搜索结果多样化研究综述

窦志成 秦绪博 文继荣

（中国人民大学信息学院 北京 100086）

（大数据管理与分析方法研究北京市重点实验室 北京 100086）

摘 要 基于传统信息检索技术的搜索引擎一般按照用户提交的查询获得一系列搜索结果，经过相关性排序之后返回给用户.相关研究表明，用户提交给搜索引擎的查询通常是短查询，且经常具有歧义性和宽泛性.另外，不同的用户在使用同一个查询词时，其查询需求也往往是不同的.此时若搜索引擎只进行相关性排序，则会面临搜索结果冗余性过大、无法有效覆盖用户意图的问题，对用户体验产生负面影响.为了满足用户需求，现有的搜索引擎在搜索排序过程中必须有效应对查询歧义性问题.搜索结果多样化是解决这个问题的方法之一，其目标是尽量提升搜索结果的多样性，让搜索结果尽可能多地覆盖不同用户的查询意图，以确保至少有一个结果可以命中实际的用户需求.和传统的搜索排序模型仅考虑文档和查询的相关性不同，多样化排序模型还需要进一步考虑文档的新颖性或者结果集对不同用户意图的覆盖度.现有的多样化算法，根据是否显式地衡量查询包含潜在用户意图所对应的子话题，可以分为隐式多样化模型和显式多样化模型.隐式多样化模型通常只考虑文档之间的相似性，即认为一份结果文档集中的每一个结果文档，彼此之间的不相似度越高，多样化程度越高；而显式多样化模型则显式地考虑不同查询所对应的不同用户意图（以子话题表示），通过衡量结果文档集对不同子话题覆盖的广度来衡量整体的多样化程度.根据多样化特征是人工指定的还是通过监督式学习自动获取的，可以分为启发式模型和学习式模型（也称为非监督式模型和监督式模型）.启发式模型使用人工指定的文档相似度或子话题覆盖程度等相关特征来判断每一个文档对多样化的贡献，而学习式模型则可以通过监督式学习，自动地学习到最优化的多样化指标.这两种分类方法相互正交，理论上显式多样化模型优于隐式多样化模型，监督式模型优于非监督式模型，但实际上模型的具体表现也可能受到训练数据总量、子话题质量等其他因素影响.另外，最近学界对强化学习的研究也对信息检索领域产生了一定的影响，由于多样化排序过程也可以被视为一个依次选择每一个候选文档的决策过程，因此也有学者将强化学习引入了搜索结果多样化领域.本文介绍了搜索结果多样化的定义，并对现有多样化算法进行分类整理，详细介绍了每个类别中的代表性方法.我们还进一步介绍了搜索结果多样性评价方法，并给出了一系列方法的对照实验结果.最后，我们对搜索结果多样化技术研究的方向进行了展望.

关键词 信息检索；用户意图；多样化；个性化；强化学习

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2019.02591

A Survey on Search Result Diversification

DOU Zhi-Cheng QIN Xu-Bo WEN Ji-Rong

(*School of Information, Renmin University of China, Beijing* 100086)

(*Beijing Key Laboratory of Big Data Management and Analysis Method, Beijing* 100086)

Abstract Search Engines based on traditional Information Retrieval (IR) technology usually use the user-issued queries to retrieve a set of search results, and return the results after relevance ranking. Research shows that queries issued by users are mostly short queries which are ambiguous or broad when they come to specifying users' information need. Different users who issue the same query may have different information needs in their mind. If search engines only rank the

收稿日期：2018-07-02；在线出版日期：2019-04-04.本课题得到国家重点研发计划（2018YFC0830703）和国家自然科学基金（61872370）资助.
窦志成，博士，教授，中国计算机学会（CCF）会员，主要研究领域为信息检索、数据挖掘、大数据等. E-mail：dou@ruc.edu.cn. **秦绪博**，博士研究生，主要研究方向为信息检索、数据挖掘等. **文继荣**，博士，教授，国家千人计划专家，中国计算机学会（CCF）会员，主要研究领域为信息检索、数据库、大数据分析、数据挖掘.

results with relevance features, the result set may be too redundant and unable to cover the real user intents, which may influence the user's experience. The present search engines have to tackle the query ambiguity problem in order to satisfy the users' needs. Search result diversification is one of the approaches to solve this problem. Search result diversification aims to retrieve diverse results to satisfy as many different information needs as possible to make sure there will be at least one result hitting the real user intent. Compared with the traditional ranking models which only consider the relevance between the query and the documents, search result diversification models further take result diversity into consideration. Considering modeling the subtopics corresponding to the user intents contained in the queries or not, the search result diversification methods can be divided into two categories, the implicit and explicit methods. Implicit diversification models usually consider the similarities between those result documents, the more dissimilar the documents are, the more diversified the document set will be; while explicit diversification models consider the different user intents (which are represented as subtopics) behind the queries explicitly and model the document set diversity by the coverage of different subtopics corresponding to the queries issued by users. These methods can also be divided into unsupervised or supervised methods considering using handcrafted features or machine learning optimized features. Unsupervised methods use handcrafted diversification features to model the contribution of different documents, while supervised methods can learn the most optimized features automatically with machine learning models. Those two categorizations are orthogonal, and in theory the explicit methods should be better than implicit methods, and the supervised ones better than unsupervised ones. While in practice, the effectiveness of diversification models may be influenced by other factors such as the amount of training data and the quality of subtopics. Since the development of reinforce learning was involved in information retrieval (IR), recent researchers have also tried to use the reinforce learning methods to diversify search result as the diversified ranking progress can be treated as decision making. In this paper, we first present the definition of search result diversification, then categorize the existing diversification methods and introduce the representative methods. We further introduce the diversity evaluation metrics and experimental results of comparing those methods. At last we introduce several trending research directions for search result diversification.

Keywords information retrieval; user intent; diversification; personalization; reinforced learning

1 引 言

传统的信息检索技术通过给定的用户查询获得搜索结果,并按照相关性排序后返回给用户.相关研究[1]指出,大部分用户提交给搜索引擎的都是短查询,并且具有歧义性和宽泛性.例如,不同的用户可能会使用同一个查询来检索不同的信息.例如,一个用户用查询词"苹果"查找关于"苹果手机"相关的信息,而另外一个用户则可能用同样的关键词来查询"水果苹果".同时,用户往往也希望检索的结果中包含满足自身需求的各个方面的信息,而不仅仅是冗余和重复的信息.例如,查询苹果手机的客户,可能希望返回的结果中能够包含"苹果手机介绍"、"苹果手机购买"、"苹果手机评测"等各种信息.

传统的搜索排序通常仅考虑结果文档与查询之间的相关性,因此当面对有歧义的查询时,传统的信息检索技术将面临返回结果冗余性过大、覆盖面过窄、难以命中用户意图等问题.而且不同用户意图在相关性排序中的地位并不平等,越是热门的话题(或主流用户意图)越容易得到更高的分数.这会导致对搜索结果的排序将过分偏向于某些热门话题的结果文档.

搜索结果多样化是解决上述问题的主要手段之一.它的基本思想是,让搜索引擎返回结果的时候,同时考虑到搜索结果的相关性和多样性.在满足相关性的同时,尽可能地提升结果的多样性,使搜索结果中包含不同的子话题,覆盖尽可能多的用户需求.

表 1 展示了对查询“人大”的搜索结果进行多样化的实例. 如上文所述，在“两会”召开期间，搜索“人大”，传统的排序算法在仅关注文档相关性的情况下，会更加倾向于把“全国人民代表大会”排在搜索结果的前面，如表 1(a)所示. 此时，如果用户想要搜索的是关于“中国人民大学”的相关内容，就会发现排序靠前的结果不是自己所需要的，其用户体验将受到影响.

而一个基于搜索结果多样化模型的排序方法，会倾向于返回如表 1(b)所示的结果，将“全国人民代表大会”和“中国人民大学”相关的结果文档，穿插排列形成一个结果列表. 这样，无论用户的真实需求是“全国人民代表大会”还是“中国人民大学”，都能在比较靠前的位置找到自己需要的结果. 经过多样化排序之后的搜索结果列表，可以大大降低搜索结果无法满足不同用户需求的风险.

表 1　查询“人大”的排序结果示例

(a) 多样化前的排序结果

排序	结果文档内容
1	D_1：全国人民代表大会
2	D_2：全国人民代表大会
3	D_3：全国人民代表大会
4	D_4：中国人民大学
5	D_5：中国人民大学

(b) 多样化后的排序结果

排序	文档内容
1	D_1：全国人民代表大会
2	D_2：中国人民大学
3	D_3：全国人民代表大会
4	D_4：中国人民大学
5	D_5：全国人民代表大会

目前，搜索结果多样化的相关研究已经成为信息检索领域近年来的重要热点之一，学者们提出了一系列的多样化模型和多样性评价方法. 现有的多样化模型可以按照如下两种方法分类：从是否考虑到查询中包含的潜在用户意图(即子话题)的角度看，可以将分为隐式(implicit)多样化模型和显式(explicit)多样化模型；从多样化规则是人工制定的还是机器学习自动生成的，又可以分为非监督式(unsupervised，或启发式)和监督式(supervised，或学习式)多样化模型. 这两种分类方式是正交的. 同一种模型，既可以按照前者判定它是隐式的还是显式的模型，也可以按照后者判定为启发式的或学习式的模型.

本文第 2 节将探讨搜索结果多样化的定义，对搜索结果多样化的目标进行较为抽象的概括；第 3 节依次介绍目前已有的多样化模型；第 4 节介绍信息检索领域已有的评测指标和它们在多样化领域的变种；第 5 节对上文所述的多样化模型的性能进行分析汇总；第 6 节介绍目前已有的对实际运行的分布式系统进行多样化研究的成果；最后我们总结当前搜索结果多样化的研究现状，并展望今后的发展方向. 需要说明的是，虽然隐式/显式和启发式/学习式这两种分类方式是正交的，但由于目前已有的绝大多数模型都是启发式的，且学习式的模型也是基于启发式模型发展而来的，因此第 3 节将首先分别介绍启发式的隐式与显式模型，然后再介绍监督式学习模型. 此外，除了传统的基于监督式学习的模型之外，近年来强化学习也在多样化领域得到了应用，第 3 节也将一并介绍此类模型.

2　搜索结果多样化的定义

传统意义上的概率排序原则(Probability Ranking Principle)[2]将返回的结果文档本身看作是彼此相互独立的，而搜索结果多样化研究则应当考虑结果文档彼此之间的差异性，类似的思想最早在 1964 年由 Goffman 提出[3]. 对于搜索结果多样化的研究，关键在于如何在对返回的全体搜索结果进行排序时，权衡每一个结果文档与查询的相关性，和其对搜索结果整体多样性的影响. 早期的多样化排序模型最具有代表性的是 Carbonell 和 Goldstein 提出的最大边界相关性(Maximal Marginal Relevance，MMR)[4]模型. 这一算法初次发表于 1998 年，第一次使得搜索引擎可以返回一个多样化的结果列表，尽可能地提高排在前面的结果文档能满足用户需求的概率.

Drosou 等人[5]从以下三个角度对搜索结果多样化进行了定义：

(1) 基于内容(Content-based)的定义

在这种定义下，搜索结果的多样性被视为返回结果列表中的文档之间的多样性：返回的每一个结果，不仅要跟查询具有相关性，彼此之间也应当尽可能地两两互不相似. 以 MMR 为代表的隐式多样化模型都是从这个角度来实现多样化排序的.

(2) 基于新颖性(Novelty-based)的定义

新颖性的概念与内容多样性类似. 与多样性的不同之处在于，基于新颖性的一类多样化定义，指的是当前返回的结果文档不仅要与此前已有的结果文档不相似，而且要包含此前已经返回的其他文档所

没有包含的新的信息.

(3)基于覆盖度(Coverage-based)的定义

在大部分搜索结果多样化研究中,用户的查询都可以被重构为一系列潜在的用户意图组成的集合,此类潜在的用户意图通常用子话题(subtopic)表示,每一个子话题代表一个用户意图.基于覆盖度的定义主要考量全体结果文档对与查询相关的子话题的覆盖程度,在这种定义下,一个结果集能够覆盖的相关子话题越多,其多样化程度越好.目前大部分显式多样化模型都采用了基于覆盖度的多样化定义,例如著名的 xQuAD 模型[6].

除了上述定义方式之外,还可以从用户意图的角度来看待搜索结果的多样性.Radlinski 等人[7]提出,搜索结果的多样性可以分为外在多样性(Extrinsic Diversity)和内在多样性(Intrinsic Diversity).正如上文所述,用户对于搜索结果多样性的需求,主要来自于两个方面:外在的多样性表示查询本身的歧义性,即同一个查询可能对应多个主体对象;内在的多样性指用户自身需求的多样化,即对于同一个主体对象,用户可能会需要多个不同方面的信息.

以上所述的所有多样化定义,最终都可以给出以下的形式化表述:给定一个查询 q,返回一个多样化的搜索结果 $R(q)$.$R(q)$中的全体文档应当满足以下需求:既要跟查询 q 本身相关性较高,又要尽可能地覆盖 q 的不同方面,降低自身的冗余度.搜索结果多样化的目的即为尽可能满足用户多样化的信息需求,让排序靠前的结果尽可能覆盖查询对应的多个事物及其不同方面.

3 搜索结果多样化算法

现有的搜索结果多样化算法可以粗略地按照两个维度进行分类:根据是否显式地考虑与查询相关的不同解释和不同方面(即子话题)[8],可以分为隐式(implicit)方法和显式(explicit)方法两类;按照规则生成的方式分类,搜索结果多样化算法也可以被分为启发式(或非监督式,unsupervised,即人工指定)的方法和学习式(或监督式,supervised,使用机器学习模型)的方法.这两种分类方法是正交的,互不重叠——一种多样化方法可以既是隐式或者显式的方法,也是启发式或者学习式的方法.表 2 展示了目前已有的传统多样化方法的分类汇总,某些非传统算法(例如短文本流多样化算法和基于强化学习的多样化算法)未包含在内,本节将单独叙述.启发式的多样化方法比学习式的方法发展得更早,模型种类更多,且目前的学习式方法通常都以启发式方法为基础引入监督式学习的成分,因此接下来本文将主要以启发式方法为切入点介绍隐式和显式多样化排序方法,然后再单独介绍基于监督式学习模型的学习式方法.除了上述分类介绍的启发式和学习式模型之外,还有一些针对短文本流进行多样化的模型未列入表中.此外,近年来也有一部分学者将强化学习引入了搜索结果多样化中,由于强化学习方法并不属于上述分类,因此也没有列入到上述表格中,后文将单独介绍基于强化学习的多样化方法.

表 2 传统多样化算法一览

	启发式	学习式
隐式方法	MMR[4]	SVM-DIV[12], R-LTR[13], PAMM[14], NTN[15]
显式方法	IA-Select[9], xQuAD[6], PM2[10], HxQuAD[11], HPM2[11]	DSSA[16]

3.1 隐式多样化排序方法

这类方法的特点是:它们将注意力放在文档本身,通过对文档之间差异性进行判断,依次选出下一个和查询相关并且和已选文档差异性较大的文档,以此来获取与查询意图相关而自身彼此间又具备差异性的结果文档集.

如上文所述,Carbonell 和 Goldstein 提出的最大边界相关性(Maximal Marginal Relevance,MMR)[4]最早提出了将文档自身的多样性与文档和查询之间的相关性的线性组合用于文档多样化排序的核心思想.其优势在于可以维持相关性和多样性的平衡,防止由于排序时过度偏向相关性而造成查询结果过度同质化.本文中介绍的所有多样化模型都使用了 MMR 模型的这一核心思想.

MMR 模型可以用以下公式进行表述:

$$\mathrm{MMR}^{\mathrm{def}}=\arg\max_{d_i\in R\setminus S}\left[\lambda P(d_i|q)-(1-\lambda)\max_{d_j\in S}P(d_i|d_j)\right] \tag{1}$$

上述公式中,R 表示全体搜索结果,S 表示当前已选中的全体文档,d_i 表示下一个候选文档,d_j 表示当前已选中的全体文档集合中的文档.$P(d_i|q)$表示候选文档和查询 q 的相关性,而 $P(d_i|d_j)$表示该文档与已有文档的相似性,显然候选文档与已有文档的相似性越低,其相对于整个已选中的文档集的新颖性也就越高.参数 $\lambda\in[0,1]$用于调节相关性和新颖性之间的比例.

在此之后的大部分隐式多样化方法，都是基于MMR模型进行改进的，不同模型之间的差异主要体现在对文档相似度计算方法的不同. 例如Zhai等人[17]在MMR算法的基础上进行扩展，通过语言模型框架来对相关性和新颖性进行建模. 他们发现在多样化排序时新颖/冗余性的作用比相关性更大，因此在对结果文档进行排序时必须考虑文档冗余的风险. 随后Zhai和Lafferty[18]利用贝叶斯决策理论将排序过程模拟为一个决策过程，定义一个损失函数(loss function)代表用户对文档的偏好. 该函数认为如果用户已经看到过一个文档的内容(或者类似内容)，那么选择该文档的风险将大于选择一个用户从未看过的全新内容的文档. 由此该算法将多样化排序问题转化为最小风险问题. Zhang等人[19]在2005年提出了一种根据文档内容来构建图的方法. 他们为搜索结果多样化问题定义了两个指标，多样性(diversity)和信息丰富性(information richness). 前者代表多个文档中包含的不同话题，后者代表某个具体文档中的覆盖话题情况. 他们由此构造了有向图Affinity Graph(AG)，提出了相应的多样化算法. Zhu等人[20]认为排在结果文档序列前列的文档应该尽可能地不相同，才能更广泛地覆盖查询对应的多个用户意图. 他们提出的Grasshopper算法是基于吸收马尔可夫链的随机游走模型. 该方法将已选择的文档状态改为吸收状态，降低与已排序项重复的文档的重要程度，从而防止了将冗余文档排在前面的情况.

以上的一系列方法都是在MMR模型基础上提出相应改进措施的方法，相关实验证明这些方法都具有有效性. 但是这些方法只是单纯从文档内容本身出发，尽可能提高文档之间的不相似性，模型本身并不知道选择的文档具体覆盖了什么用户意图，也没有具体衡量应当优先覆盖哪些用户意图. 从不相似性的角度来看，这些模型所返回的结果文档集确实具有多样性. 但是搜索结果多样化的根本目的在于覆盖尽可能多的潜在用户意图，这些模型返回的结果文档仍然不能充分满足用户对多样化信息的需求. 因此，学者们开始研究如何显式地引入与查询相关的外部信息，即代表不同用户意图的子话题(subtopic)来实现搜索结果多样化. 与隐式多样化方法相对，这类方法称为显式多样化排序方法.

3.2　显式多样化排序方法

顾名思义，显式多样化排序方法即显式地考虑查询所包含的子话题(subtopic)，让返回的结果文档集里较为靠前的结果可以尽可能覆盖较多不同的子话题. 这些子话题代表了查询所对应的不同事物及其不同方面，因此提高子话题覆盖程度即可提高搜索结果多样化程度. 如上文所述，本文中提到的几种显式多样化模型也使用了MMR模型的核心思想，主要是在文档多样性的计算方面做出一些变动.

通常情况下，显式多样化方法在实际运行的过程中，可以分为两个步骤：首先针对给定的查询，抽取其所对应的子话题；然后利用这些子话题，计算各个文档对不同子话题的覆盖程度，实现多样化排序.

3.2.1　显式多样化排序的子话题抽取

这一部分主要介绍关于子话题抽取部分的相关内容. 不同的算法通常采用不同的方式对查询生成子话题，例如对于NTCIR查询“三毛”(id=13)，基于Google Suggestions可以获得以下的一系列子话题：三毛流浪记、三毛作品、三毛从军记、三毛语录等.

Wang等人[21]在2013年，对子话题挖掘的相关方法进行了总结：子话题的挖掘，可以按照来源，分为内部抽取、外部抽取和混合抽取.

内部抽取指的是从检索到的文档中进行集中抽取，抽取到的子话题是隐式表达的，一般是潜在的子话题. 例如Dang等人[22]通过查询重构，对锚文本应用聚类方法，来抽取查询的不同方面(aspects). 他们在研究中使用锚文本作为对短查询进行重构的结果，然后对重构的结果进行聚类. Li等人[23]通过对结果文档进行聚类，通过聚类结果对查询进行分类，将获得的类别(Categories)作为潜在的用户意图，以得到对应的子话题.

与之并列的则是外部抽取. 外部抽取指的是从查询日志等外部Web资源中入手，外部抽取到的子话题是显式表达的，通常是重新排列组合的查询. Agrawal等人[9]发现文档和查询都可能包含多个用户意图，他们设计的基于用户意图的IA-Select算法通过ODP分类信息(ODP taxonomy)为查询和文档划分类别(category)来获取子话题. Rafiei等人[24]利用不同结果文档的点击率来解决搜索结果多样性问题，对结果文档的每次点击被视为查询的相关性投票，用矩阵最优化的方法来获取用户意图. Yin等人[25]从流行的子话题入手，利用已有的分类类别(例如Google Insights)对网页进行子话题分类，然后对结果文档按照子话题进行分类实现多样化.

这里需要指出内部抽取和外部抽取各有各的特

点和优势，目前学界普遍倾向于采用混合抽取的方式，即同时进行内部抽取和外部抽取来最大化地获得子话题. 子话题的抽取来源本身并不存在一个适用于所有情况的最优解，不同的查询在不同类型的数据上可能表现出不同的多样化特征. Dou 等人[26]在 2011 年首先提出了基于四种不同来源(锚文本、查询日志、聚类文档、网站)进行子话题混合抽取的方案，并提出了一种开放性的通用多样化框架，明确地整合不同数据来源的子话题以共同完成多样化. Zheng 等人[27]提出了一种新的方法，通过将概念之间的层级关系引入到子话题挖掘模型中来挖掘结构化数据中的子话题. 研究发现概念自身之间存在着类别和层级关系，把这些层级关系纳入到模型中可以显著地提高子话题抽取的效率. Nguyen 等人[28]也将查询日志和结果文档本身结合起来，通过随机游走的方法分析和抽取得到随时间变化的时效性子话题.

表 3 展示了三种抽取方式的主要工作所对应的子话题来源.

表 3 三种抽取方法的主要子话题来源一览表

抽取方式	作者	子话题来源
内部抽取	Dang 等人[22]	锚文本聚类
内部抽取	Li 等人[23]	结果文档聚类
外部抽取	Agrawal 等人[9]	查询和文档分类
外部抽取	Rafiei 等人[23]	不同结果文档的点击率
外部抽取	Yin 等人[25]	已有的分类类别
混合抽取	Dou 等人[26]	锚文本、查询日志、聚类文档、网站
混合抽取	Zheng 等人[27]	按层级分类的概念
混合抽取	Nguyen 等人[28]	查询日志和文档集(随时间变化)

总的来说，显式多样化模型的表现直接受到子话题抽取结果的影响. 学者们已经提出了一系列抽取子话题的方法，并进行了总结，而实践中较为常用的方法是基于查询建议(如 Google Suggestions)获取子话题的方法，下文中介绍的一些显式多样化模型(如 HxQuAD，DSSA 等)都使用这一方法来获取子话题. 另外，针对不同的子话题来源是否有相应的抽取方法，以及如何更好地总结抽取自不同来源的子话题等，都是需要进一步研究的问题.

3.2.2 显式多样化排序模型

这一部分主要讨论两种最具有代表性的模型，分别为 Santos 等人[6]提出的 xQuAD 模型，和 Dang 等人[10]提出的 PM2 模型. 目前已经有许多学者在这两种模型的基础上提出了一系列改进模型. 这一部分将依次对这两种模型进行介绍.

现有的显式多样化模型仍然是在 MMR 模型的基础上发展而来的，遵循 MMR 的基本规则，将对结果文档好坏的评价视作相关性和多样性的线性组合，如以下公式所示：

$$(1-\lambda)P(d|q)+\lambda P(d,\overline{S}|q) \tag{2}$$

其中，$P(d|q)$表示文档 d 的相关性，$P(d,\overline{S}|q)$表示文档 d 被发现且不同于搜索结果中已有文档的概率，该项可以用于表示文档 $P(d,\overline{S}|q)$这一项可以用于表示文档 d 的多样性. 对于不同的多样化模型，一般主要在文档多样性的衡量方式上有所不同. 接下来将主要介绍两种最具有代表性的模型：xQuAD 和 PM2.

对于多样化模型，本文主要关注其对多样性部分的改进，因此下文将着重讨论类 MMR 模型线性组合中的多样性部分，暂不讨论相关性.

3.2.2.1 xQuAD 模型

Santos 等人[6]在模型中提出，将一个不明确、有歧义的查询分解成一个子查询的集合，每一个子查询对应一种子话题，分别代表原始查询在不同方面的潜在信息需求.

对于 $P(d|q)$，将查询 q 按照子话题拆分为一系列子查询 $Q=\{q_1,\cdots,q_i\}$，假设在查询 q 中，文档 d 与已经查询到的文档集合 S 相互独立，则 xQuAD 模型的文档评分可以用以下公式来表述：

$$(1-\lambda)P(d|q)+\lambda\sum_{q_i\in Q}[P(q_i|q)P(d|q_i)\prod_{d_j\in S}(1-P(d_j|q_i))] \tag{3}$$

与 MMR 模型相同，xQuAD 模型也使用参数 λ 调节新颖性与相关性的平衡. 从该公式可以看出，xQuAD 模型在衡量搜索结果多样性的过程中，实际上考虑了以下三个方面：

(1) 子话题本身的重要性，即子话题在原始查询中所占的权重 $P(q_i|q)$；

(2) 文档对于子话题的覆盖程度，即文档与子话题间的相关性 $P(d|q_i)$；

(3) 文档的新颖性$\prod_{d_j\in S}(1-P(d_j|q_i))$，其中$(1-P(d_j|q_i))$表示此前已经选中的文档集 S 中的文档 d_j 没有满足子话题 q_i 的概率，显然$\prod_{d_j\in S}(1-P(d_j|q_i))$即表示此前所有已经选中的文档都没有满足 q_i 的概率，即为当前选中文档 d 的新颖性.

xQuAD 模型将对于文档新颖性的计算从文档间的直接比对转化成了对文档 d 满足子话题的边际效益衡量，提供了一种衡量文档新颖性的有效方法，它并不需要直接比对结果集中的各个文档，也不

需要对所有结果文档中的词项进行搜索，只需要基于结果集 S 中的每个文档与给定子话题的相关性更新该子话题的新颖性即可. 为此只要在倒排表中额外查找一些符合该子话题的文档即可，并不会增加更多时间开销. 该模型提出于 2010 年，其在当年的 TREC 会议上评测得到了当时最佳的多样化效果.

3.2.2.2　PM2 模型

由 Dang 和 Croft[10] 在 2013 年提出的 PM2 模型，跟 xQuAD 模型一样也将重心放在了如何从子话题的角度入手，对文档的多样性进行衡量. PM2 属于较为典型的话题比例模型(topic proportionality model). 这一类模型的基本思路是：一个多样化的结果文档集，它的文档所包含的子话题分布应该满足一定的比例，结果列表中返回的覆盖某一子话题的文档数，跟这个子话题的权重(即重要程度)应当成正比.

例如，对于查询“java”，在返回的所有结果中，关于 JAVA 编程语言的文档占了 90%，而关于爪哇岛(岛屿名称同样也叫作 java)的文档占了 10%，那么返回的顺序靠前的结果文档中，关于 JAVA 编程语言和爪哇岛的结果比例大致应该为 9∶1，在每次选择当前最佳文档 d^* 时，应该优先选择最能够改善当前子话题分布比例的文档.

在确定子话题分布的比例时，PM2 参照了新西兰议会选举中使用的 Sainte-Lague 模型. 该模型根据党派收到的选票数量来确定党派的席位比例. 它的基本思想是，一个党派获得新的席位的优先级与党派得票总数成正比，而与党派当前已经占有的席位成反比. 如果一个党派得票较多，而当前已经占有的席位较少，那么它将会优先被分配更多的席位. 同理，对于搜索结果来说，某个子话题的重要性越高，在已选中的结果文档集中对应的文档数目越少，则该子话题就越需要改善. PM2 算法选取最佳文档有两个步骤：首先找出当前最应该改善的子话题 q^*，然后下一轮选择偏向于该子话题并且与其他子话题有一定相关性的文档作为当前最佳文档 d^*.

需要注意的是，在搜索结果多样化的过程中有以下两点与现实中的政党席位分配不同：首先，在实际的搜索结果中，一篇文档对应的子话题很可能不止一个；其次，对于给定子话题，如果没有非常理想的候选文档，则优先保证相关性，即宁可选择一个冗余文档，也不选择一个不相关的文档.

PM2 模型的具体计算方式如下所示：

$$qt_i=\frac{P(q_i|q)}{2\times s_i+1} \tag{4}$$

qt_i 表示当前的子话题 q_i 目前需要改善的份额(quotient)，其中 $s_i=\sum_{d\in S}\frac{P(d|q_i)}{\sum_{q_j\in T}P(d|q_j)}$. 这里 $P(d|q_i)$ 是文档 d 和子话题 q_i 相关的概率，$\frac{P(d|q_i)}{\sum_{q_j\in T}P(d|q_j)}\in[0,1]$ 代表相对于所有子话题，q_i 所占据的文档 d 的席位. 将已选择文档集 S 中所有的文档 d 的席位累加起来，就得到了 S 被子话题 t_i 所占用的程度 s_i. 显然 qt_i 与子话题 q_i 在查询中的重要程度 $P(q_i|q)$ 成正比，与 s_i 成反比.

每次 PM2 模型都选择当前 qt_i 值最大的子话题作为当前选择的最佳子话题 q_i^*，然后选择与当前子话题 q_i^* 相关性最大，并且与其他子话题比较相关的文档，作为当前的最佳文档 d^*. 同样基于 MMR 的思想，算法引入选择参数 λ 来控制最佳子话题和其他子话题在选择最佳文档时的贡献程度. 当 $\lambda>0.5$ 时，算法优先选择与 q_i^* 相关性较高的文档.

Dang 等人在论文中提出，目前要针对某个话题抽取相关子话题仍然存在一些现实的困难[10]. 为此他们专门提出了一个假设：对于某个文档 d，其与某个话题 A 的相关性应当正比于其与 A 对应词汇的相关性. 基于这一假设，他们提出了一种基于词汇的多样化方法，即针对给定话题，直接用贪心算法从排序的结果文档集中抽取对应词汇，然后用得到的词汇表示话题. 实验结果证明这种基于话题词汇的方法要优于直接考虑子话题本身的方法，同时也证明了 PM2 模型的优越性.

以上两种模型是显式多样化排序模型中最具有代表性的两种方法. 众多学者在之后的研究中，也都对这两种方法进行了进一步的改进.

3.2.2.3　基于子话题多层级分类方法的 HxQuAD 和 HPM2 模型

对于包括 xQuAD 和 PM2 在内的多样化模型，最为重要的一个步骤就是按照不同的粒度将一个查询拆分成子话题. 对于大部分查询，使用粒度更加细致的子话题拆分方案可以得到更好的子话题覆盖效果. 但是粒度拆分过细会导致模型选取的多个细分子话题无法覆盖不同的上级子话题，削弱多样化效果. 例如对于查询“Defender”，对应的子话题之一“LandRover Defender”，可以进一步细分出二级子话题“LandRover Defender USA”和“LandRover

Defender Parts”等. 如果简单地把这些不同层级的子话题平摊在一个列表里,那么多样化算法连续选取的多个细分子话题很可能都是跟同一个上级子话题相关的,不利于多样化效果的提升.

Hu 等人[10]在已有的模型基础上,提出了对不同子话题进行多层级分类的方法. 这种方法按照多个不同层级的粒度对查询拆分子话题,这样可以同时具备粗粒度和细粒度子话题拆分的优点. 继续以查询“defender”为例,可以建立一种查询——一级子话题和二级子话题的两层分类的话题树模型. 根节点表示查询本身,第一层结点表示查询“defender”对应的子话题,例如“LandRover Defender”,“Windows defender”等,而第二层叶子结点则代表按照子话题进行进一步细分的二级子话题,上文中提到的“LandRover Defender USA”在该模型中与以及自话题“LandRover Defender”对应,如图 1 所示.

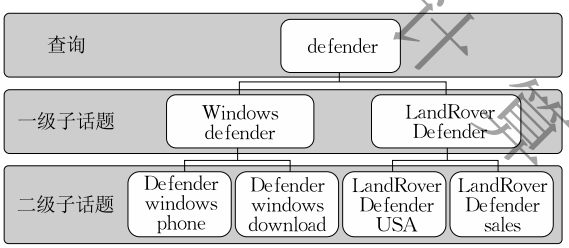

图 1 子话题分层的话题树结构

这个分层结构的模型可以按不同粒度和层级关系对子话题进行分类,进而大幅增加了正确多样化意图的概率.

层次化模型使用已有的一些方法[29]来挖掘查询中的层次化信息,根据已有的相关工作,Hu 等人在研究中直接使用了 Google 的搜索建议作为层次化子话题的来源. 今后的工作可以在子话题挖掘方法上做进一步的探索. 这种对用户意图多层分类的多样化框架,在使用中可以通过对已有的多样化模型进行扩展而发挥效用. 模型假设同一层次的子话题之间彼此独立,使用贝叶斯公式来衡量下级子话题权重. 例如对子话题 q_1 的下级子话题 $q_{1.1}$,其相对整个查询的权重表示如下:

$$P(q_{1.1}|q)=P(q_{1.1}|q_1)P(q_1|q) \tag{5}$$

即下级子话题相对于查询的权重,由它相对于上级子话题的权重,和上级子话题相对于查询的权重共同决定. 在此基础上,可以对已有多样化模型进行扩展.

例如在经典的 xQuAD 模型的基础上引入分类方法,即可扩展得到 HxQuAD 模型,类似地,还有在经典 PM2 模型的基础上加入子话题分层次框架而来的 HPM2 模型.

如果不引入分层结构,传统多样化方法很难对不同层次的子话题实现区分和平衡,往往会重复地使用多个对应相同一级子话题的细分二级子话题. 而相关研究[30]指出,现实存在的真实用户意图所对应查询的不同方面,恰好可以按照两层模型进行分类和扩展. 相关实验结果显示,基于多层级子话题的层次多样化框架,相比子话题列表平摊的传统多样化模型,在 ERR-IA、α-NDCG 等主要指标上都有着明显的优势. 即便只使用单层子话题,层次多样化框架依然可以带来有效的性能改进,因为话题树上的单层节点本身即可隐式地使用话题层级信息辅助多样化排序.

3.2.2.4 针对短文本流的多样化

上文所述的 xQuAD 和 PM2 等多样化模型,都属于针对非流式文本的静态多样化模型,对于较大的,静态的长文档集有较好的效果,而文档动态的潜在子话题的特征与分布通常会被模型忽略,因此并不能直接用于对短文本流进行多样化排序,此类问题还需要进一步研究.

短文本流多样化问题,可以详细定义为针对特定时间点的有歧义的查询,对流式输入的短文本进行搜索结果多样化. 该问题的输入包含有歧义的查询,输出为随时间变化的短文本多样化排序结果,该结果应当尽可能覆盖查询在近期时间的不同方面. 对短文本流进行多样化排序有很多具体的应用,例如 Twitter 或微博的搜索功能就需要返回多样化的随时间变化的短文本搜索结果.

针对长文本流的多样化研究,最早由 Minack 等人[31]和 Chen 等人[32]进行了相关研究. Minack 等人提出了两种针对数据流的增量多样化算法:最大-最小增量(Max-Min Incremental,MMI)和最大-总和增量(Max-Sum Incremental,MSI);Chen 等人提出了针对文本流的多样化敏感(Diversity-Aware)Top-k 算法. 上述算法都假设文本内容丰富,且直接计算文档不相似性即可衡量文档集整体的多样化程度,该假设对短文本并不成立,因此这类方法并不适合短文本.

解决该问题的一个自然的方法是使用话题模型,例如著名的 LDA 模型[33],而在文本流中,针对给定查询的子话题分布会随时间改变,现有的静态话题模型并没有考虑到这种情况. 因此在 LDA 等静态话题模型的基础上,学者们又提出了一些动态

话题模型，例如ToT(Topic over Time)[34]、Twitter-LDA[35]等，但是大部分动态话题模型都是针对长文本设计的，只有Twitter-LDA等少数几个模型可用于短文本，且这些模型都不以多样化为目标.

为了解决这个问题，Liang等人[36]提出了一种基于流式短文本的多样化算法.该多样化算法主要分为以下几个部分：首先通过一种基于动态Dirichlet分布的多项话题模型(Dynamic Dirichlet Multinomial Mixture Model，D2M3)获得短文本流所包含的潜在子话题；接下来使用子话题执行多样化排序.多样化算法首先使用话题模型获取当前时间t下对应查询q的潜在子话题及其概率分布，文本流的子话题分布可能随时间变化，在不同时刻各个子话题的权重也将有所变化.在使用子话题执行排序的过程中，算法将首先运用一个时间敏感型搜索模型来获取前k个相关文档，然后通过一个基于PM-2模型和子话题动态分布(由动态话题模型得到)的流式多样化算法(Streaming Diversification Algorithm，SDA)来对前k个文档进行多样化排序.

此处需要强调的是，获取短文本子话题分布和对短文本进行多样化排序，是两个独立的过程，即D2M3和SDA之间不存在耦合关系，可以进行拆分和组合.

相关实验在Twitter文本数据集上进行，实验结果显示在nDCG等主要评价指标上，SDA与D2M3组合的模型领先于所有传统的静态多样化模型和传统的长文本流多样化模型，也领先PM2与传统话题模型的组合模型.基于PM-2修改的SDA算法，对短文本流多样化有着良好的性能，显著领先于其他模型.此外，将SDA与其他话题模型组合，仍然能得到领先于传统模型的性能，但SDA与D2M3组合可以显著扩大性能优势.

除了上述的主要方法之外，针对显式多样化中的一些细分的具体问题，也有学者提出了一些多样化方法.例如，为了用尽可能少的文档覆盖尽可能多的用户意图，Carterette等人[37]提出了主题分面检索(Faceted topic retrieval)的概念，即假设某个查询本身的含义是明确的，只是存在针对该查询不同方面的解释；而对于平衡参数λ，一般的获得方法是对照查询结果调参而获得最优值，而针对歧义程度不同的查询，对应最优化的λ应该也不一样，因此Santos等人[38]提出了一种监督式方法，先对查询依次标注最优参数，然后再根据特征集和已经被标注的查询获取回归模型，并用于预测对未来的“不可见(unseen)”查询的参数λ的最优值，进而提高多样化结果.

启发式的多样化方法面临一个共同问题——只能使用人工选择的特征，且调参的代价很高.于是一部分学者便将注意力放在了使用监督式学习方法来进行排序并自动学习相关参数的多样化模型，即学习式排序模型.

3.3　基于监督式学习的多样化排序模型

3.3.1　学习式排序模型概述

如前文所述，基于MMR模型的隐式与显式多样化模型，都属于启发式的方法，目前已经得到了广泛的运用.启发式的方法依赖于一系列效用函数，而这些函数只能使用事先选定的少数特征，并且调参成本很高(尤其是在复杂的搜索中).相对于启发式方法，基于机器学习模型的学习式方法可以在函数中自动训练多种特征，并且通过加入新的训练数据就可以自动优化参数.

对于搜索结果相关性排序的学习式方法研究此前已经有了长足进步，这类方法在学术界和工业界都得到了积极地发展与应用，但是针对搜索结果多样性的学习式方法的相关研究却相对较少[39].造成这一现象的主要的原因是，此前已有的针对搜索结果相关性的学习式排序方法普遍遵循文档独立性假设，即认为文档之间是相互独立的，文档的评分只与其自身内容与查询的相关性有关.然而如上文所述，基于MMR的搜索结果多样性模型，需要考虑到结果集中文档与文档之间的关系，这显然与文档独立性假设相违背.此外，考虑文档之间的关系而进行排序学习也存在一个现实的困难：这属于一个典型的子空间选择(subset selection)问题，该问题已经被证明为NP-难解问题.

Yue和Joachims[12]首先进行了一些前期的探索.他们提出了SVM-DIV模型，将机器学习的任务形式化为通过预测多样化子集来建立多样化排序，并引入了结构化的支持向量机模型(Structural SVM).相对于基于MMR模型的启发式方法，此方法只专注于对多样性的捕捉，没有引入相关性考量.此类排序模型仍然沿用了传统的针对结果相关性进行排序的形式，没有考虑到结果之间的关系，因而排序的效果不尽如人意.突破瓶颈的关键在于，如何让学习式模型也能像MMR模型一样同时考虑到相关性与多样性，并将计算的复杂度压低在可以接受的范围之内.

接下来，本文将依次介绍学习式排序模型的发

展,主要介绍的模型如表 4 所示,表中各项以发表时间的先后顺序进行排序.

表 4 主要监督式学习模型及其特点一览表

模型	作者	特点	子话题
SVM-DIV	Yue 和 Joachims[12]	只捕捉多样性	隐式
R-LTR	Zhu 等人[13]	突破了文档独立性假设,考量文档间关系	隐式
PAMM	Xia 等人[14]	直接优化多样性评价指标,无需单独设置损失函数	隐式
NTN	Xia 等人[15]	自动捕捉最佳新颖性特征	隐式
DSSA	Jiang 等人[16]	捕捉子话题,显式监督式学习	显式

3.3.2 关系排序模型(R-LTR)与 PAMM 模型

受到 MMR 模型的启发,Zhu 等人在 SIGIR2014 上发表了论文,提出了关系学习排序模型(Relational Learning to Rank,R-LTR)[13].他们对互联网搜索中用户至上而下的浏览动作进行分析,考虑到多样性排序中基于贪心算法的近似选择过程,将多样化排序当做一个顺序的选择过程.R-LTR 模型突破了传统的学习式方法所基于的文档独立性假设,并且能在有限时间内完成对目标的合理近似假设.

跟大部分多样化排序模型一样,R-LTR 的基本结构与原始的 MMR 模型结构相同,定义为文档与查询的相关性得分,和在给定已选文档集的情况下当前文档的新颖性得分的加和.跟 MMR 模型一样,相关性得分只考虑文档自身的内容,而新颖性则依赖于当前待选文档与给定的文档集之间的不相似性.在 R-LTR 模型的训练中,定义样本空间为对结果集的所有可能的排序的组合,而每个组合的出现概率则由自上至下的顺序文档选择过程中对各个结果文档的打分来决定.这里使用 Plackett-Luce 模型进行极大似然估计,即定义一个网页列表之后的任何可能的排列所出现的概率.训练的目标是针对多样化排序较为理想的文档组合,尽可能地提高这些组合出现的概率.

R-LTR 模型出现之后,还有其他的一些方法也是基于 R-LTR 发展而来的[40].此类方法的共同特点是,将机器学习的问题形式化为对损失函数进行最优化.但是这些损失函数通常与搜索结果多样化的评价指标并没有直接关联.因此其他学者们开发出了新的机器学习方法,目标是在关系排序的学习中直接对搜索结果多样化的相关评价指标进行优化.

为了改进 R-LTR 模型的缺陷,Xia 等人[14]提出了一种基于感知机的算法,目的是利用 MMR 模型在训练的过程中直接优化任何多样化评估指标.在 R-LTR[13]和感知机模型[41]的启发下,他们提出了 PAMM(Perceptron Algorithm using Measures as Margins)模型.

类似于 R-LTR 模型的训练过程,PAMM 利用文档的顺序选择过程,在学习的过程中,首先为任意排序结果定义一个概率,在给定标注数据下生成"正面排序(即标注正确的排序结果)"和"负面排序(标注错误的排序结果)",然后按照排序概率计算正面排序和负面排序间的间距,最后更新模型参数以确保正面和负面排序距离最大化.其损失函数如下所示:

$$\sum_{n=1}^{N}(1-E(X^{(n)},\hat{y}^{(n)},J^{(n)})) \tag{6}$$

该函数是直接通过多样化评价指标定义的,表示预测与人工标注不一致时导致的损失.与 R-LTR 模型类似,PAMM 也使用基于 Luce 模型定义的函数 F,表示生成排序列表 y 的概率,它的产生方式是迭代地从剩余文档集中选择排序靠前的文档,并使用 MMR 的边际相关性函数作为选择基础,通过优化,最终将问题转化成使用感知机框架,尽可能地降低表达式上界的过程,实际上可以看作是尽可能地最大化正面排序 $E(X^{(n)},y^{+},J^{(n)})$ 和负面排序 $E(X^{(n)},y^{-},J^{(n)})$ 间距的过程.

相对于此前已有的启发式和学习式模型,PAMM 模型具有以下几个优点:

(1)可以直接产生一个基于 MMR 的模型;

(2)可以在训练中,直接对任何多样化评测指标进行优化;

(3)在训练中,既可以使用正面结果排序,也可以使用负面结果排序.

在已有数据集 TREC Web Track 2009-2011 上,R-LTR 优于包括结构化 SVM 在内的各种方法,而 PAMM 则在 R-LTR 的基础上更进一步.PAMM 模型还有助于对于一个选定的多样化评价准则,进行有针对性的提升.

3.3.3 使用神经张量网络(NTN)抽取新颖性特征

以 R-LTR 和 PAMM 为代表的上述基于关系的模型仍然存在以下问题:如何定义一系列可以有效捕获复杂文档关系的新颖性特征.

跟传统的相关性排名学习不同,抽取搜索结果多样化所需的新颖性特征是非常困难的.在多样化排序模型中,只有少数的新颖性特征可以被利用[15].例如对于 R-LTR 和 PAMM,文档的新颖性

衡量仅仅使用了七种特征，大部分特征都是文档之间基于 TF-IDF 向量或者话题向量的余弦相似度等预先定义的相似度指标，并且分别用于文档的标题、主体和锚文本. 在实际应用中排序算法的模型研究依赖于这些人工选择特征的有效性，并且不同的排序任务需要不同的特征，几乎不可能靠人工抽取获得文档最优化的新颖性特征. 因此，需要发展一种可以自动学习新颖性特征的算法.

为了解决这个问题，Xia 等人[15]接下来又提出了一种基于神经张量网络(Neural Tensor Network，NTN)的新颖性模型. 跟此前手工定义文档相似度函数或新颖性特征的方法不同，这种基于 NTN 的新方法可以通过训练数据自动地学习非线性的文档新颖性函数. 它首先使用一个非线性张量层，通过候选文档和其他文档的联系生成新颖性信号. 然后，应用最大池化来选择最有效的新颖性信号，最后把选中的信号进行线性组合以生成最终的文档新颖性评分.

在深度学习领域，神经张量网络(Neutral Tensor Network)最初应用于在知识图谱中，主要的作用是衡量两个实体之间的关系. 对于两个维度为 l_e 的实体 e_1，e_2，NTN 的目标是预测它们之间是否存在确定性的关系 R.

这种方法可以自然地扩展到衡量搜索结果多样化排序中的新颖性特征抽取过程. 即一个候选文档的新颖性信息可以被表述为该文档和其他文档之间的双线性张量积. 将 NTN 文档新颖性模型与上文所述的学习式新颖性排序算法结合，可以自动地学习得到文档的新颖性函数. 在新的定义下，文档的 MMR 评分可以用以下函数表示：

$$f(d,S)=g_r(x)+g_n(\boldsymbol{v},S)$$
$$=g_r(x)+\boldsymbol{\mu}^{\mathrm{T}}\max\{\tanh(\boldsymbol{v}^{\mathrm{T}}W^{[1:z]}[v_1,\cdots,v_{|S|}])\} \quad (7)$$

此处 g_r 表示文档的相关性，g_n 表示文档的新颖性，矩阵$[\boldsymbol{v}]$的每一列都对应文档集中的一篇文档，$W^{[1:z]}$ 是一个张量，$\boldsymbol{\mu}$ 是张量每一层的权重. 可以看到此处使用了 NTN 来对新颖性进行定义.

使用 NTN 的优势在于，传统的方法只能通过一个预定义的函数(跟启发式方法一样)或者通过一系列新颖性特征的简单线性组合来衡量新颖性，而张量则可以多层次地衡量候选文档和当前文档集的关系. 直观地看，模型可以解释成张量的每一层都代表了查询的一个方面或者一个子话题. 每一个张量层都衡量了候选文档和当前已选文档集的多样性关系. 因此，通过多层张量层，模型可以计算出基于多个多样化方面的新颖性评分.

将 NTN 与已有的学习式多样化模型结合，可以得到新的模型. 例如与 R-LTR 结合，可以得到模型 R-LTR-NTN；类似地，与 PAMM 结合，可以得到 PAMM-NTN 模型. 在 ERR-IA、α-NDCG 等几个主流评价指标上，相关实验结果都显示，NTN 可以有效抽取文档新颖性特征，提高已有的学习式多样性排序模型的性能.

3.3.4　基于注意力机制的显式多样化学习方法

以 R-LTR 和 PAMM 等为代表的监督式学习的多样化排序模型，较启发式的传统方法在性能上具备显著的优势，并且对于搜索引擎来说，随着数据集的增加，可以不断地进行线上学习，进一步地提高模型的性能. 但是从是否将用户意图纳入考量的角度来看，这些方法都继承了 MMR 的指导思想，属于隐式的学习方法. 也就是说，跟 MMR 本身一样，这些方法只是简单地考虑文档之间彼此的相似性，并没有考虑到真实的用户意图. 尽管它们在相似性指标衡量上有着明显的优势，但它们跟以 MMR 为代表的隐式启发式方法一样面临一个问题：如何跨越降低文档冗余和提高用户意图覆盖之间的鸿沟. 因此未来的发展方向是提出一种显式考虑子话题的监督式学习的多样化排序方法.

在 SIGIR2017 上，Jiang 等人[16]提出了一种新的显式监督式学习框架，这种框架同时结合了显式非监督学习方法和隐式监督式学习方法的优点. 一方面这种模型显式地衡量子话题，可以通过优化提高用户意图的覆盖程度，另一方面可以自动地学习文档的多样化函数，并且可以捕获文档和子话题之间的复杂交互. 这种框架称为 DSSA(Document Sequence with Subtopic Attention). 具体地说，为了选择下一个新文档，模型首先考量已选择的文档序列，以捕获它们的内容和与子话题的关系. 接下来在已有文档包含的信息的基础上使用注意力机制决定接下来的子话题.

该框架的本质可以进行如下概述：沿着已经选中的文档，模型对过往的文档序列进行编码，然后注意力机制将会监视文档序列满足每一个子话题的程度. 如果一个文档覆盖了此前覆盖度较低的子话题，那么这个文档将会被评为高分. 最后通过对注意力的适应性学习覆盖多个子话题. DSSA 框架提供了一种直接的显式衡量子话题的监督式学习方法，同时将相关性和多样性纳入了考量，而多样性的考量也是综合衡量文档对子话题覆盖程度的结果.

目前,注意力(Attention)机制在图像识别和自然语言处理领域已经有了成功的应用,对于搜索结果多样化中的文档选择问题同样表现良好:将注意力集中在子话题上,可以在结果列表中优化内容.在DSSA框架中,注意力机制应用于子话题,可以衡量不同的角度和用户意图.

DSSA框架的基本形式与上文论述的几种显式模型类似:

$$S_{d_t}=(1-\lambda)s_{d_t}^{\mathrm{rel}}+\lambda s_{d_t}^{\mathrm{div}}(0\leq\lambda\leq1) \tag{8}$$

为了解决这些问题,DSSA框架在原始的显式非监督学习框架方程的基础上对其进行了扩展.

框架可以分为三个主要的部分:(1)文档序列表示组件 H,(2)子话题注意力组件 A 以及(3)评分组件 S^{rel} 和 S^{div}.

文档序列表示组件 H 将文档序列 C_{t-1} 中的信息编码成固定长度的隐藏层 h_{t-1}. h_{t-1} 可以被看作 C_{t-1} 更高层次的表现. H 通过RNN实现,作用是对已有的文档序列的信息进行编码.这里将第 t 位的文档序列表示如下:

$$\boldsymbol{h}_t=\tanh(W^n[h_{t-1};e_{d_t}]+b^n) \tag{9}$$

此处 e_{d_t} 表示当前文档的分布式表示,b^n 为偏差项. h_{t-1} 储存了前述 $t-1$ 个文档的信息.接下来使用注意力机制来衡量各个子话题对于第 t 个文档的重要性,注意力机制从排名位的首位一直扫描到最后一位,基于已有的文档序列判断接下来应当增强哪一个子话题的权重.

扫描完成之后,模型使用注意力权重来计算指定文档的多样性得分:将文档与子话题的分布式表示计算相似性,并额外组合一些该子话题的相关性指标(如TF-IDF等),即可获得文档与指定子话题的相关性得分.将文档与所有子话题的相关性得分按注意力权重线性组合,即得到文档最终的多样性分数.

在TREC Web Track 2009-2012数据集上的测试结果显示,DSSA在主流的几个多样化排序指标,例如ERR-IA,α-nDCG,NRBP等,全面地领先于目前已有的启发式显式多样化模型,例如xQuAD,PM2及其相关的衍生模型,也领先于PAMM-NTN等隐式多样化学习模型.DSSA的实验结果显示,在学习式模型中显式地衡量子话题可以避免启发式显式多样化模型对函数和参数定义的局限性.然而,准确地衡量文档和子话题之间的复杂关系,仍然具有很大的挑战性.目前DSSA模型要训练的参数较多,因此需要庞大的训练数据.由于DSSA目前使用的训练数据较少,因此如果能够收集到更多数据进行训练,应该可以更加充分地释放该模型所具有的潜力.另外,DSSA只处理多样化排序函数的学习,即假设文档和查询使用的分布式表示已经被创建好了.而实际上,这一步也可以直接使用一个模型来学习如何表示这些文档和查询,并且与DSSA进行组合,进而获得更理想的训练数据.

3.3.5 小 结

尽管显式启发式方法、隐式学习式方法和显式学习式方法都对多样性判断方法做了较大的改动,但是实际上,上述模型本质上都是在MMR模型的基本思想基础上进行改进,即多样化模型在对文档进行排序时,选择边际相关度最大的文档可以降低整个结果集合的信息冗余性.

以上所有模型无论使用哪种方法抽取多样性特征,都使用参数 λ 来控制相关性和多样性的比重.对于参数 λ 的具体取值,根据实际应用场景,对于启发式模型,该参数一般根据文档和查询的实际情况通过人工调参得到,而学习式模型通常可以通过在机器学习所使用的验证集上进行验证,自动学习得到该参数的最佳值.

如上文所述,多样化模型可以分为隐式启发式模型、显式启发式模型,隐式学习式模型和显示学习式模型四类.隐式启发式模型的缺点是性能上限最差,但模型使用的成本较低,可以在不依赖外部信息的情况下执行多样化排序.

显式启发式模型具体衡量了文档的子话题覆盖,因此在有着充分的子话题信息的情况下可以拥有比隐式启发式模型更好的效果.但是显式启发式模型的运行依赖于子话题信息的数量和质量,如果没有覆盖面足够广泛的高质量子话题来源,其性能的稳定性将无法保证.

隐式学习式模型与显式学习式模型都属于监督式学习模型,因此都依赖于人工标注的数据,在此基础上显式学习式模型同时也需要子话题信息.无论是否支持端到端学习,监督式学习模型都需要在上线运行之前使用标注数据进行预训练,以防止损害用户体验.而当训练数据累积较大时,两种监督式学习模型都可以表现出优于隐式或显式启发式模型的性能.

3.4 将强化学习应用到搜索结果多样化中

强化学习(Reinforced Learning)是一种用于决策的重要的机器学习方法,其过程可以简要叙述如下:每经过一个时间步(Time Step),由代理(Agent)

对环境(Environment)进行观察(Observation)，做出决策动作(Action)，然后从环境获得回报(Reward)作为下一步决策的依据.而搜索引擎对结果文档的排序过程，无论相关性排序还是多样性排序，都可以视作一个决策过程，即文档序列的连续选择过程.由于用户浏览搜索结果的过程是一个自上而下的过程，对于返回搜索结果序列的每一个排位，搜索引擎都应当选择当前对用户信息收益最大的文档.因此，强化学习在信息检索领域也有着日益广泛的应用.

具体到信息检索的应用场景中，整个过程实际上就相当于搜索引擎作为代理，与环境(即用户)进行交互.搜索引擎选取文档交给用户，让用户来判断文档是否满足自己的信息需求，然后将结果反馈给搜索引擎.由此可见，将强化学习应用于信息检索乃至多样化排序过程中，首先要解决的问题就是如何对要解决的问题进行形式化定义.目前，对于如何定义时间步，主要有三种不同的定义方式：按会话(Session)定义，按结果页(Result Page)定义，和按单个结果(Item)定义.针对多样化排序问题，由于多样化排序的目标通常是尽可能在靠前的位置满足用户的信息需求，因此应选择按单个结果定义的时间步.

对于多样化排序问题，另一个需要首先定义的方面就是如何获得决策中的回报，即如何衡量多样化排序的效果.最为直接的方法就是使用真实的用户数据来作为衡量回报的标准.在 2008 年，Radlinski 等人[42]提出了一种基于多臂老虎机(Multi-Armed Bandit，MAB)算法进行线上学习式排序的模型.该模型定义了一系列有着信息需求的用户，并按顺序给用户提供一系列文档集.模型从上到下对每一个排位进行排序，对于每一个时间步 t，模型给用户按顺序展示一个文档序列，然后记录用户是否点击了其中的某一个文档——用户可以点击多次，也可以一次都不点击，但对每一个序列，用户最多只能点击一个文档.对于每一个排位，如果用户有点击至少一个文档，就将当前的收益(Payoff)记作 1，没有则记作 0，模型学习的目标即为对收益总和进行最大化.用户 i 点击某一个文档 j 的概率 p_{ij} 由真实用户的历史数据计算得到，并引入一个 MAB 算法以模拟真实用户可能发生用户意图变化的情况.这一排序算法称为 RBA(Ranked Bandit Algorithm).实验证明，该算法即使是在用户意图变化的情况下，依然可以实现最优化的最坏情况(worst-case)性能.

但是，使用真实用户数据存在着一个问题——相对于人工标注数据，真实用户数据的成本较为高昂，而商业搜索引擎又不可能以牺牲实际的用户体验为代价，在线上实时收集真实用户点击数据.因此，更为实际的一种方式，就是以模拟(Simulation)的方式获取回报数据，即通过人工标注数据和多样化评价指标测算出每一步决策的回报.

Xia 等人在 SIGIR2017 上提出了一个新的模型 MDP-DIV[43]，将马尔可夫决策过程(Markov Decision Process，MDP)引入搜索结果多样化过程，把结果文档的多样化排序过程视作一个连续的 MDP，进而将多样化排序问题形式化为一个学习 MDP 模型的问题.

对 MDP-DIV 模型，它采用以下方式，对搜索结果多样化排序任务进行形式化定义：每一步的 MDP 状态 S_t 定义为一个三元组，由已选中文档序列 Z_t，候选文档集 X_t 和编码向量 $\boldsymbol{h}_t$ 组成，$\boldsymbol{h}_t$ 用于衡量用户对每一个排位文档的信息需求效益.在此基础上，动作 a_t 和状态转移函数 T 定义如下：动作 a_t 定义为从候选文档集合中选取一个文档 $x_{m(a_t)}$，将其加入到 Z_t 中，此处应注意 Z_t 是一个有序的序列，每次插入只能将选取的候选文档放在 Z_t 的末尾，同时完成插入之后 $x_{m(a_t)}$ 将从候选文档集 X_t 中移除，最后使用学习得到的相关参数更新 $\boldsymbol{h}_t$.

如上文所述，MDP-DIV 模型是通过模拟的方法来获得回报的强化学习模型，它计算回报的方法是通过一个回报函数来获取每一步动作为用户带来的信息边际效益.该回报函数可以通过某一个具体的多样化评价指标来定义和计算，例如可以使用决策 a_t 带来的 α-DCG 指标增加值来定义回报函数.除了 α-DCG 之外，还可以根据实际需要，选择任何一个可以针对每个排位计算对应值的多样化指标，例如 S-recall 等.最后，对于给定的状态 s_t 和动作 a_t，可以计算得到策略函数 π.模型通过反复运行来优化学习相关参数，其中对回报的计算使用马尔可夫随机梯度下降方法完成(因为 α-DCG 的函数直接求导较为困难).

相对于上文所述的其他多样化模型，MDP-DIV 模型提供了一个动态地衡量用户在自上而下地浏览搜索结果时所获得的信息收益的优雅的方法，并且相对于 R-LTR 和 PAMM 等多样化模型，它还具备以下的优点：首先，MDP-DIV 可以实现端到端的多样化模型学习，不需要手工指定相关性或者新颖性指标；第二，MDP-DIV 模型使用的随机梯度上升算

法,不仅可以衡量每一步决策之后产生的直接回报,还可以衡量决策过程中的长期回报;第三,MDP-DIV 使用"用户信息收益"这一统一指标,不需要像已有的启发式模型一样将相关性和多样性分开计算,在概念上更加简洁,且在实际应用中更加有效.

相对于 RBA 算法,MDP-DIV 模型使用人工标注数据进行模拟回报计算,实际效果相对于真实用户数据不可避免地会有所下降,但是训练成本显著低于真实用户数据.表 5 从强化学习的角度,对 RBA 和 MDP-DIV 进行了对比.

表 5 从强化学习角度对比 RBA 和 MDP-DIV

模型	RBA	MDP-DIV
时间步	每一个文档排位	每一个文档排位
动作	把一系列候选文档交给用户,并记录点击数据	选择一个文档,将其从候选文档集取出,附加(append)到已排序文档序列末端,并更新相关参数
回报	每一个排位的收益(用户是否有点击)	这一步决策带来的边际信息效益
回报来源	真实用户数据	使用多样化指标模拟

在 TREC Web Track 数据集上进行的实际测评中,由于 MDP-DIV 模型的回报函数可以选取不同的评价指标,其在不同测试中也有着不同性能表现,例如对 ERR-IA@5 指标,使用 α-DCG 作为回报函数的性能就要比 S-recall 稍好,但两种回报函数的 MDP-DIV 模型性能均领先于 xQuAD 和 PM-2 等启发式模型,也领先于 R-LTR 和 PAMM 等学习式模型.

最近在 SIGIR2018 上,Feng 等人又对 MDP-DIV 模型进行了进一步的改进,提出了 M2DIV 模型[44].典型的搜索结果多样化排序过程往往会使用贪心算法选择文档以实现多样化排序,即每一步都选择当前最能提升已选文档序列多样化性能的候选文档.上文所述的许多常见的算法,包括启发式的 xQuAD、PM-2 等,和学习式的 R-LTR 以及 MDP-DIV 本身,都采用的是贪心选择策略.在线上的学习过程中,贪心选择策略可以有效地简化问题,提高效率.

但是将贪心选择策略应用于多样化排序,存在着一个明显的问题:贪心算法在对问题求解时,每一次都会求出针对当前子问题的局部最优解,它能够得到全局最优解的条件是,所求解的子问题具有无后效性.具体到多样化排序,贪心算法能够得出全局最优解的条件就是文档与文档之间彼此无关,但这一假设是不成立的.因为用户阅读不同文档的信息效益并不是彼此独立的,在多样化问题中,对某一个排位的文档选择必然会对其他候选文档的信息效益产生影响,而与此同时,对 N 个文档进行排序共存在 $N!$ 种不同的情况,因此在实际应用中显然不可能直接对整个排序空间进行搜索.综上所述,对于多样化排序问题,贪心算法从理论上即不可能得到全局最优解.

为了解决这一问题,M2DIV 模型应运而生,M2 中的第一个 M 表示 MDP(与 MDP-DIV 模型相同),第二个 M 表示蒙特卡洛树搜索(MCTS).受到 AlphaGo[45] 和 AlphaGo Zero[46] 在围棋博弈中所获成功的启发,Feng 等人对 MDP-DIV 模型进行了进一步的改进,在文档选择的过程中,通过 MCTS 取代了贪心算法,实现对整个排序空间的启发式搜索,进而实现比贪心算法更优化的决策过程.

M2DIV 进行 MDP 的过程与 MDP-DIV 基本相同,此处不再赘述,接下来简要介绍 MCTS 的过程:从搜索树的根节点 R 出发,MCTS 递归地选择当前最佳的子节点,直到找到叶子结点 L;如果 L 不是一个终止节点,那就为 L 建立一个或更多子节点,根据预测策略选择一个最优子节点 C;接下来从 C 出发进行模拟和评估,对于 AlphaGo Zero,此处即为模拟对弈的结果;模拟完成之后,对 C 的所有直接和间接上级节点进行反向传播,更新其储存的值.

与 MDP-DIV 相比,M2DIV 的状态和动作定义等都基本相同,主要改变了策略 p 和价值函数 V 的定义.相对于 MDP-DIV 直接使用多样化指标作为价值函数,M2DIV 使用了长短期记忆(LSTM),以向量化的起始状态 s 作为输入,并将价值函数定义为 LSTM 权重和的非线性变换.LSTM 的相关参数在训练中学习.在 MDP 执行的过程中,MDP-DIV 使用的策略是基于贪心算法的选择策略,如上文所述,无法得到全局最优解.为了解决这个问题,M2DIV 使用了一个新的基于 MCTS 的搜索过程来对策略 π 进行增强——基于 MCTS 的搜索策略 π 可以对整个排序空间进行搜索,进而获得比贪心算法更优化的全局解.

在每一次迭代中,M2DIV 模型的 MCTS 搜索策略将迭代进行 K 次搜索,从根节点 s_R 出发,当搜索到叶子节点 s_L 时(每个节点与一个状态对应),若当前节点可以继续向下扩展就使用价值函数 V 进行评估,然后向下扩展一个节点,否则就使用人工标注的标签配合预定义的性能评测指标(此项只针对训练过程).当整个训练过程结束时执行反向传播,更新搜索过程中经过的所有节点和边——在

M2DIV 定义的 MCTS 过程中，搜索树的边 $e(s,a)$（从状态 s 指向状态 $T(s,a)$）储存了动作的价值 $Q(s,a)$，访问计数 $N(s,a)$ 和优先概率 $P(s,a)$. 当执行完 K 次迭代之后，对于根节点的搜索策略 π 即可直接通过搜索树中各条边储存的 N 计算得到，而模型需要的参数均通过强化学习训练得到.

在训练过程中，M2DIV 模型的参数均通过执行 MCTS 搜索过程得到，而在实际线上运行的过程中，由于 MCTS 需要额外的时间开销，因此在线上运行中根据实际情况可以选择直接使用已经训练好的策略 p，而不需要再进行搜索. 基于 TREC Web Track 数据集的测试结果显示，M2DIV 在 MDP-DIV 的基础上，实现了进一步的性能提升，也同样优于上文所述的启发式和学习式模型. 测试中发现，使用已经训练好的模型参数，而在线上决策过程中使用训练好的策略 p 而不再进行 MCTS 搜索的“无 MCTS 版”M2DIV 模型，其性能会比线上运行时也进行 MCTS 搜索的 M2DIV 模型稍逊，但依然显著领先于包括 MDP-DIV 在内的其他模型. 这是因为 MCTS 在训练的过程中即可发挥效用，学习到更加准确的模型参数，进而提高策略函数 p 的性能. 此处需要说明的是，尽管该模型的训练使用了子话题信息作为训练参数，但是必须强调若按照是否显式考虑子话题进行分类，则该模型仍然是隐式模型，子话题信息的作用仅仅是用于在每一步中对回报函数（即多样化评价指标）进行计算. 因此，一个可能的改进方向即为在 MDP 的策略中显式考虑子话题信息.

3.5　搜索结果个性多样化

如上文所述，搜索结果多样化是基于以下原则进行的：用户往往会更加倾向于使用简单直白的文字描述他们需要的信息. 这就意味着，用户向搜索引擎提交的查询本身代表的用户意图具有歧义性和不确定性.

而搜索结果多样化就是要在“用户意图是有歧义且不确定性的”这一基本假设成立且不改变的前提条件下，尽可能多地覆盖与查询相关的实体和这些实体的不同方面，使得返回的搜索结果能够命中用户需求. 与之对应的另一个重要的研究领域，则是搜索结果的个性化（Personalization）[47]. 跟搜索结果的多样化一样，搜索结果的个性化要解决的同样是用户查询的歧义性和不确定性问题. 但是在解决问题的思路上，搜索结果的个性化与搜索结果的多样化存在一些方向上的差别.

面对用户意图的不确定性，搜索结果个性化的目标则是直接消除这种不确定性本身. 搜索结果个性化的信息来源并不仅仅局限于用户提交的查询本身，还包括其他的多种信息来源，例如来自用户的会话反馈、用户画像、社交链接，以及其他的用户上下文和兴趣模型等. 通过这些数据，个性化系统可以对用户的真实意图进行进一步的判断，直接减少乃至消除用户意图的不确定性. 简而言之，多样化系统接受并适应用户意图的不确定性，而个性化系统则尝试增强系统对于用户的认识以克服不确定性. 与尽可能地满足更多的用户不同，个性化的目标则是对用户本身的特点和偏好进行识别，进而最大化特定用户的满意程度.

表面上看，搜索结果的个性化和多样化仿佛是矛盾的，因为个性化和多样化本身的目标似乎是互斥的. 如果用户提交查询的歧义性和不确定性被个性化方法消除了，那么再对搜索结果进行多样化似乎就没有意义了. 然而实际上，目前学者们已经发现，多样化和个性化并不是必然不兼容或者互斥的，实际上它们本身不仅可以共存，而且还可以互相增强. 此前已经有学者提出，应用搜索结果多样化技术来改善搜索结果个性化的效果，最早由 Radlinski 和 Dumais[48] 提出对短查询重构来引入多样化，对搜索引擎返回的前 N 个结果进行重排，然后再使用用户信息进行个性化. 随后，Vallet 和 Castells[49] 在 SIGIR2012 上首先提出了搜索结果个性多样化（Personalized Diversification）的概念.

这一概念提出的背景是，个性化搜索相效于普通搜索，实际上不确定程度要更高. 对于输入的个性化信息的解释是搜索结果个性化任务中显著而固有的部分，也是困难的部分. 事实上，相同的个性化上下文信息可能会跟几种不同的解释相对应，这跟多样化问题面临的情况是一致的. 对个性化的系统假设进行多样化，可以降低猜错用户兴趣的概率. 进一步地讲，搜索系统可以接触到的用户行为与兴趣信息通常只覆盖了全部用户信息的一小部分，因为对于上下文的解释要远远比用户提出的查询请求自身更加不完整. 另外用户的意图通常不是均匀分布的，所以对于用户信息不同方面的多样化跟搜索结果多样化中对于查询不同方面的多样化，本质上是相似的.

Vallet 和 Castells 首先在已有的多样化方法的基础上添加了一种个性化成分，实现了对于多样化方法的泛化，并建立了一种通用的框架以实现对于搜索

个性化信息的多样化.个性多样化模型是在已有的多样化模型(如 xQuAD,IA-Select 等)的基础上,引入代表用户的随机变量 u,假设 q 和 u 彼此独立,在模型原有的多样化部分加入新的子项 $p(d|u)$,用以衡量用户 u 对文档 d 表示喜爱的概率,即用户选择某一文档的可能性.这个指标可以由基于某些用户喜好学习系统的个性化系统提供.

实验结果显示,对于个性化的 IA-Select 和 xQuAD 框架,相对于原始框架,引入个性化变量可以提高 3%～11%的精确性指标,以及 3%～8%的多样性指标.这一结果证明了即使是在引入个性化用户信息的情况下,对搜索结果进行多样化仍然具有正面的意义.

Vallet 和 Castells 的成果更加有力地证明了搜索结果多样化与个性化之间存在的内在联系,两者并不是彼此孤立的,并没有彼此不相关的实现目标.在这些成果的基础上,Liang 和 Ren 等人[50]提出了一种新的基于结构化监督式学习的搜索结果个性多样化方法.

在搜索结果个性多样化的过程中引入监督式学习,具有一系列显著的优势:监督式学习模型可以充分地利用搜索引擎运行的过程中所积累的一系列标注过的训练数据和相关的一系列优化方法,对于提高模型的性能具有非常显著的意义.在这种监督式学习方法中,首先将搜索结果个性多样化的任务形式化为对于给定的特定用户和查询预测一个多样化文档集的预测任务.在此基础上,使用结构化 SVM 框架(SSVMs)对模型进行训练.训练的主要步骤是,首先提出一个跟用户兴趣相关的类 LDA 的话题模型,通过它可以为每个文档推测出一个多项式的话题分布,并决定一个文档是否满足特定的用户.接下来在训练的过程中,可以使用通过字段统计信息、非监督式个性多样化算法和话题模型中直接抽取的特征.

此外,在 SSVMs 中,额外定义了两类限制,以确保搜索结果具有多样性并与用户兴趣相关.性能测试结果证明,基于 SSVMs 的监督式学习个性多样化算法,在性能上领先于此前已有的最佳的非监督式个性多样化、单纯的多样化和单纯的个性化算法.附加的两个限制在监督式学习中起到了重要的作用.Liang 等人的研究成果证明了监督式学习方法在搜索结果个性多样化方面具有显著的发展潜力,后续研究中可以尝试引入更多的监督式学习方法.

4 搜索结果多样化的主要评价指标

在信息检索领域,传统的评价指标主要关注搜索结果的相关性,即返回的结果是否与用户提交的查询相关.而在搜索结果多样化的问题中,不仅要评价结果的相关性,同时也要对整个结果列表的多样性进行评价.将多样性纳入考虑之后,当前学者们对于信息检索系统的研究将主要着重于对搜索引擎的以下能力进行改进:

(1)寻找与查询相关文档的能力;

(2)对相关文档进行排序的能力;

(3)结果文档集满足多样化需求的能力;

(4)对结果文档集进行多样化排序的能力.

对于(1)和(3)的衡量相对比较简单,直接使用传统的准确率(Precision)和召回率(Recall)来衡量即可.进一步地,还可以在通用的指标基础上,发展出针对子话题的准确率与召回率指标.Zhai 等人[10]提出的 S-recall、S-precision 和 WS-precision 方法分别计算了子话题的召回率、准确率和冗余情况,这些是最早出现的关注子话题的搜索结果评价方法.

但是,这些指标都只关注了文档集的覆盖程度,完全没有衡量排序效果.后来学者们在传统的评价指标的基础上进行了拓展以评价搜索结果多样化排序效果.诸如 MAP、DCG、ERR 等指标都会受到排序结果和文档相关性的影响,在这些指标的基础上进行发展,可以进一步得到 MAP-IA,α-NDCG 和 ERR-IA 等指标,这些指标可以同时反映文档集的多样性和文档排序效果.

基于排序的多样性指标,诸如 MAP-IA、α-NDCG 和 ERR-IA 合并了相关性、多样性和结果排序,它们都可以作为优化多样化系统效率的指标.实际上,为了尽可能提高对系统效率衡量的精确度,在模型的测试中一般会同时衡量多个主要指标.

接下来,本文将对以下几种主流的评价指标进行较为详细的介绍:

(1) MAP-IA 和其他意图敏感(Intent-Aware)指标

MAP(Mean Average Precision)[51]是一个在索引检索中常用的评测指标.如上所述,精度(Precision)只考虑到了结果文档集中相关文档的个数,并没有考虑到文档集的排序.显然搜索引擎返回的结果列表一定是有序的,并且相关文档排序越靠前越好.在此基础上,发展了平均精度(Average Precision)的概念,其定义可以用以下公式表示:

$$AP=\frac{\sum_{j=1}^{k}P(j)\times Relevance(j)}{\sum_{j=1}^{k}Relevance(j)} \tag{10}$$

式中 j 表示文档排位，$Relevance(j)$表示对应文档的相关性，然后 $P(j)=\sum_{i=1}^{j}Relevance(i)$. 对MAP，则是对所有查询的 AP 值求平均数.

包括 MAP 在内的经典信息检索指标，主要集中关注文档相关性，并没有考虑到用户意图. 假设有查询 q 分别属于两类子话题 c_1 和 c_2，并有 $P(q|c_2)\gg P(q|c_1)$. 现在有两个文档，d_1 只与 c_1 相关性极佳，而 d_2 只与 c_2 相关性较好(不如 d_1 对 c_1)，传统评价指标会倾向于给(d_1，d_2)较高的评分，但是大部分用户通常会觉得(d_2，d_1)的顺序更具有实用性. 因此，对搜索结果的评价需要考虑到用户意图.

Agrawal 等人[9]提出了新指标 MAP-IA，在改良了传统 MAP 方法的基础上，增加了用户意图敏感(Intent-Aware，IA)的成分. 对于给定的子话题分布 $P(c|q)$和各个文档的子话题标签，可以计算特定结果序列的 MAP 期望值. 对于查询所包含的每一个用户意图，可以将前 k 个文档中每一个与用户意图不匹配的文档标记为不好的结果，并计算出其与用户意图独立的 MAP(Q，$k|c$)，求其平均值即可得到 MAP-IA(Q，k)，公式表述如下：

$$\text{MAP-IA}(Q,k)=\sum_{c}P(c|q)\text{MAP}(Q,k|c) \tag{11}$$

与之类似的还有另外两种评价指标 nDCG-IA 和 MRR-IA.

(2) α-nDCG

DCG(Discounted Cumulative Gain)[51]是信息检索领域一个经典的评价指标. 对于给定文档集 Q 和它上面一个理想的排序 R，前 k 位对应的 DCG 值计算如下所示：

$$\text{DCG}(Q,k)=\sum_{j=1}^{k}\frac{2^{r(j)}-1}{\log(1+j)} \tag{12}$$

此处 r 表示对 Q 中第 j 位的评价，理想排序集合 R 包括所有对给定查询按评分从大到小排序的文档. 将其引入作为正规化指标，即为 NDCG(Normalized Discounted Cumulative Gain)的定义：

$$\text{NDCG}(Q,k)=\frac{\text{DCG}(Q,k)}{\text{DCG}(R,k)} \tag{13}$$

Clark 等人[52]在 DCG 基础上进行拓展得到了新的评价方法 α-NDCG 该方法主要倾向于奖励还有新发现的子话题的文档，而对子话题缺乏新颖性的文档进行惩罚. 这里 α 表示对冗余的子话题进行惩罚的变量因子.

(3) ERR-IA

ERR(Expected Reciprocal Rank)首先由 Chapelle 等人提出[53]. RR(Reciprocal Rank)是指第一个正确答案的排名的倒数，对其取均值，就可以得到 MRR，公式如下：

$$\text{MRR}=\frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{14}$$

表达式中，$rank_i$表示第 i 个查询语句的第一个正确答案的排名.

之前的模型只考虑了位置自身的价值信息与位置上文档的相关信息，并没有考虑文档之间的相关性信息. Chapelle 等人提出了新的模型 Cascade Models，它假设用户从排名由高到底依次查看文档，一旦文档满足了用户的需求，则停止查看后续的文档.

这里用 R_i表示用户只查看在位置 i 上的文档就满足了需求的概率，文档相关度越高，R_i越大，此处 R_i可以设置为任何关于文档相关度的函数. ERR 即为用户需求被满足时停止的位置的倒数的期望.

用户在位置 r 停止的概率公式如下：

$$PP_r=\prod_{i=1}^{r-1}(1-R_i)R_r \tag{15}$$

ERR 的计算公式如下所示，此处 $\varphi(r)$可以使用任何关于 r 的函数：

$$\text{ERR}=\sum_{r=1}^{n}\varphi(r)PP_r \tag{16}$$

与 MAP-IA 类似，在 ERR 的基础上，单独计算各个子话题的 ERR 值，然后计算其加权平均值，即为 ERR-IA.

(4) I-rec

意图召回率(Intent Recall，I-rec)[54]，有时也叫子话题召回率[55]，表示一个搜索结果排序中，不同的用户意图所占的比率. 其详细定义如下：令 d_r表示第 r 位的文档，$I(d_r)$表示跟文档 d_r相关的用户意图集合，于是对于特定截断长度 K 的 I-rec 可以使用如下公式计算：

$$I\text{-rec}@K=\frac{\left|\bigcup_{r=1}^{K}I(d_r)\right|}{|\{i\}|} \tag{17}$$

I-rec 的核心思想在于奖励占比较低的用户意图，本身并不考虑相关文档的位置，也无法处理意图概率和相关性级别的评估.

(5) D-指标(D-measure)

Sakai 等人[54]提出，设 $g_i(r)$为第 r 位的文档对用户意图 i 获得的增益(Gain)评分，它通过对每个用户意图的评估而获得，由此可以定义第 r 位文档

全局增益(Global Gain)评分：

$$GG(r)=\sum_{i\in\{i\}}\Pr(i|q)g_i(r) \tag{18}$$

将此全局增益评分，用于替换其他评价指标中的文档增益，即可获得评价指标 D-measure. 例如使用全局增益结合 nDCG 指标，即可得到：

$$D\text{-nDCG@}K=\frac{\sum_{r=1}^{K}GG(r)/\log(r+1)}{\sum_{r=1}^{K}CG^{*}(r)/\log(r+1)} \tag{19}$$

类似地，令 $J(r)=1$ 表示第 r 位文档与$\{i\}$中的任意用户意图 i 相关，反之则以 $J(r)=0$ 表示. 令 $C(r)=\sum_{k=1}^{r}J(k)$表示前 r 位相关文档总数，由此可定义 D-Q：

$$D\text{-}Q@K=\frac{1}{\min(K,R)}\sum_{r=1}^{K}J(r)\frac{C(r)+\beta CCG(r)}{r+\beta CCG^{*}(r)} \tag{20}$$

(6) $D\sharp$-指标($D\sharp$-measure)

$D\sharp$-指标[54]专注于对用户意图召回率进行增强，侧重于奖励跟热门意图高度相关的文档，其定义为 D-measure 和 I-rec 的线性组合. 这一指标在 TREC Web Track 或 NII Testbeds 中的几个任务中得到了广泛地应用.

5 搜索结果多样化实验结果与分析

下面将使用上文介绍的指标，对几种多样化模型的实验结果进行对比分析. 本文介绍的一部分评测结果，来自于已发表的相关文献[16]. 此外，我们还使用相同的数据集，额外做了一些针对 MDP-DIV 模型(使用 α-DCG 指标)的对照实验，由于时间仓促，M2DIV 模型的对照实验数据没有包括在内. 参照 Hu 等人[11]的相关工作，这里使用的数据集来自 TREC Web Track 2009 到 2012，一共包括 198 个带有多样性判断的查询，每一个查询都包含 3 到 8 个人工标注的子话题，相关性评分以二分形式在子话题层面给出，所有相关实验均在 ClueWeb09① 数据集上完成. 实验使用的子话题来自于 Google 搜索提供的查询建议，并简单地认为所有子话题权重相等.

对于搜索结果多样化评测，可以使用的数据集包括 TREC Web Track 和 NTCIR Imine Tasks 等公开数据集. 相对于其他信息检索领域，由于搜索结果多样化的评测要求标注数据给出详细的子话题覆盖信息，因此可用的数据集相对较少. 在实践中，由于查询总数较多，且子话题信息标注较为完备，大部分学者都倾向于使用 TREC Web Track 数据集.

Web Track 官方使用的多样性评价指标包括 ERR-IA、α-NDCG 和 NRBP[56]，因此实验中也使用这些指标. 以上指标通过显式地奖励新颖性和惩罚冗余性来评价多样性. 此外，实验结果中同样包括了 NTCIR Intent 与 Imine 任务所使用的 $D\sharp$-指标，以及传统的多样性指标 Pre-IA 和 S-rec.

实验使用已有文献[11]中使用的非多样化搜索结果作为基线模型，该结果为 Lemur 服务提供的基于语言模型搜索到的结果，标记为 Lemur. 参与评价的模型包括典型的显式启发式模型(如 xQuAD)及其相关衍生模型，典型的隐式学习式模型(如 R-LTR 和 PAMM)，以及显式学习式模型(DSSA). 结果如表 6 所示.

表 6 主要多样化模型性能评测

模型	ERR-IA	α-NDCG	NRBP	Pre-IA	S-rec
Lemur	0.271	0.369	0.323	0.153	0.621
xQuAD	0.317	0.413	0.284	0.161	0.622
PM2	0.306	0.411	0.267	0.169	0.643
HxQuAD	0.326	0.421	0.294	0.158	0.629
HPM2	0.317	0.420	0.297	0.172	0.645
R-LTR	0.303	0.403	0.267	0.164	0.631
PAMM	0.309	0.411	0.271	0.168	0.643
PAMM-NTN	0.311	0.417	0.272	0.170	0.648
DSSA	**0.356**	**0.456**	**0.326**	0.185	0.649
MDP-DIV(α-DCG)	0.329	0.451	0.286	**0.226**	**0.676**

可以看到，相对于未经多样化的基线成绩，主要的多样化模型都具有明显的性能优势，这证明了多样化排序对于改善搜索结果质量的重大意义.

两种显式启发式模型的代表 xQuAD 和 PM2 在不同的评测指标上互有优劣，而加入了用户意图分层机制之后，HxQuAD 和 HPM2 相较于原有模型都有着明显的性能提升，这证明了子话题按层次划分对于多样化性能的积极意义. 例如，对于某个一级子话题 q_1 拥有较多的二级子话题，而另一个一级子话题 q_2 下属的二级子话题较少的话，那么不经分层的显式多样化模型就可能会更加偏向于 q_1，缺乏对 q_2 的覆盖，进而导致搜索结果整体多样性的削弱. 用户意图分层机制在解决这个问题上有较为明显的效果.

从评价指标上来看，R-LTR、PAMM 和 PAMM-NTN 等学习式模型相对于基线成绩，同样具备很大的优势. 理论上讲，由于搜索引擎可以进行在线实时学习，通过机器学习模型抽取多样化特征并直接针对多样化评测指标进行训练的学习式模型应当显著

① http://boston.lti.cs.cmu.edu/Data/clueweb09/

优于启发式模型.但由于此类模型属于隐式学习式模型,其实际指标与 PM2 和 xQuAD 等显式启发式模型互有优劣,所以并没有明显的优势.

上述结果一方面实际证明了监督式学习方法自动抽取最优化的多样化指标,确实具备理论上的优越性,另一方面也显示出,隐式模型由于无法充分地利用子话题信息,在多样化排序中具备较大的局限性.

如果能将显式启发式模型与监督式学习相结合,则应当可以取长补短,充分发挥机器学习模型的潜能.而 DSSA 的表现也证明了这一点,其作为显式学习式模型,相对于此前的所有显式启发式模型和隐式监督式学习模型在性能上都具有非常明显的优势,而 MDP-DIV(α-DCG)则与 DSSA 在不同指标上互有胜负,同样领先于其他所有模型.

实验结果表明,在显式多样化模型的基础上引入机器学习方法,在搜索结果多样化领域有着巨大的优势和发展前景.需要说明的是,对于同类型的模型,学习式模型全面优于启发式模型,显式多样化模型全面优于隐式多样化模型,但是性能优势是建立在引入额外的外部数据的基础上的.例如,显式多样化模型需要查询对应的子话题信息,学习式模型在投入线上运营之前需要首先投入额外的计算资源,使用人工标注数据进行预训练,且必须经过充分的可靠性验证才能投入商业运营.因此,对于实际运行的搜索系统,具体使用哪种多样化模型,应结合自身实际运营条件具体分析,不能一概而论.

MDP-DIV 模型的表现,则证明了强化学习在信息检索,尤其是多样化问题上的巨大潜力.由于蒙特卡洛树搜索在训练过程中需要非常长的训练时间,我们没有对 M2DIV 模型进行验证,而相关文献已经表明 M2DIV 模型相比 MDP-DIV 模型可以进一步提高性能.这里需要再一次强调,MDP-DIV 模型和 M2DIV 模型,按照是否显式考虑子话题来分类的话,并不完全属于隐式多样化模型,在模型训练的过程中也需要使用到子话题信息.尽管这些子话题信息并没有显式地参与到决策过程中,但是它们被用于计算回报(即选定的多样化指标差值),相当于以回报的形式间接地对决策过程进行了干预.如果参考 DSSA,将子话题信息直接融入到强化学习的决策过程中,应该可以进一步地提升模型性能.

6　搜索结果多样化在实际部署的大型分布式系统中的性能评估

如上文所述,针对搜索结果多样化问题,已经有许多学者提出了相应的模型以对搜索引擎获得的搜索结果进行多样化重排,但是值得注意的是,此前并没有相关研究实际地考虑如何将多样化系统应用在现实中的大型搜索系统上.在竞争激烈的市场上,商用搜索引擎必须拥有足够高的搜索效率才有可能获得商业上的成功.换而言之,针对搜索结果多样化模型的性能考量,不仅应当面向其多样化性能,同样也不应忽视多样化算法本身的运行效率.

此前 Deng 等人[57]已经对多样化算法本身进行了形式化定义和静态时间复杂度分析,但单机运行的多样化模型性能并不能代表线上搜索时的实际性能.现代搜索引擎对计算机系统的计算能力和储存能力都提出了极高要求,因此它们普遍搭建在大型分布式系统上,为了保证能及时处理搜索请求,这类系统往往规模巨大,包含成千上万个结点.应用于搜索引擎的主流分布式系统有基于文档的分区(document-based partitioning)和基于词项的分区,通常倾向于使用前者,因为其具有更加节省资源和负载更加均衡的优势[58].

对于基于文档的分区架构,执行查询的过程可以高度并行化——查询任务可以被分割到代理(broker)节点和搜索节点中,代理节点负责将查询转发到搜索集群中的搜索节点中,然后由搜索节点查询和汇总文档索引,返还给代理节点.最后代理节点汇总并进行全局排序之后将结果返还给用户端.

在 2016 年,Naini 等人[59]第一次对实际运行在分布式系统中的线上搜索多样化问题进行了相关研究.他们首先对 Zucoon 等人[60]提出的使用贪心局部搜索(GLS)以求解最佳选址问题进行多样化排序的算法,在实际系统上的运行效率进行了评估,引入两轮聚类改善了算法的运行效率,然后针对目前线上搜索使用的分布式系统的实际情况,提出了基于代理和基于节点的两种不同的分布式多样化策略.

Zucoon 等人于 2012 年提出的算法,是将搜索结果多样化视作一个最佳寻址问题,优化目标是让每一个子话题尽可能地“接近”每一个排位上的目标文档,并针对这一问题提出了一个基于贪心局部搜索(GLS)的算法.这一策略可以与已有的隐式多样化模型,例如 MMR 等结合,进一步提高多样化性能.但是在实际的应用中,要对 N 个文档进行多样化排序找出最佳的前 k 个结果,该算法每进行一轮迭代就有高达 $O(N^2k^2)$ 的时间复杂度,其运算效率远不足以满足线上搜索的需要,更糟糕的是,这一时间复杂度是建立在任意两个文档的距离 $w(d_i,d_j)$(即两个文档间的相似度)都已知的前提条件上,在

线上运行的条件下,这意味着更多的内存和缓存消耗.综上所述,该算法在实际应用中开销过高.

Naini 等人提出对该算法进行运行效率上的优化,他们首先使用 K-means 方法对候选文档进行了聚类,然后直接将目标函数中的 $w(d_i,d_j)$ 改为 $w(d,ctr(c))$,即储存并计算所有文档到各自聚类中心的距离.通过这一方法,对目标函数进行计算的时间复杂度从 $O(k\cdot N)$ 下降到了 $O(k^2)$(通常 k 远小于 N).在此基础上再一次对结果文档先按相关性聚类,再从每一类中选择文档,得到候选文档集 $TopC$,最后实现的算法 C^2-GLS,其总和时间复杂度为 $O(Rk^3)$,此处 $R=|TopC|$ 且 $R\ll N$,且存储规模也从 $O(N^2)$ 下降到了 $O(Nk)$.

接下来他们在此基础上对搜索结果多样化在分布式系统上的架构设计进行了研究(此前并没有这类研究成果),提出了两种不同的基于分布式系统的多样化策略:基于代理(broker-based)的多样化方法,和基于节点(node-based)的多样化方法.前者的流程是:首先各个搜索节点获得查询结果之后,先进行局部相关性排序,然后汇总给执行查询的代理节点,代理节点进行归并排序之后再执行多样化,并返回结果;后者的流程则是各个搜索节点在局部排序的时候就直接进行多样化,返回给代理节点之后排序并交给用户端.需要说明的是,对于相同的查询,两种策略的多样化结果可能会有所不同.

在 TREC 数据集上针对算法的测试证明,聚类并不能提高 GLS 的多样化性能,但适当的预处理可以大幅提高多样化算法的运行效率,C^2-GLS 相对 GLS 总运行时间减少了近 90%.这里需要说明的是,相对于隐式多样化方法,显式多样化方法具有全方位的优势,xQuAD 不仅性能上显著优于 GLS 系列算法,且运行时间也比原始 GLS 少 99%.但是由于显式多样化模型需要额外的子话题信息,实际线上运营的模型不一定能满足显式多样化条件,所以对 GLS 的算法效率改进仍然具有相当程度的意义.针对两种策略的测试结果显示,在多样化性能方面,由于基于代理的多样化方法更容易获得全局最优解,所以相比基于节点的多样化方法有更好的性能;另一方面,当返回结果总量非常大的时候,基于节点的方法在网络开销上有着较为明显的优势.

7 未来研究方向

本文介绍了搜索结果多样化的背景和需求,探讨了相关定义,进而对目前主流的搜索结果多样化方法、评价指标与性能做了介绍.总结目前已有的相关模型,我们认为今后对搜索结果多样化的研究,将主要围绕以下几个方向展开:

(1)进一步拓展子话题的来源,发掘一种可以方便地挖掘高质量子话题的方法.虽然学者们已经提出了很多种挖掘子话题的方法,但是由于种种原因,在实际研究中使用较多的仍然是诸如 Google Suggestions 之类的来自于搜索引擎查询建议的挖掘方法,这使得挖掘出的子话题质量难以进一步提升.因此,有必要研究如何高效率地挖掘高质量的子话题信息;

(2)搜索结果多样化与其他研究方向的融合,将其他领域的研究成果带入多样化领域.目前,启发式方法和学习式方法(包括监督式学习和强化学习)的研究进度之间仍然存在着明显的断层,相对于启发式方法,学习式方法对于子话题和用户意图的衡量仍然处在明显的落后状态,除了 DSSA 之外鲜有相关报道.今后可以尝试将不同种类的多样化模型融合以实现取长补短的效果,例如可以将 DSSA 模型与 M2DIV 模型相融合,用显式的子话题信息改善 M2DIV 模型的策略部分.此外,深度学习、强化学习、生成对抗网络(GAN)等新的机器学习技术近几年来的快速发展,已经在不同的领域产生影响,诸如深度文本匹配等技术已经在信息检索领域得到了广泛应用.但目前大部分应用仍然局限在搜索结果的相关性排序,如何将这些应用拓展到多样性排序的领域,有待进一步的研究;

(3)如何将多样化与个性化结合,在个性化搜索系统中引入多样化.已有的一些研究成果已经证明,多样化和个性化方法虽然看起来南辕北辙,但实际上两者之间并没有不可逾越的界限.多样化与个性化并非是对立的互斥关系,而是可以相互协同,相互促进,将多样化与个性化方法结合起来,在明确用户意图这一根本目的上,是可以取得优于单纯多样化或者单纯个性化的效果的;

(4)将多样化模型引入到推荐系统中.事实上,多样化模型不仅可以用于搜索引擎,推荐系统中同样也有必要引入多样化排序.对于基于信息流的媒体/社交平台,例如今日头条、知乎等,由于存在新注册用户或者游客,因此其推荐系统并非在所有条件下都有可依赖的个性化信息,而且过度依赖用户点击数据等信息会让用户的时间线过度单一,给用户体验带来负面影响.因此,有必要在推荐系统中引入多样化,通过提升用户时间线多样性以提升用户体验;

(5)针对短文本流等其他类型搜索结果的多样

化研究.传统的多样化模型针对的是静态长文本,而对于 Twitter、微博、京东商城等网站的搜索系统而言,其搜索结果的形式是短文本流(如微博等)和命名实体(购物网站的商品)等其他形式的信息,传统多样化模型针对此类搜索结果的多样化效果往往不尽如人意,需要对此做出相应的改进.

需要注意的是,目前对搜索结果进行多样化的方法,大部分都是通过对结果文档集进行重新排序而实现的.这也就暴露了一个问题:如果结果文档集本身就缺乏多样性,对子话题的覆盖程度不高的话,那么这类方法的效率就会很低.最后,搜索结果多样化的根本目的是在实际运行的信息检索系统中提高搜索结果的信息量,改善用户体验,实际运行的信息检索系统往往跟实验环境有很大的区别.在进一步探索和改进多样化模型时,不仅应当关注模型的多样化性能本身,也应当关注模型在线上搜索系统中的运行性能和可用性.如何衡量多样化模型在线上搜索系统中的实际运行效果,让多样化模型在真实生产环境下能更加充分地发挥作用,仍需要进一步的研究.

参　考　文　献

[1] Silverstein C, Marais H, Henzinger M, Moricz M. Analysis of a very large Web search engine query log. ACM SIGIR Forum, 1999, 33(1): 6-12

[2] Robertson S E. The probability ranking principle in IR. Journal of Documentation, 1977, 33(4): 294-304

[3] Goffman W. On relevance as a measure. Information Storage and Retrieval, 1964, 2(3): 201-203

[4] Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries //Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia, 1998: 335-336

[5] Drosou M, Pitoura E. Search result diversification. ACM SIGMOD Record, 2010, 39(1): 41-47

[6] Santos R L T, Macdonald C, Ounis I. Exploiting query reformulations for Web search result diversification// Proceedings of the 19th International Conference on World Wide Web. Raleigh, USA, 2010: 881-890

[7] Radlinski F, Bennett P N, Carterette B, Joachims T. Redundancy, diversity and interdependent document relevance. ACM SIGIR Forum, 2009, 43(2): 46-52

[8] Santos R L T, Peng J, Macdonald C, Ounis I. Explicit search result diversification through sub-queries//Gurrin C, He Y, Kazai G, et al, eds. Proceedings of the 32nd European Conference on Advances in Information Retrieval (ECIR'2010). Berlin, Germany, 2010: 87-99

[9] Agrawal R, Gollapudi S, Halverson A, Ieong S. Diversifying search results//Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (WSDM'09). Barcelona, Spain, 2009: 5-14

[10] Dang V, Croft B W. Term level search result diversification// Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13). Dublin, Ireland, 2013: 603-612

[11] Hu S, Dou Z, Wang X, et al. Search result diversification based on hierarchical intents//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM'15). New York, USA, 2015: 63-72

[12] Yue Y, Joachims T. Predicting diverse subsets using structural SVMs//Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland, 2008: 1224-1231

[13] Zhu Y, Lan Y, Guo J, et al. Learning for search result diversification//Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. Queensland, Australia, 2014: 293-302

[14] Xia L, Xu J, Lan Y, et al. Learning maximal marginal relevance model via directly optimizing diversity evaluation measures//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. Santiago, Chile, 2015: 113-122

[15] Xia L, Xu J, Lan Y, et al. Modeling document novelty with neural tensor network for search result diversification// Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. Pisa, Italy, 2016: 395-404

[16] Jiang Z, Wen J, Dou Z, et al. Learning to diversify search results via subtopic attention//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. Tokyo, Japan, 2017: 545-554

[17] Zhai C, Cohen W, Lafferty J. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval// Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Toronto, Canada, 2003: 10-17

[18] Zhai C, Lafferty J. A risk minimization framework for information retrieval. Information Processing and Management: An International Journal-Special Issue: Formal Methods for Information Retrieval, 2006, 42(1): 31-55

[19] Zhang B, Li H, Liu Y, et al. Improving Web search results using affinity graph//Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Salvador, Brazil, 2005: 504-511

[20] Zhu X, Goldberg A B, Gael J V, Andrzejewski D. Improving diversity in ranking using absorbing random walks//Proceedings of the North American Chapter of the Association for Computational Linguistics. Rochester, USA, 2007: 97-104

[21] Wang C, Lin Y, Tsai M, Chen H. Mining subtopics from different aspects for diversifying search results. Information Retrieval, 2013, 16(4): 452-483

[22] Dang V, Xue X, Croft W B. Inferring query aspects from reformulations using clustering//Proceedings of the 20th ACM International Conference on Information and Knowledge Management. Glasgow, UK, 2011: 2117-2120

[23] Li Z C, Chen F, Xing Q L. THUIR at TREC 2009 Web Track: Finding relevant and diverse results for large scale Web search//Proceedings of the 18th Text Retrieval Conference. Gaithersburg, USA, 2009: 145-153

[24] Rafiei D, Bharat K, Shukla A. Diversifying Web search results//Proceedings of the 19th International Conference on World Wide Web. Raleigh, USA, 2010: 781-790

[25] Yin D, Xue Z, Qi X, Davison B D. Diversifying search results with popular subtopics//Proceedings of the 18th Text Retrieval Conference. Gaithersburg, USA, 2009: 500-278

[26] Dou Z, Hu S, Chen K, et al. Multi-dimensional search result diversification//Proceedings of the 4th ACM International Conference on Web Search and Data Mining. Hong Kong, China, 2011: 475-484

[27] Zheng W, Fang H, Yao C. Exploiting concept hierarchy for result diversification//Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM'12). New York, USA, 2012: 1844-1848

[28] Nguyen T N, Kanhabua N. Leveraging dynamic query subtopics for time-aware search result diversification//Proceedings of the 36th European Conference on IR Research on Advances in Information Retrieval-Volume 8416. Amsterdam, The Netherlands, 2014: 222-234

[29] Dou Z, Hu S, Luo Y, et al. Finding dimensions for queries//Proceedings of the 20th ACM International Conference on Information and Knowledge Management. Glasgow, UK, 2011: 1311-1320

[30] Dang V, Croft W B. Diversity by proportionality: An election-based approach to search result diversification//Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. Portland, USA, 2012: 65-74

[31] Minack E, Siberski W, Nejdl W. Incremental diversification for very large sets: A streaming-based approach//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, 2011: 585-594

[32] Chen L, Cong G. Diversity-aware top-k publish/subscribe for text stream//Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. Melbourne, Australia, 2015: 347-362

[33] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation. The Journal of Machine Learning Research, 2003, 3: 993-1022

[34] Wang X, McCallum A. Topics over time: A non-Markov continuous-time model of topical trends//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 424-433

[35] Zhao X W, Jiang J, Song Y, et al. Topical keyphrase extraction from Twitter//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Portland, USA, 2011: 379-388

[36] Liang S, Yilmaz E, Shen H, et al. Search result diversification in short text streams. ACM Transactions on Information Systems, 2017, 36(1): 1-35

[37] Carterette B, Chandar P. Probabilistic models of ranking novel documents for faceted topic retrieval//Proceedings of the 18th ACM Conference on Information and Knowledge Management. Glasgow, UK, 2011: 1287-1296

[38] Santos R L T, Macdonald C, Ounis I. Intent-aware search result diversification//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, 2011: 595-604

[39] Xu Jun, Lan Yan-Yan. Diversification: A new direction of learning to rank. Communications of the CCF, 2016, 12(7): 50-52(in Chinese)
(徐君，兰艳艳. 多样化——排序学习发展的新方向. 中国计算机学会通讯，2016，12(7)：50-52)

[40] Li L, Zhou K, Xue G R, et al. Enhancing diversity, coverage and balance for summarization through structure learning//Proceedings of the 18th World Wide Web (WWW) Conference. Madrid, Spain, 2009: 71-80

[41] Clarke C L, Kolla M, Vechtomova O. An effectiveness measure for ambiguous and underspecified queries//Proceedings of the 2nd International Conference on the Theory of Information Retrieval. Cambridge, UK, 2009: 188-199

[42] Radlinski F, Kleinberg R, Joachims T. Learning diverse rankings with multi-armed bandits//Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland, 2008: 784-791

[43] Xia L, Xu J, Lan Y, et al. Adapting Markov decision process for search result diversification//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. Shinjuku, Japan, 2017: 535-544

[44] Feng Y, Xu J, Lan Y, et al. From greedy selection to exploratory decision-making: Diverse ranking with policy-value networks//Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. Ann Arbor Michigan, USA, 2018: 125-134

[45] Silver D, Huang A, Maddison C J, et al. Mastering the game of go with deep neural networks and Tree Search. Nature, 2016, 529(7587): 484-489

[46] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of go without human knowledge. Nature, 2017, 550(7676): 354-359

[47] Shen X, Tan B, Zhai C. Implicit user modeling for personalized search//Proceedings of the 20th ACM Conference on Information and Knowledge Management. Bremen, Germany, 2005: 824-831

[48] Radlinski F, Dumais S. Improving personalized Web search using result diversification//Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, USA, 2006: 691-692

[49] Vallet D, Castells P. Personalized diversification of search results//Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Portland, USA, 2012: 841-850

[50] Liang S, Ren Z, Rijke M. Personalized search result diversification via structured learning//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14). New York, USA, 2014: 751-760

[51] Manning C D, Raghavan P, Schutze H. Introduction to Information Retrieval. UK: Cambridge University Press, 2008

[52] Clarke C L A, et al. Novelty and diversity in information retrieval evaluation//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Singapore, 2008: 659-666

[53] Chapelle O, Metlzer D, Zhang Y, Grinspan P. Expected reciprocal rank for graded relevance//Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09). Hong Kong, China, 2009: 621-630

[54] Sakai T, Craswell N, Song R, et al. Simple evaluation metrics for diversified search results//Proceedings of the 3rd International Workshop on Evaluating Information Access. Tokyo, Japan, 2010: 42-50

[55] Zhai C X, Cohen W W, Lafferty J. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval//Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Toronto, Canada, 2003: 10-17

[56] Clarke C L A, Kolla M, Vechtomova O. An Effectiveness measure for ambiguous and underspecified queries//Proceedings of the 2nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory. Cambridge, UK, 2009: 188-190

[57] Deng T, Fan W. On the complexity of query result diversification. ACM Transactions on Database Systems, 2014, 39(2): 46

[58] Dean J. Challenges in building large-scale information retrieval systems: Invited talk//Proceedings of the 2nd ACM International Conference on Web Search and Data Mining. Barcelona, Spain, 2009: 1

[59] Naini K D, Altingovde I S, Siberski W. Scalable and efficient Web search result diversification. ACM Transactions on the Web, 2016, 10(3): 30

[60] Zuccon G, Azzopardi L, Zhang D, Wang J. Top-k retrieval using facility location analysis//Proceedings of the 34th European Conference on Advances in Information Retrieval. Barcelona, Spain, 2012: 305-316

DOU Zhi-Cheng, Ph. D., professor. His research interests include information retrieval, data mining, and big data analytics.

QIN Xu-Bo, Ph. D. candidate. His research interests include information retrieval and data mining.

WEN Ji-Rong, Ph. D., professor. His main research interests include information retrieval, database, big data analytics and data mining.

Background

The search engines in real search scenario often receive short queries issued by users, which are usually ambiguous or multi-faceted. In addition to being relevant to the query, the retrieved documents are expected to be as diverse as possible to satisfy different users' information needs as many as possible. Search result diversification has become an important part of the modern search engines, and many diversification methods are proposed to generate diverse results. Existing studies have shown that the explicit approaches, which model subtopics under the query, have better performance than those implicit approaches. While the supervised approaches can avoid heuristic and handcrafted functions and parameters rather than unsupervised approaches. The two ways of categorization are orthogonal, an approach can be both explicit or implicit, and supervised or unsupervised. In this paper, we present the formal definition of search result diversification tasks, and introduce the present approaches and measures of diversification. We also provide experimental results to compare the performance of those approaches.