身就是一个困难问题。伪相关文档毕竟不是真正的相关文档，它们会引入不相关文档而降低准确度。

2.1.3　基于人工编辑语料的方法

随着信息技术和互联网的发展，现在有很多编辑良好的描述词与词之间关系的数据，利用这些资源可以发现词与词间的语义联系，构造相关查询。这类方法通过利用词典（例如 WordNet[33]）或其他人工编辑好的数据（如 Wikipedia[34-35]、Open Directory Project[36]）查找相关查询词或短语。这类方法的结果往往比较准确，但是难以处理那些尚未编辑的新出现查询词，而新词却在用户搜索中占很大比例。

尽管基于文档的方法可以找出与当前查询相关的一系列词或短语，但是要完成查询推荐还需要将

这些相关词或短语组合成合适的搜索引擎查询。搜索引擎查询不同于人类自然语言中的问题,它有其自身的特点,如何组合成合适的查询本身也是一个难题。另一方面,搜索引擎查询日志中记录了用户构造的各种真实查询,通过分析查询日志更容易找出并推荐合适的查询。

2.2 基于日志的方法

当用户搜索时,搜索引擎通常会将用户的行为记录下来,这些记录数据构成了搜索引擎查询日志。查询日志中每条记录对应于用户的一次行为,例如,开启一次新的查询、点击一个搜索结果链接、翻看新一页的搜索结果等等。查询日志中记录了用户点击文档和其他用户搜索行为信息,这些重要信息是基于文档的方法难以提供的。利用查询日志中记录的各种信息,可以挖掘出不同查询间的联系。现有方法主要利用查询间的共有属性来挖掘查询间的联系紧密程度,这些属性特征主要包括:查询共同出现在同一搜索过程(Session)的次数、查询共有的相同或相似的点击URL、日志中不同查询间的文本相似度、两查询出现频率随时间分布的相关性。根据所依赖的特征不同,这里将基于日志的方法分为四类方法:基于Session的方法、基于点击URL的方法、基于文本相似度的方法和基于时间分布的方法。

2.2.1 基于Session的方法

用户在搜索过程中为了同一个检索目标所做的一系列检索行为构成一个Session。很多时候,一个Session中会包含多个查询,这表明用户对Session中初始查询的检索结果不满意,后来他有重新构造一个或多个表达同一搜索意图的查询。

用户搜索Session中的信息可以从多个方面帮助查询推荐。首先,可以用之前用户的搜索经验帮助后来的用户,直接向当前用户推荐之前用户最终找到正确答案所用的查询。基于这一思想,Cucerzan和White[6]提出一套规则判别Session中的最终结果网页,进而向用户推荐能直接返回最终结果网页的查询。其次,经常出现在同一Session中的两个查询很有可能是语义相似的,因为它们多次表达同一查询意图。由此,可以根据Session中查询的共现信息利用关联规则[26]、互信息[9]、相似度算法[7,37,41]度量查询间相关性。最后,Session相对于单个查询,提供了更多有助于明确查询意图的信息,根据整个Session而非单个查询进行推荐将会更加准确[7,38]。

基于Session的方法需要首先将查询日志划分成多个Session,而Session划分好坏会影响查询推荐的准确率。传统方法根据同一用户两个相邻查询间的时间间隔判断这两个查询是否处于同一Session中,如果时间间隔大于一个设定的阈值,则在这两个查询间进行Session切分。单纯依靠时间间隔进行Session划分并不十分精确,近年来提出了一些更有效的Session划分方法,相关工作可参见文献[39]。

2.2.2 基于点击URL的方法

查询日志中记录了每次查询时用户点击的URL,这些URL可用来挖掘查询间的关系紧密程度。如果两个查询所对应的点击URL很多都是相同或相似的,那么这两个查询就有很大的相关性。根据此思想,很多查询推荐算法被提出。

最开始的工作主要利用相同点击URL衡量查询相关性。其中,王继民和彭波[40]提出一种基于查询共有相同点击URL数的查询推荐的方法。进一步地,查询 Q_i 可以表示成由其所对应URL构成的向量(U_{i1},U_{i2},…,U_{im}),然后应用向量空间检索模型[42]计算不同查询间的相似度。其中 U_{ia} 表示第 a 个URL的权重,U_{ia} 可以用最简单的第 a 个URL在查询 Q_i 出现的次数表示,也可以根据tf-idf思想做适当改进。

查询日志统计分析[43]显示一次查询平均只有几次点击,表示查询的向量往往非常稀疏。实际上,对一个查询,用户往往只点击前1,2页中的某几个结果,很多相似查询没有相同的点击URL。为了应对这一问题,不同研究者提出了不同的方法。首先,可以在计算查询相似度时,把相似URL也考虑进来,拥有内容相似点击URL的查询也应该是相似的。既然可以根据点击URL算出查询间的相似度,反过来依据URL所对应的查询同样可以求得URL间相似度,这样不断迭代就可以同时得到查询间和URL间的更精确的相似度。基于这种思想,Antonellis等人[10]提出利用一种改进的SimRank相似度算法度量查询相关性。另一方面,如果能把查询或URL的空间维度降低,就能避免数据稀疏的问题。这方面的相关研究有基于查询聚类的方法[27,38]和基于矩阵分解的方法[44-45]。但是,上述方法在提高准确度的同时也加大了算法计算复杂度。

2.2.3 基于文本相似度的方法

查询也是由词和短语构成的对象,传统的文本信息检索模型或文本编辑距离同样可以用来度量查

询相似度。但是搜索引擎查询通常很短，平均长度不到三个词[1-2]，直接对日志中查询计算相似度的效果并不好。例如，查询“电脑”和“计算机”相关但是却没有相同查询词，查询“汽车引擎”和“搜索引擎”不相关却有 50％的查询词重叠。

如果能对日志中查询的文本内容进行扩充，就能避免上述问题。为此，很多研究用伪相关文档构造表示查询的文档 QD，进而利用 QD 间的相似度计算其所对应查询的相似度。不同方法中伪相关文档的定义不同，Sahami[25]提出用搜索引擎返回的排名靠前的 n 个结果作为伪相关文档，Baeza-Yates 等人[8]用有用户点击的结果作为伪相关文档并将点击频率因素考虑进来。直观来看，用户点击过的文档比直接用所有排名靠前文档似乎更相关一些，但实际上搜索引擎排序对用户点击同样会产生很大影响[46]，用户点击频率排序经常跟搜索引擎排序是一致的。采用一些更可靠的分析用户点击的模型[47]可能有助于提高该类方法。

2.2.4　基于时间分布的方法

有研究提出相似查询的搜索频率在时间分布上应该是相似的，例如查询“沃尔玛”和“山姆会员店”在不同时间段的分布都是比较均匀的，而查询“北京奥运会”和“中国 金牌榜”这样的查询频率分布在同一特定时间有明显的尖峰。此外，查询推荐也应该考虑查询频率在时间上的分布情况，有的查询有其重要时间段，在其重要时间段的推荐将更有效。例如，查询“巧克力”在 2 月 14 日是个比较好的推荐查询，因为巧克力作为礼物在情人节很流行。

查询频率在时间上的分布可以用查询时间分布向量 $f_q=(f_{q1},f_{q2},\cdots,f_{ql})$ 表示，其中 f_{qi} 表示查询 q 在第 i 个单位时间段内的搜索频率。为了度量不同查询在时间分布上的相关性，Chien 和 Immorlica[28]提出用查询时间分布向量的皮尔森(pearson)相关度表征查询相似度。基于 Chien 和 Immorlica 的工作，Zhang 等人[48]提出一种考虑重要时间段的方法。此外，如同 Web 搜索中的 PageRank 一样，查询在日志中出现频率也常作为一种静态排序因子用于查询推荐[49]。查询频率的时间分布特征的确是查询推荐中的一个重要特征，但是只依靠此类特征判断查询相关性是不充分的，这类方法可以作为其他方法的一种补充应用在查询推荐系统中。

基于日志的方法根据用户搜索历史推荐查询，相对于基于文档的方法其构造的查询更符合用户查询的特点。但是查询在日志中的出现频率呈指数分布，大多数查询在日志中出现次数不多[43]。这使得基于日志的方法时面临更严重的数据稀疏问题，当前方法在处理流行查询时可以取得不错的效果，但是对付出现次数较少的查询准确度较差。

2.3　小结

各种查询挖掘方法各有优缺点，表 1 对此进行

表 1　查询处理相关工作优缺点总结

方　法		优　点	缺　点
基于文档的方法	全局文档分析	可以直接利用大量语料数据和传统信息检索模型对各类词语进行分析	计算量大，现实应用难以承受。受自然语言处理技术限制，效果不是很好。
	局部文档分析	可对各类查询或查询词进行处理，只分析查询相关文档，相对于全局文档分析，降低计算开销。	如何准确获取查询相关文档成为另一个引入的难题，计算开销仍然较大。
	利用语义资源	直接利用人工编辑好的各种语义资源，处理简单快捷，结果准确	难以处理网络实时出现的各种新词汇，而有很大比例查询是对新事物的搜索。
基于日志的方法	基于 Session 信息	Session 信息体现了用户对各种查询的理解，反映了查询意图和查询间的语义相关性。	需要先对查询日志进行准确的 Session 划分，Session 比查询更少，直接依据统计信息依然存在信息稀疏问题。
	基于点击 URL	综合利用了查询、用户和文档间的关系信息，以相关点击文档信息扩充查询内容，便于现有技术处理	依赖于搜索引擎检索效果，而且用户点击结果中包含偏向性和噪音，同时很多查询面临缺乏点击的信息稀疏问题。
	基于文本相似度	可以直接利用现有信息检索模型对查询进行处理	由于查询自身文本信息太少，数据稀疏且经常有多义查询词，结果不准确。
	基于时间信息	考虑了不同查询的时间分布差异	只能作为其他结果的补充，本身难以取得实用效果。

了总结。采用基于文档的方法分析查询,可以利用互联网上各类大规模数据和一些人工编辑好的资源。然而普通文档中的文本内容与一般查询不同,从普通文档中挖掘查询信息受到很多制约。搜索引擎查询有其自身的特点,搜索引擎查询日志中记录了用户构造的各种真实查询,相对于基于文档的方法更适合分析挖掘查询性质。但是查询日志中大多数查询出现次数不多,面临严重的数据稀疏问题,当前方法在处理流行查询时可以取得不错的效果,但是对于出现次数较少的查询准确度较差。

从本质看,查询推荐是一种以查询作为处理对象的信息检索应用问题。然而,不同于传统文本信息检索,查询推荐面临文本内容少、复杂多样、信息稀疏等挑战。直接应用传统信息检索模型并不能很好解决这些问题,为此学术界提出多种方法应对这些挑战。下面分别针就上述问题使用不同方法做一个小结。

(1) 搜索引擎查询平均长度一般只有两三个词,缺乏传统信息检索模型所需的特征信息。针对该问题有两类解决方法。

• 第一类,利用查询相关文档扩充查询文本内容,将查询间的相似度计算就转化为查询相关文档集间的文本相似度计算[25]。查询相关文档可以定义为搜索引擎返回的前 N 篇文档,或者是查询日志中的用户点击结果。

• 第二类,利用文档或日志中的相关信息将查询或查询词映射到另一个特征空间[8],这样不同但相似的查询或查询词间的匹配度也不为零。

(2) 为了应对信息稀疏问题,近几年学术界提出了不少解决办法,这些方法基本上可以归为三类。

• 第一类,将查询映射到一个目录中,利用目录的树状结构度量查询间相似度[37];这个目录可以采用 WordNet、ODP 等已经编辑好的目录,也可以根据具体问题重新构建。

• 第二类,通过聚类或 LSI 等方法将信息稀疏的高维特征空间映射到低维空间中[38, 45]。

• 第三类,将查询组织到关系图中,其中节点表示查询,边表示查询间的关系。在查询关系图中,通过其他查询间接关联的查询之间也有一定相关性。利用查询间关系信息可以计算间接相关查询间的相似度[29, 50-51]。

(3) 搜索引擎查询复杂多样且经常意图不清。为了辨识用户搜索意图,利用整个 Session 信息而非单个查询信息是经常采用的方法[7, 38],用户的查询历史记录也可用于帮助明确用户意图。另一方面,查询受时间、地域等因素影响。对一些查询在推荐时考虑更多特征信息将取得更好的效果[28, 48]。

综上所述,查询推荐是一项应用驱动的技术,为了在具体应用环境中取得更好的效果,各种数据和方法都可以采用。各类数据和方法都与各自的优缺点,真实查询推荐系统中应该综合利用各种数据和方法。有研究提出综合两种方法进行查询推荐,现有方法[49, 52]主要将基于文档的查询相似度和基于日志的查询相似度做一个线性加权,基于文档的相似度权重和基于日志的相似度权重根据经验值设定。总的来说,现有的综合文档和日志的方法还比较简单,也缺乏相应的理论基础。另一方面,由于真实搜索引擎的搜索量巨大,查询推荐的可扩展性和实时性也是非常重要的。能更好解决查询稀疏性、多样性等问题的方法也常常更加复杂,因此这类算法经常分为在线运算和离线运算两部分。一个实际运行的查询推荐系统需要在推荐效果和效率方面做好平衡。

3 评价方式

查询推荐的质量与其应用目标相关,在不同的应用环境中,推荐查询的目的会有所不同。在搜索引擎中,查询推荐的目的在于帮助用户构造更合适查询,从而更快地找到所需信息。一个好的推荐查询不仅要与初始查询相关还应该贴合用户查询意图并有很好的检索效果[21],这样才能帮助用户。在广告检索系统中,查询推荐的目的在于提高搜索引擎的广告收益,这也不仅要求语义相关还应该考虑查询价格、广告商预算等多种因素[53]。针对一个具体的应用,查询推荐的质量评价需要考虑多方面因素,而且很多时候要做到完全准确客观的评价是很困难的。

如第三部分"相关概念"所述,本文的查询推荐指根据初始查询找出或构造一组相关查询的技术。因此,这里所述的评价体系只考虑推荐出的查询是否与原查询相关,不考虑其他因素。实际上,绝大多数查询推荐研究工作都采用这种方法进行评价。下面分别从测试集构建和评价指标设计两个方面介绍查询推荐评价体系。

3.1 测试集

要评价一种查询推荐技术的好坏,首先需要构

建一个用于评测推荐结果相关与否的测试集，测试集包括查询推荐方法依赖的数据和标示相关性的答案。这里，查询推荐方法依赖的数据可以是基于文档的方法所依赖的文档集，或者是基于日志的方法所用的查询日志；标示答案的相关性只考虑推荐与原查询是否语义相关，不考虑推荐查询能否使搜索引擎返回与用户意图更相关结果等其他因素。下面分别介绍已有的可用数据和相关性答案的构建方法。

根据已知知识，现在尚未有公认的查询推荐公共评价语料。因此，尽管现在已经有大量查询推荐相关工作发表，但仍然缺乏不同方法间的客观比较。对于基于文档的方法，所需的文档集可以用传统的方法收集或直接使用 TREC 语料、维基百科等现成数据。但是对于现在主要研究的基于日志的方法，由于查询日志涉及用户隐私等原因，除了搜索引擎公司（如微软、雅虎、谷歌）的实验室，多数大学和研究机构的实验室难以获取真实查询日志。已知的公开或曾公开过的查询日志有：搜狗公司公开的 2008 年部分中文查询日志①，AOL 曾公开过的 2006 年部分英文查询日志②。这些日志包括用户 ID、查询词、点击 URL、点击 URL 在返回结果中的排名等基本信息。

在查询推荐实验中，有两类构建相关性答案的方法：人工评价和自动评价。人工评价指：由人工标注每个推荐查询与原查询是否语义相关。自动评价则不需要人工标注。下面分别简要介绍这两类方法：

• 采用人工评价方法时，为了减少标注工作量，通常采用类似于 pooling [54] 的评测集构建方法。即对每个评测查询，取各种对比方法的前 n 个推荐结果构成一个集合，然后请人评价集合中每个结果的相关度。相关度的标注可以只是“相关”、“不相关”两类，也可以根据某些标准[9, 45]分成多个等级。

• 自动评价利用其他资源判定查询间的相关度，一般利用查询日志、人工编辑目录、WordNet、维基百科等数据构建相关性答案。文献[10]用查询日志做自动评价，将查询日志分为两部分：一部分做训练集，一部分做测试集。对测试集中出现的查询 Q，找出用户在同一 Session 中使用查询 Q 后又构造的其他查询 $\{Q_1, Q_2, \cdots, Q_3\}$。对一种待评价方法 R，如果 R 的推荐结果中包括 $\{Q_1, Q_2, \cdots, Q_3\}$ 中的查询，则认为推荐成功。有些研究利用人工编辑目录（ODP[45, 55]，Google Directory[56]）评价推荐结果的相关程度：对于两个要度量其相关度的查询 Q_1 和 Q_2，它们的相似度通过它们的共享目录类别个数度量，Q_1 和 Q_2 共同属于的类别越多，它们就越相关。

3.2　评价指标

查询推荐是以查询作为处理对象的信息检索问题，因此一般查询推荐系统的评价都采用了文本检索系统里面的指标。常用的评价指标有：

• 召回率（*Recall*）和正确率（*Precision*）。对一个查询推荐系统的推荐结果：

$$Precision = \frac{\#\text{推荐出的相关查询数}}{\#\text{推荐出的所有查询个数}},$$

$$Recall = \frac{\#\text{推荐出的相关查询数}}{\#\text{各种方法推荐出的相关查询数}},$$

其中 $P@N$（Precision at N）是一个经常被采用的指标，即前 N 个推荐结果中相关查询所占的比例。

• AP 和 MAP。AP（Average Precision）是对单个查询在不同召回率点上的正确率进行平均得到的平均正确率。对所有测试查询的取平均值，则得到 MAP（Mean AP）。令 $R(k)$ 表示第 k 个相关结果所在位置，N 表示测试查询的个数，则：

$$AP = \sum_k P@R(k),\ MAP = \frac{1}{N}\sum_{i=1}^{N} AP_i$$

• DCG 和 $NDCG$。有时候推荐查询的相关度有多个等级，并非只有“相关”和“不相关”两类，另外用户往往更关注于排名靠前的结果。这时需要更准确的评价指标，相关等级越高的结果越多越好，相关等级越高的结果排名越靠前越好。DCG（Discounted Cumulative Gain）和 $NDCG$（Normalized DCG）就是这样一类评价指标。假设＜V1, V2, …, Vn＞是查询 q 的返回结果，令 $R(k)$ 表示 Vk 的得分，则在第 k 个位置上的 DCG 表示为：

$$DCG@k = \sum_{i=1}^{k} \frac{1}{\log_2(1+i)}(2^{R(i)} - 1)$$

$NDCG$ 是归一化后的 DCG。对每个查询，计算一个标准答案的 $IDCG$（Ideal DCG），用返回结果的 DCG 除以标准 DCG，就是 $NDCG$。

$$NDCG@k = \frac{DCG@k}{IDCG@k}$$

① http://www.sogou.com/labs/dl/q.html

② http://gregsadetsky.com/aol-data/

- 覆盖率(Coverage)

$$Coverage=\frac{\#返回正确答案的查询数}{\#测试查询总数}$$

4 结论

查询推荐及其相关技术是搜索引擎的必备技术之一。一方面,查询推荐可以帮助用户构造更准确查询,节省搜索时间。另一方面,查询推荐技术也帮助搜索引擎广告系统匹配更多更准确的广告,增加搜索引擎利润已经广告商的潜在收益。此外,查询推荐相关技术还被广泛应用于查询优化、拼写错误检查、查询扩展、个性化搜索等领域。

本文阐述了查询推荐技术的相关概念及研究和发展过程,对查询推荐的主要方法和关键技术进行综述,介绍了目前学术研究中采用的方法和评价体系。本文将查询推荐技术按照所依赖的数据不同分为基于文档的方法和基于日志的方法,并说明了查询推荐中面临的挑战和现有的各种解决思路。

相对于传统文本检索,查询推荐还是一个相对较新的研究领域,还有很多地方有待进一步探讨和研究。首先,查询推荐尚缺乏统一的评价标准和评测语料,各类技术方法难以做客观的比较。其次,各类基于不同数据和算法的查询推荐方法各有优缺点,需要一种有效的理论方法将它们融入一个统一的系统中。最后,中文搜索引擎查询推荐也有待完善[57]。不同于其他语言,中文搜索引擎查询有其自身特点,例如,多数中文查询是一个不包含空格的完整词或短语而不是多个查询词的集合[43]。尽管有一些利用中文查询日志所作的查询推荐研究[41,58],但是这些方法基本与语言无关,结合中文特点的查询推荐仍然亟待研究。

参考文献

[1] A. Spink, B. J. Jansen. A Study of Web Search Trends [J]. Webology. 2004, 1(2), Article 4. Available at: http://www.webology.ir/2004/v1n2/a4.html

[2] 余慧佳,刘奕群,张敏,茹立云,马少平. 于大规模日志分析的网络搜索引擎用户行为研究[C]//第三届学生计算语言学研讨会(SWCL2006). 2006.

[3] R. Song, Z. Luo, J.-R. Wen, Y. Yu, and H.-W. Hon. Identifying ambiguous queries in web search [C]//WWW'07. New York: ACM, 2007: 1169-1170.

[4] Andre Broder. A taxonomy of web search [J]. SIGIR Forum 36(2), 2002.

[5] M. Strohmaier, M. Kröll, C. Körner. Intentional Query Suggestion: Making User Goals More Explicit During Search [C]//Proceedings of the 2009 workshop on Web Search Click Data. 2009:68-74.

[6] Silviu Cucerzan, Ryen W. White, Query suggestion based on user landing pages [C]//Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. New York: ACM, July 23-27, 2007: 875-876.

[7] C. K. Huang, L. F. Chien,, & Y. J. Oyang. Relevant term suggestion in interactive Web search based on contextual information in query session logs [J]. Journal of the American Society for Information Science and Technology. 2003, 54(7): 638-649.

[8] R. Baeza-Yates, C. Hurtado, and M. Mendoza. Query Recommendation Using Query Logs in Search Engines [C]//EDBT 2004 Wordshops. LNCS 2368, 2004:588-596.

[9] R. Jones, B. Rey, O. Madani, W. Greiner. Generating Query Substitutions [C]//WWW2006 New York: ACM, 2006: 387-396.

[10] I. Antonellis, H. Garcia-Molina and C. C. Chang, SimRank++: query rewriting through link analysis of the click graph[C]//Proceedings of VLDB. 4 Dec 2008: 408-421.

[11] D. Kelly, K. Gyllstrom, E. W. Bailey. A comparison of query and term suggestion features for interactive searching [C]//SIGIR '09. New York: ACM, 2009:371-378.

[12] E. Eftheimiadis. Query expansion [J]. Annual Review of Information Science Technology. 1996, 31: 121-187.

[13] J. Koenemann. Relevance feedback: usage, usability, utility [D]. Ph. D. Dissertation, Rutgers University, Dept. of Psychology. 1996.

[14] A. Spink, R. M. Losee. Feedback in information retrieval [J]. Annual Review of Information Sciences Technology. 1996, 31: 33-78.

[15] J. Xu, W. B. Croft. Query expansion using local and global document analysis [C]//Proceedings of 19th ACM-SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1996: 4-11.

[16] N. J. Belkin. Intelligent information retrieval: Whose intelligence? Herausforderungen an die Informationswissenschaft [C]//Proceedings des 5. Internationalen Sypmosiums für Informationswissenschaft (ISI '96). J. Krause, M. Herfurth, and J. Marx, Eds. Universitatsverlag Konstanz, 1996: 25-31.

[17] N. J. Belkin, C. Cool, J. Head, J. Jeng, D. Kelly, S. J. Lin, Lobash, L. Park, P. Savage-Knepshield, and C. Sikora. Relevance feedback versus Local Context Analysis as term suggestion devices[C]//Proceedings of the Eighth Text Retrieval Conference TREC8. Washington, D. C., 2000.

[18] N. J. Belkin. Helping people find what they don't know [J]. Communications of the ACM. 2000, 43(8): 58-61.

[19] Peter Anick. Using Terminological Feedback for Web Search Refinement - A log-based Study. In SIGIR2003 [C]//New York: ACM, 2003:88-95.

[20] A. Feuer, S. Savev, J. A. Asiam. Evaluation of phrasal query suggestions [C]//CIKM'07. New York: ACM, 2007: 841-847.

[21] R. W. White, M. Bilenko, S. Cucerzan. Studying the use of popular destinations to enhance web search interaction [C]//Proceedings of SIGIR '07. New York: ACM, 2007: 159-166.

[22] P. Vakkari. Changes in search tactics and relevance judgments when preparing a research proposal: A summary of the findings of a longitudinal study [J]. Information Retrieval. 2004, 4(3-4), 295-310.

[23] J. Jeon, W. B. Croft, and J. H. Lee. Finding similar questions in large question and answer archives [C]//CIKM '05. New York: ACM. 2005: 84-90.

[24] P. A. Chirita, C. S. Firan, and W. Nejdl. Personalized query expansion for the web[C]//SIGIR '07. New York: ACM, 2007:7-14.

[25] M. Sahami and T. D. Heilman, A web-based kernel function for measuring the similarity of short text snippets [C]//Proceedings of the 15th international conference on World Wide Web. New York: ACM, 2006: 377-386.

[26] B. M. Fonseca, Golghe, P. B., Moura, E. S. d., and Ziviani, N. Discovering Search Engine Related Queries Using Association Rules [J]. J. Web Eng., 2004, 2(4):215-227.

[27] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log [C]//SIGKDD '00. New York: ACM, 2000: 407-416.

[28] S. Chien and N. Immorlica. Semantic similarity between search engine queries using temporal correlation [C]//Proceedings of WWW-05, International Conference on the World Wide Web. New York: ACM, 2005: 798-799.

[29] Paolo Boldi, Francesco Bonchi, Carlos Castillo, Debora Donato, Sebastiano Vigna, Query suggestions using query-flow graphs[C]//Proceedings of the 2009 workshop on Web Search Click Data. Barcelona, Spain, February 09-09, 2009: 56-63.

[30] Y. Qiu and H. P. Frei. Concept based query expansion [C]//SIGIR 1993. New York: ACM, 1993: 160-169.

[31] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis [J]. JASIS, 41(6): 391-407. January 1999.

[32] R. W. White and G. Marchionini. Examing the effectiveness of real-time query expansion [J]. Inf. Process. Manage. 2007, 43(3): 685-704.

[33] E. Stoica, M. Hearst, and M. Richardson. Automating creation of hierarchical faceted metadata structures [C]//NAACL/HLT 2007. 2007.

[34] J.-R. Shieh, Y.-H. Hsieh, T.-Ch. Su, Ch.-Y. Lin, J.-L. Wu. Building term suggestion relational graphs from collective intelligence[C]//WWW 2009. New York: ACM, 2009:1091-1092.

[35] Yang Xu, Gareth J. F. Jones, Bin Wang. Query dependent pseudo-relevance feedback based on wikipedia [C]//SIGIR 2009. New York: ACM, 2009:59-66.

[36] Y. Chen, G.-R. Xue, and Y. Yu. Advertising keyword suggestion based on concept hierarchy [C]// Proceedings of the international conference on Web search and web data mining. New York: ACM, 2008: 251-260.

[37] Eric C. Jensen, Steven M. Beitzel, Abdur Chowdhury, Ophir Frieder, Query Phrase Suggestion from Topically Tagged Session Logs[C]//Proceedings of the 7th International Conference on Flexible Query Answering Systems (FQAS 2006), Milan, Italy, June 2006: 185-196.

[38] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen and H. Li, Context-aware query suggestion by mining click-through and session data[C]//Proceeding of the 14th ACM SIGKDD. New York: ACM, 2008: 875-883.

[39] D. Gayo-Avello. A survey on session detection methods in query logs and a proposal for future evaluation [J]. Information Science: an International Journal. Elsevier Science Inc. May, 2009, 179(12): 1822-1843.

[40] 王继民,彭波. 搜索引擎用户点击行为分析 [J]. 情报学报. 2006,25(2).

[41] 李亚楠,许晟,王斌. 基于加权 SimRank 的中文查询推荐研究[J]. 中文信息学报. 2010, 24(3): 4-10.

[42] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval [M]. New York: ACM, and England: Addison-Wesley, 1999.

[43] Yanan Li, Sen Zhang, Bin Wang, Jintao Li, Characteristics of Chinese Web Searching: A Large-Scale Analysis of Chinese Query Logs [J]. Journal of Com-

putational Information Systems, 2008, 4(3): 1127-1136.

[44] D. Gleich, L. Zhukov. SVD based term suggestion and ranking system [C]//ICDM'04. IEEE, 2004.

[45] H Ma, H Yang, I King, M R Lyu. Learning latent semantic relations from clickthrough data for query suggestion [C]//CIKM' 08. New York: ACM, 2008:709-718.

[46] Thorsten Joachims, Laura Granka, Bing Pan, Accurately Interpreting Clickthrough Data as Implicit Feedback [C]//SIGIR'05. New York: ACM, 2005.

[47] Shihao Ji, Ke Zhou, Ciya Liao, Zhaohui Zheng, Gui Rong Xue, O. Chapelle, Gordon Sun, Hongyuan Zha. Global ranking by exploiting user clicks[C]//SIGIR'09. New York: ACM, 2009: 35-42.

[48] W. Zhang, J. Yan, Sh.-Ch. Yan, N. Liu, Zh. Chen. Temporal query substitution for ad search[C]//SIGIR'09. New York: ACM, 2009: 798-799.

[49] J.-M. Yang, R. Cai, F. Jing, Sh. Wang, L. Zhang, W.-Y. Ma. Search-based query suggestion [C]//CIKM'08. New York: ACM, 2008.

[50] Yanan Li, Bin Wang, Sheng Xu, Peng Li, Jintao Li. QueryTrans: Finding Similar Queries Based on Query Trace Graph[C]//Proceedings of the 2009 IEEE / WIC / ACM Joint Conference on Web Intelligence. September 15th-18th, 2009:260-263.

[51] Q. Mei, D. Zhou, and K. Church. Query suggestion using hitting time [C]//CIKM '08. New York: ACM, 2008:469-478.

[52] Zh.-Y. Zhang, O. Nasraoui. Mining Search Engine Query Logs for Query Recommendation [C]//WWW'06//New York: ACM, 2006: 1039-1040.

[53] Azarakhsh Malekian, Chi-Chao Chang, Ravi Kumar, Grant Wang. Optimizing query rewrites for keyword-based advertising [C]//EC '08: Proceedings of the 9th ACM conference on Electronic commerce. 2008: 10-19.

[54] E. M. Voorhees. The philosophy of information retrieval evaluation[C]//In Proceedings of the Second Workshop of the Cross-Language Evaluation Forum, (CLEF 2001), 2001: 355-370.

[55] R. Baeza-Yates and A. Tiberi. Extracting semantic relations from query logs[C]//KDD'07. New York: ACM, 2007:76-85.

[56] Hongbo Deng, Irwin King, Micheal R. lyu. Entropy-biased Models for Query Representation on Click Graph [C]//SIGIR' 09. New York: ACM, 2009: 339-346.

[57] 姜文彬. 搜索引擎相关关键词推荐的对比研究 [J]. 现代图书情报技术. 2009, 180:35-40.

[58] Xiaoyan Shen, Bo Cheng, Junliang Chen, Xiangwu Meng. An Effective Method for Chinese Related Queries Recommendation [C]//Proceedings of the 2008 Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing table of contents. 2008: 381-386.