

基于主题模型的深层网数据源选择算法

王秋月, 曹巍*, 史少晨

(中国人民大学 信息学院, 北京 100872)

(*通信作者电子邮箱 caowei@ruc.edu.cn)

摘要: 联邦搜索是从大规模深层网上获取信息的一种重要技术。给定一个用户查询, 联邦搜索系统需要解决的一个主要问题是数据源选择问题, 即从海量数据源中选出一组最有可能返回相关结果的数据源。现有的数据源选择算法大多基于数据源的样本文档集和查询之间的关键词匹配, 通常无法很好地解决少量样本文档的信息缺失问题。针对这一问题, 提出了基于隐含狄利克雷分布(LDA)主题模型进行数据源选择的方法。首先, 使用 LDA 主题模型获得数据源和查询的主题概率分布; 然后, 通过比较两者主题概率分布的相近性来对所有数据源进行排序。通过将数据源和查询映射到低维的主题空间来解决高维词条空间稀疏性所带来的信息缺失问题。在 TREC FedWeb 2013 和 2014 Track 的测试集上分别进行了实验, 并和其他参赛方法的结果进行了比较。在 FedWeb 2013 测试集上的实验结果显示比其他参赛方法的最好结果提高了 24%; 在 FedWeb 2014 测试集上的实验结果显示比传统的基于小文档和大文档的关键词匹配方法分别提高了 22% 和 43%。另外, 使用文档片段来代替文档还可以大幅提升系统的效率, 更增加了此方法的实用性和可行性。

关键词: 深层网; 主题模型; 隐含狄利克雷分布; 数据源选择; 联邦搜索

中图分类号: TP391.3 **文献标志码:** A

Deep Web resource selection using topic model

WANG Qiuyue, CAO Wei*, SHI Shaochen

(School of Information, Renmin University of China, Beijing 100872, China)

Abstract: Federated search is a widely-used technique to find information on Deep Web. Given a user query, one of the challenges for a federated search system is to select a set of resources that are most likely to return relevant results for the query. Most existing resource selection methods are based on text-matching between the sample documents of the resource and the query, which typically suffer the problem of missing vocabulary or incomplete information. To alleviate the problem of incomplete information, Latent Dirichlet Allocation (LDA) topic model approach for resource selection was proposed. First, topic probability distributions for resources and query were inferred using LDA topic model approach. Then the similarities between the topic distributions of resources and query were calculated to rank the resources. By mapping both resources and the query into the low dimensional topic space, the problem of missing information caused by the sparsity of high dimensional word space was alleviated. Experiments were conducted on the test sets of TREC FedWeb 2013 and 2014 Tracks, and the results were compared with that of other participants in the Tracks. The experimental results on the TREC FedWeb 2013 Track show that the LDA based approach outperforms the best result of other participants by 24%; and the results on the TREC FedWeb 2014 Track show that it outperforms the best results of the traditional text-matching-based resource selection methods using either small- or big-document strategies by 22% for small-document methods and 43% for big-document methods respectively. In addition, using sampled snippets rather than documents to generate big-document representation for resources can significantly improve the efficiency of the system, thus enables the proposed approach more feasible and applicable in practice.

Key words: deep Web; topic model; Latent Dirichlet Allocation (LDA); data resource selection; federated search

0 引言

随着互联网技术的不断发展, Web 上各种信息和数据呈爆炸性的增长。这其中不仅包括能被通用搜索引擎(如 Google、百度等)索引和检索到的静态网页, 还包括更多的各

种 Web 数据库中的数据, 例如电子商务网站上各种商品的数据、科学数据库中的数据、电子图书馆的目录、飞机订票系统中的数据等。这些数据大多存于网站后端的数据库中, 很难被一般搜索引擎爬取和索引, 用户只能通过网站提供的查询界面(如 Form)来获取信息。例如, 在天猫商城的搜索框中输

收稿日期: 2015-04-07; 修回日期: 2015-05-05。

基金项目: 国家自然科学基金资助项目(61202331, 61472425); 软件工程国家重点实验室开放研究基金资助项目(SKI2012-09-33)。

作者简介: 王秋月(1974-), 女, 山西定襄人, 讲师, 博士, CCF 会员, 主要研究方向: 数据库与信息系统、信息检索、知识库、自然语言问答; 曹巍(1975-), 女, 辽宁沈阳人, 讲师, 博士, CCF 会员, 主要研究方向: 高性能数据库、数据库自我管理自调优、闪存数据库; 史少晨(1992-), 男, 江苏淮安人, 硕士研究生, 主要研究方向: 数据库管理系统、信息检索、数据挖掘。

入“iphone 5S”的关键词查询,其后台数据库将返回超过100个商品记录。但百度对相同的查询“iphone 5S site:tmall.com”只能返回不到20个结果,且没有一个是天猫商城卖家的iPhone 5S商品页面。这类数据源通常被统称为深层网(Deep Web)或暗网(Hidden Web)^[1-2]。据2000年估计^[1],深层网的数据量是搜索引擎索引的表层网数据量的500倍。Google 2007年估计Web上已有近1千万个不同的有用的Form^[3]。据2011年2月的最新估计,Web上大约有超过10亿个结构化数据集^[4]。

从深层网中获取信息的方法大概有两种:数据表层化法和联邦搜索法。数据表层化法是一种数据预取法,它为每个Form预先生成一些查询,由于每个查询对应一个带?的URL,因而可以像其他静态HTML网页一样被抓取和索引,即将深层网中的数据表层化了。它的优点是不需要改变搜索引擎现有的体系结构,因而被大部分搜索引擎所采用,例如Google的深层网爬取器^[5]。但它的缺点是无法适应底层数据的动态变化;另外也无法预计算和抓取使用POST方法的Form,因为所有POST方法提交的Form都具有相同的URL。联邦搜索法(federated search),又称分布式信息检索(distributed information retrieval)或选择性元搜索(selective meta-search),则不受此限制,并能很好地适应底层数据的动态变化。

对于用户提交的关键词查询,联邦搜索系统选择一些最有可能返回相关结果的网站,并将用户的查询提交给这些网站的查询接口,最后再把每个网站返回的搜索结果合并排序成最终结果,返回给用户。它需要解决的主要问题是数据源选择问题和结果合并排序的问题。本文的主要工作集中在数据源的选择问题上,即如何为给出的关键词查询选择一组最相关的深层网数据源。

1 数据源的表示和选择

对于一个给定的用户查询,如关键字查询,数据源选择问题是在系统已知的所有数据源中选出一组最有可能返回相关结果的数据源。

数据源的相关性与很多因素有关^[6]:数据源自身与具体查询无关的权威性或有用性;数据源所包含的数据内容与查询的匹配程度;数据源的主题与查询主题的匹配程度等。

1.1 数据源的有用性

数据源自身的与具体查询无关的权威性或有用性,例如华东师范大学组在TREC FedWeb 2014的工作中提出的搜索引擎影响因子(Search Engine Impact Factor)^[7]。它和数据源本身能提供的信息的质量和数量有关,在一定程度上可以反映用户对数据源的选择偏好。它类似于网页搜索中使用的PageRank等先验概率,与具体查询无关。文献[7]中还给出了对搜索引擎影响因子的两种估计方法:一种是使用著名的互联网分析公司comScore发布的搜索引擎公司的市场份额报告来估算每个数据源的搜索引擎影响因子;另一种方法是使用TREC FedWeb 2013发布的人工评价的查询与数据源的相关度来估计搜索引擎影响因子的大小。查询与数据源的相关度按照查询进行归一化后,再按照数据源聚集从而得到每个数据源的搜索引擎影响因子的估计值。基于他们以往的实

验结果,文献[7]的作者指出第二种估计方法比第一种更有效。事实上,文献[7]在TREC FedWeb 2014的实验结果显示用第二种方法估计出的搜索引擎影响因子在数据源选择中单独使用就可以获得非常好的效果,当然结合其他因素,如内容匹配等,可以获得更好的效果。

1.2 数据源的内容匹配

大部分数据源选择算法衡量数据源对给定查询的相关度是基于它们内容上的匹配,即数据源中的文档所包含的关键词与查询关键词的匹配程度。

对于深层网的数据源,由于各种原因人们无法预先获得其完整而准确的内容描述(如包含的所有关键词及其出现频率,以及包含的所有文档个数等),所以一般使用采样的方法(如提交随机生成的关键词查询)获取数据源中的部分数据,称之为样本文档^[8]。例如,TREC FedWeb Track^[9]向每个数据源提交了2000(2013年)或4000(2014年)个采样查询,将每个查询返回的前10个文档收集起来构成了每个数据源的样本文档集。

不同的数据源选择算法再从样本文档集中提取出需要的信息来评估每个数据源和查询的相关度。基于内容匹配的方法大概可以分成两大类:大文档策略和小文档策略。

大文档策略如CORI^[10]和语言模型^[11]等。在大文档策略中,每个数据源的所有样本文档被串接起来,视为一个“大文档”,即每个数据源被表示为一个“大文档”,那么就可以使用已有的各种信息检索模型来计算数据源(大文档)和查询之间的相关度了。例如,文献[11]使用经典的语言模型计算数据源和查询之间的相关度。

小文档策略如GAVG^[12]、ReDDE^[13]和CRCS^[14]等。在小文档策略中,每个样本文档不被串接起来构成大文档,而是单独计算和查询的相关度。最后每个数据源的相关度由它包含的所有样本文档的相关度聚集得到,例如取所有样本文档相关度的几何均值^[12],或者根据样本文档相关度的排序以及数据源的大小等估算出数据源中相关文档的个数或密度作为数据源的相关度^[13-14]。

1.3 数据源的主题匹配

基于内容匹配的数据源选择算法的一个最大问题是采样数据的不完全性会造成匹配失败。相对于数据源中包含的所有数据而言,采样得到的数据只是其中的一小部分,因而没有出现在样本文档中的关键词就很难被匹配了。为了解决小样本带来的信息不完全问题,一些数据源选择算法把数据源表示成其内容所属的主题或类别,而不仅仅是关键词的集合。将数据源的内容从高维的词条空间映射到低维的主题空间,可以更好地泛化到未知的词条和文档,从而解决小样本的信息不完全问题。

大部分基于主题匹配的数据源选择算法使用的是预先构造好的主题或分类的层次结构^[15-16],例如开放式目录(Open Directory Project, ODP)或KDD-CUP 2005竞赛中提供的预先定义好的67个类别。文献[16]使用ODP提供的在线服务为每个数据源和查询获得其所涵盖的一组ODP类别,然后用数据源和查询的类别向量之间的余弦相似度或Jaccard相似度表示数据源对查询的相关度。文献[15]则使用查询探测法(query probing)在采样的过程中同时将各数据源划分到一个

预先定义好的类别层次结构中,其中父类涵盖子类。在计算数据源和给定查询的相关度时,数据源的内容表示,如数据源的语言模型,先和其相关类别(其父类和祖先类,以及子类等)中所有数据源的内容表示进行平滑操作。平滑操作是解决数据稀疏问题的一种常用方法。

另外一些基于主题匹配的数据源选择算法使用的是从数据中挖掘出来的主题或类别。例如,文献[17]使用文档聚类的方法获得数据源所属的类别;文献[18]使用一个二层的隐含狄利克雷分布(Latent Dirichlet Allocation, LDA)主题模型来描述数据源的生成过程,在数据集上训练后可以得到主题模型。在选择数据源时,文献[18]使用训练得到的主题语言模型来平滑各数据源的语言模型,从而解决缺失词条的问题。

本文提出使用 LDA 主题模型的方法从样本数据中自动挖掘出主题,并将各数据源和查询分别表示成主题分布向量;然后用数据源和查询的主题分布向量的相近程度来表示数据源对查询的相关度。本文的方法和文献[18]中提出的方法最接近,不同的是:1)文献[18]中使用的是扩展的二层 LDA 主题模型;而本文使用的是一层 LDA 主题模型。2)因而文献[18]只考虑了小文档策略,而本文的模型既可以考虑小文档策略又可以考虑大文档策略,事实上第3节的实验结果显示基于文档片段(snippet)的大文档策略是实际中更可行有效的方法。3)文献[18]中使用训练得到的主题语言模型平滑各数据源的语言模型,从而解决信息不全的问题,而本文的方法是将数据源和查询都映射成低维空间(即主题空间)的向量,从而解决稀疏数据带来的问题。

2 基于主题模型的数据源选择算法

本章将具体讲述基于主题模型的数据源选择算法。

2.1 主题模型简介

概率主题模型有很多种,如概率潜在语义索引(Probabilistic Latent Semantic Indexing, PLSI)、LDA 等,被广泛使用在文本挖掘、分类、主题发现等各类应用中。其中 LDA 是使用最广的一种主题模型。它是一种概率生成模型,描述了给定文档集中文档的生成过程。简而言之,LDA 模型认为一篇文档的每个词都是通过“以一定概率选择了某个主题,并在这个主题中以一定概率选择某个单词”这样一个过程得到的。

具体地,如图1所示。假设文档集中有 M 篇文档,每篇文档有 N 个单词,而所有文档的内容是关于 K 个主题的。其中 K 是预先设定的参数。模型中其他两个需要预先设定的参数是 α 和 β ,它们分别是两个 Dirichlet 分布 $Dir(\alpha)$ 和 $Dir(\beta)$ 的参数。 θ 和 φ 是两个 Multinomial 分布。

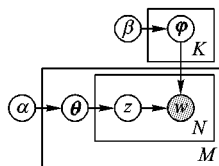


图1 LDA 的概率图表示

给定一个文档集,它的生成过程如下。

1) 为文档集中的每篇文档,从 $Dir(\alpha)$ 分布中选择一个主题概率分布 θ 。

2) 为每个主题,从 $Dir(\beta)$ 分布中选择一个主题语言模型(即单词概率分布) φ 。

3) 为文档中的每个单词:

① 先从文档的主题概率分布 θ 中选择一个主题 z ;

② 再从主题 z 的语言模型 φ 中选择一个单词 w 。

模型中, w 是观察到的数据, θ 、 φ 和 z 则是隐含的变量。可以通过对训练文档集的最大似然估计获得对 θ 、 φ 和 z 等隐含变量的估计,其中 θ 是每个文档的主题概率分布,也称之为主题模型, φ 是每个主题的语言模型。

例如,对于某一给定的文档集,表1和表2分别展示了 LDA 模型通过训练学习后获得的主题语言模型 φ 和文档的主题概率分布 θ 。从表1可以看到:主题1多包含 health、disease、cancer 等词条,是有关健康的;而主题2多包含 tv、movie、comedy 等词条,是和电视等娱乐节目有关的。同样从表2也可以发现,文档1的内容多是和主题1(即健康)相关的。

表1 示例:主题语言模型 φ

主题	词条概率分布/%						
	health	disease	cancer	...	tv	movie	comedy
主题1	30	15	5	...	0.1	0	0
主题2	0.2	0.001	0	...	15.0	5	3
主题3	0.3	0	0	...	0.7	0.002	0
...
主题K	0	0	0	...	0.0	0.001	0

表2 示例:文档的主题概率分布 θ

文档	主题概率分布/%				
	主题1	主题2	主题3	...	主题K
文档1	50	1	5	...	0
文档2	0.2	0.01	0	...	15
文档3	0.3	0	0	...	0.7
...
文档M	1	20	0	...	0

2.2 数据源的主题表示

采用 LDA 主题模型选择数据源时,首先要获得每个数据源的主题概率分布,即每个数据源中包含哪些主题的内容,并且每个主题的内容所占的比重是多少?

像大多数工作一样,先用查询采样的方法为每个数据源获得一组样本文档。基于数据源样本文档集表示,文献[18]使用一个扩展的二层 LDA 模型直接描述数据源的生成过程,即先生成数据源的主题分布,再以此为基础生成文档的主题分布,最后再生成文档中的单词。由于每个样本文档被单独考虑,这其实是一种“小文档”策略。本文采用两种策略来获得数据源的主题概率分布:“大文档”和“小文档”策略。

在大文档策略中,每个数据源的所有样本文档被串接在一起,视为一个大文档。这样就可以直接在这个“大文档”集上训练得到每个“大文档”(即数据源)的主题分布。而当数据源的样本文档比较多时,其“大文档”将很长,训练 LDA 主题模型的代价会很大。例如, FedWeb 2014 中对每个数据源提交了 4000 个采样查询,每个查询返回的前 10 个文档被爬取下来作为该数据源的样本文档,从而每个数据源的样本文档集有大约 40000 个文档,串接起来的大文档通常会有几百兆甚至几 G 字节,造成训练 LDA 主题模型时非常大的空间和时间代价。受文献[19-20]中工作的启发,本文提出使用采样查询返回的结果题目(title)和片段(snippet)来替代文档。这通常只有 200 个左右的字符,串接起来的“大文档”也不会

太大, LDA 主题模型训练所需要的时间和空间都大幅减小了。另外对有些数据源来说, 使用文档片段代替文档是必需的。例如, FedWeb 2014 中的第 122 个数据源 e122 (即 Picasa 网站, <http://picasaweb.google.com/lh/explore>), 其所有样本文档都是图片文件, 还有一些数据源是关于视频或音频的, 其所有样本文档都是视频或音频文件。对于这些只包含非文本文档的数据源, 其“大文档”表示为空, 所以无法应用 LDA 主题模型进行训练。而采样查询返回的结果片段中含有题目 (title) 和片段 (snippet) 等少量文本内容, 从而可以构成对数据源的一个“大文档片段”描述。事实上, 在 FedWeb 2013 和 2014 测试集上的实验显示, 基于文档片段训练出的主题模型在选择数据源时不仅效率高, 同样也可以达到很好的效果。具体的实验结果将在第 3 章中讨论。

在小文档策略中, 数据源的每个样本文档被视为独立的, 而不是被串接起来。所有数据源的所有样本文档放在一起构成 LDA 主题模型的训练文档集, 经过训练得到每个样本文档的主题分布。然后对于每个数据源, 可以使用其所有样本文档的主题分布的均值作为该数据源的主题分布。小文档策略中没有使用结果片段来替代完整的样本文档, 因为每个片段很短, 而短文档带来的词共现的稀疏性使得 LDA 主题模型算法在短文档上训练的效果不好。

这一步结束后就获得了如表 3 所示的每个数据源的主题概率分布 (假设有 L 个不同的数据源)。例如, 数据源 1 主要包含主题 1 (即健康) 方面的内容, 而数据源 L 是有关电视电影娱乐方面的网站。

表 3 数据源的主题概率分布

数据源	主题概率分布/%				
	主题 1	主题 2	主题 3	...	主题 K
数据源 1	10	7	5	...	0.5
数据源 2	0.2	10	30
...
数据源 L	1	20	0	...	0

2.3 查询的主题表示

给定一个用户查询, 可以将其看成一个文档, 基于前述在样本文档集上训练出的主题语言模型, 推断出用户查询的主题概率分布, 即该查询涉及哪些主题的内容, 每个主题所占的比重是多少? 例如, 查询“baseball team”所涉及的主题主要是体育, 而“sony vaio laptop”的主题主要是 IT 技术、电子产品等等。

但是用户查询一般很短, 例如 FedWeb 2013 和 2014 的测试查询大部分由 2~3 个关键词组成, 也有很多只包含一个关键词, 因而有必要对查询进行扩展。查询扩展是信息检索领域提高检索性能 (增大查全率同时也会改善查准率) 的常用技术。查询扩展有很多方法, 如使用同义词典或相关性反馈以及伪反馈等, 在原始查询中增加一些语义相近或经常共现的词等。本文使用 Google 来扩展查询。具体地, 调用 Google Search API, 输入用户提出的关键词查询; Google 会返回一个结果数组; 排在前 10 名的结果项中的片段 (snippet) 被提取出来, 合并成一个文档, 然后将其中词频最高的前 50 个单词添加到原查询中形成扩展后的查询。表 4 给出了两个查询的扩展示例。

表 4 基于 Google 检索结果实现的用户查询扩展

查询序号	查询形式	用户查询
1	原始	LHC collision publications
	扩展后	the lhc alice collisions proton lead publications with physics made publication correlations and experiment this collider first large hadron sep was comparison for time studying also both areas home area submitted corrections tev content competitive where other main jan reported them give insight nov into ions protons feat will that
2	原始	how to write journalistic review
	扩展后	the writing and review reviews journalistic news write editor based report subject your but journalism oct knowledge great for guardian book sunday movie from different deputy creative presentation how theater guide cricketer films most transcript are books common restaurant runs voice weaving critical identified includes inches clearly following objective diversity cultural

扩展后的包含 50 个关键词的查询被当作一个新文档, 以先前训练出的主题语言模型 ϕ 为基础, 推断出其主题概率分布。

2.4 数据源与查询的相关度排序

获得各数据源和查询的主题概率分布后, 可以通过比较两个概率分布之间的相近性来衡量各数据源和查询的相关度。这背后的基本假设是主题分布和查询的主题分布越接近的数据源, 越有可能含有和查询相同主题的内容, 即越有可能返回和查询相关的结果。

两个概率分布的相近性一般用 KL 距离 (Kullback-Leibler divergence) 的负值来衡量。式 (1) 被用来计算数据源 R 和查询 Q 的主题概率分布之间的 KL 距离:

$$D_{KL}(Q \| R) = \sum_{i=1}^K P(t_i | Q) \frac{P(t_i | Q)}{P(t_i | R)} \quad (1)$$

其中: K 是主题的个数, LDA 模型中的一个输入参数。

对给定的查询, 首先计算出它的主题概率分布与每个数据源的主题概率分布的相近性; 然后, 将所有数据源按照计算出的相近性从高到低排列。例如, 给定查询“sony vaio laptop”, 从 FedWeb 2013 给出的 157 个数据源中选出的与查询相关度最高的 3 个数据源如表 5 所示。直观来看, 系统有效地选出了和 IT 相关的咨询、广告和购物等网站。

表 5 数据源按与查询“sony vaio laptop”的相关度排序 (部分)

排名	数据源	相关度	简述
1	CNET	-0.249	美国领先的 TMT 媒体网站
2	Craigslist	-0.522	旧金山一个网上大型免费分类广告网站
3	eBay	-0.541	美国最大的 C2C 购物网站

3 实现与实验

为了验证使用 LDA 主题模型选择深层网数据源的方法的有效性, 本文进行了一系列的实验, 并和其他数据源选择方法进行比较。

3.1 TREC FedWeb Search Track

首先, TREC FedWeb Search Track^[9] 提供的数据集和查询集被选作测试集。在此之前, 测试联邦搜索数据源选择算

法的数据集最常用的是卡耐基梅隆大学 Jamie Callan 教授组提供的分布式信息检索测试集 (Distributed IR Testbeds)^[21]。但它是将已有的文档集按照一定规则划分成多个子集,每个子集代表一个数据源。例如,它将 TREC CD 1、2 和 3 上的所有文档按照数据来源划分成 100 个数据集,每个数据集代表一个搜索引擎(数据源);或者将 TREC CD 4 上的文档使用 *k*-means 聚类算法聚成 100 个数据集,代表 100 个数据源,则每个数据源中的文档是主题相似的。文献[22]第一次提出了构建一个大规模的基于 Web 上真实数据源的测试集,用于评测联邦搜索的各种算法,并在 2013 和 2014 年的 TREC 会议上分别举办了两次 FedWeb Search Track,进行公开评测。和文献[21]不同的是,文献[22]使用的是 Web 上真实的数据源,不是人工构造的数据源;另外,其数据集的规模更大。

FedWeb 2013 数据集包含 157 个数据源,大概可以归为 24 类,如新闻、娱乐、购物、社交、学术、健康、书籍、技术、视频、音频、图像、问答、百科知识等^[23]。每个数据源被提交了 2000 条采样查询,每条查询的前 10 条检索结果所指向的 HTML、TXT、PDF、WORD、EXCEL、PPT、图像、视频、音频等各类文件被下载下来,作为该数据源的样本文档,合计约 260 万个文档,总大小约 180 GB。另外,每条采样查询的前 10 条检索结果的片段(snippet)也被存储下来。在 2000 条采样查询中,1000 条查询是从 ClueWeb09-A 文档集中随机抽取出的单词,它们对所有被采样的数据源都是一样的;另外 1000 条查询则是从各数据源前面采样得到的文档中随机抽取单词构成。FedWeb 2014 使用的数据源和 FedWeb 2013 的数据源相同,只是数目减少到 149 个而不是 157 个,因为某些数据源在 2014 年不能被爬取了^[24]。所有数据源被重新采样,而且每个数据源的采样查询增加到 4000 条。同样地,其中 2000 条采样查询对所有数据源都是一样的;而另外 2000 条采样查询则是从各数据源已有的样本文档中随机抽取单词构成。合计约 500 万个文档,总大小约 370 GB。

FedWeb 2013 和 2014 分别发布了 200 条和 75 条测试查询,但是每年只选取其中的 50 条查询给出了人工评估结果。具体地,每条查询提交到各个数据源上返回的前 10 条结果(文档和结果片段)的相关度被人工标注。标注结果是 TREC Web Track 中采用的分级相关度(graded relevance level):Non(不相关),Rel(相关),HRel(高度相关),Key(顶级相关)和 Nav(导航级相关)。每个级别的相关度被附以一个权值,如 $w_{\text{non}} = 0, w_{\text{rel}} = 0.25, w_{\text{hrel}} = 0.5, w_{\text{key}} = 1, w_{\text{nav}} = 1$ 。这是 FedWeb 2013 中使用的相关度权值,在 FedWeb 2014 中有点儿变化,即 $w_{\text{rel}} = 0.158, w_{\text{hrel}} = 0.546$ 。它们是在 FedWeb 2013 标注结果的基础上,使用文献[25]中提出的用户分歧模型(User Disagreement Model)推算出的权值。接下来,评估系统使用相关度权值按照文献[26]中提出的方法计算出每个数据源返回的前 10 条结果的分级精度(Graded Precision)。它是一个 0~1 的数值,乘以 100 并四舍五入就可以被转化为 0~100 的一个正整数。这个值被作为数据源和查询之间的相关度的衡量。显然这是一个分级相关度,从而可以计算 nDCG(normalized Discounted Cumulative Gain)等指标,来获得对提交结果的一个整体评价指标,方便系统之间的比较。FedWeb 2013 和 2014 使用 nDCG@20 作为比较各数据源选择算法效果的主要评价指标。

3.2 系统实现与性能分析

MALLET^[27] 软件工具箱里的主题模型工具被用来进行第 2 章中所讨论的有关 LDA 主题模型的各种操作,包括在样本文档集上进行主题模型的训练和对查询作主题分布的推断。在应用 MALLET 进行主题模型训练或推断之前,要先对文档进行一些预处理。具体地,各种类型(f、xml、pdf、ppt、doc、xls 等)的文档先被分别解析提取出其中的文本内容部分,去除停用词(使用的是 Indri 附带的停用词表),并使用 Krovetz 词根分析法将所有单词的不同形式转化为其词根形式。对查询进行扩展以及计算各数据源和查询之间主题分布的相似度并排序是用 Java 编程实现的。

系统的总体运行时间由 3 部分构成:1)在样本文档集上训练 LDA 主题模型;2)查询扩展并用 1)得到的主题模型推断查询的主题分布;3)计算每个数据源的主题分布和查询主题分布的相似度并排序。其中 2)和 3)的运行时间都很短,在几秒到几十毫秒之间,而且对第 2.2 节中讨论的三种不同策略(即大文档、大文档片段和小文档策略)都是一样的,没有太大差别。第 1)步,即 LDA 主题模型的训练,是占用时间和空间最多的。一般地,LDA 训练算法的时间和空间复杂度与训练文档集的大小(即单词总数) N ,词汇表的大小(即不同单词的个数) W ,文档个数 D ,主题数目 K ,以及迭代次数 I 成正比。一般 LDA 训练算法的时间复杂度是 $O(INK)$,空间复杂度是 $O(K(D+W)+N)$ ^[28]。因而当文档集越大时,训练 LDA 模型所需要的时间和空间就越多。例如在 TREC FedWeb 2013 数据集上,用大文档、小文档和大文档片段 3 种不同策略得到的训练文档总长度(即 N)、最大文档长度和词汇表大小(即 W)如表 6 所示。同时,表 6 中还给出了当主题数目设为 10,训练的迭代次数设为 250,从第 50 次开始每迭代 20 次优化一下 LDA 模型的超参数(hyperparameter)时,MALLET 训练主题模型所需要的运行时间和内存大小。可以看到,由于大文档和小文档策略比大文档片段策略的训练文档集大很多,所以训练时间和内存占用都非常大。

表 6 3 种不同策略的训练集大小和训练性能的比较

策略	训练集大小			训练性能	
	文档集总长度	最长文档长度/MB	词汇表大小/MB	时间/min	内存/GB
大文档	1.5 GB	58.0	11.5	1440	28
小文档	1.5 GB	1.1	11.5	420	28
大文档片段	64 MB	1.5	1.4	20	2

虽然 LDA 模型的训练需要花费很多时间,例如即使在大文档片段策略下,训练 10 个主题仍然要用 20 多分钟的时间,但这一步只是在系统一开始为数据源建立主题分布时做,然后存储起来,不需要为每个查询重复该步骤,所以不会影响查询响应时间。另外,有条件的话,还可以使用多台机器,用并行算法提高 LDA 模型的训练效率^[28]。

3.3 实验结果与分析

表 7 显示了基于 LDA 主题模型的数据源选择算法在 TREC FedWeb 2013 测试集上的实验结果,比较了第 2.2 节中讨论的三种不同策略(即大文档、大文档片段和小文档策略)在数据源选择上的性能。比较指标是 nDCG@20。大文档策略即将每个数据源的所有样本文档串接成一个大文档;大文

档片段策略是将每个数据源的所有样本文档片段 (snippet) 而不是文档本身串接成一个大文档; 小文档策略则是每个样本文档独立处理而不被串接起来。

表 7 TREC FedWeb 2013 上的实验结果

主题 数目 K	nDCG@20		
	大文档	小文档	大文档片段
10	0.263	0.110	0.245
20	0.282	0.163	0.203
30	0.337	0.221	0.296
50	—	0.244	0.314
100	—	0.260	0.372
150	—	—	0.366
200	—	—	0.307
250	—	—	0.328

随着主题数目 K 增大, 训练 LDA 主题模型所需要的时间和空间会在表 6 的基础上进一步增大, 在现有的实验条件下无法完成, 所以这些结果在表 7 中被标为“—”。实验结果显示大文档策略比小文档策略在数据源选择上更有效, 但是大文档所花费的代价要高得多。而大文档片段策略的代价最小, 同时效果上只稍逊于大文档策略, 因而更适合在实际中使用。

本方法没有正式参加 TREC FedWeb 2013 Track。但根据 [23] 中报告的参赛结果, 表 8 给出了本文的算法和其他参赛算法的结果比较。对每个参赛队提交的结果, 表 8 只列出了其最好的一组结果。最后一行是本文的算法在 FedWeb 2013 测试集上的最好结果, 它是使用大文档片段策略, 主题数目设为 100 得到的。可以看到, 本文算法的结果显著好于 FedWeb 2013 所有参赛算法的结果。

表 8 与其他 TREC FedWeb 2013 参赛算法的结果比较

参赛组	Run ID	nDCG@20	所用的资源
University of Padova	UPDDW13mu	0.299	documents
Track 组织者	RS_clueweb	0.298	snippets
University of Stavanger	UiSP	0.276	documents
University of Delaware	udelFAVE	0.244	documents
University of Twente	utTailyM400	0.216	documents
Centrum Wiskunde and Informatica	cwi13SniTI	0.123	snippets
International Institute of Information Technology	iiitnaive01	0.107	snippets, Wikipedia, WordNet
East China Normal University	ECNUBM25	0.105	snippets, Google search
Indian Statistical Institute	incgqdv2	0.037	Google Query
Stanford University	StanfordEIG10	0.018	documents
中国人民大学	FW13Search100	0.372	snippets, Google search

本方法在参加 TREC FedWeb 2014 Track 的数据源选择任务中, 正式提交了 6 组运行结果。它们的评测结果见表 9: 其中 3 组以 Search 命名的结果, 即 FW14Search100、FW14Search75 和 FW14Search50, 使用的是大文档片段策略; 另外 3 组以 Docs 命名的结果, 即 FW14Docs100、FW14Docs75 和 FW14Docs50, 使用的是小文档策略。提交结果中没有大文档策略的运行结果, 因为 FedWeb 2014 数据集中每个数据源的样本文档数目是 FedWeb 2013 数据源样本文档数的两倍, 生成的大文档太大, 无法在合理的时间内训练和推断主题模型。

表 9 TREC FedWeb 2014 上的实验结果

Run ID	主题 数目 K	nDCG@20	nDCG@10	nP@1	nP@5
FW14Search100	100	0.505	0.425	0.278	0.384
FW14Search75	75	0.461	0.366	0.256	0.345
FW14Search50	50	0.517	0.426	0.271	0.404
FW14Docs100	100	0.444	0.337	0.165	0.239
FW14Docs75	75	0.422	0.306	0.106	0.198
FW14Docs50	50	0.419	0.292	0.174	0.203

和 FedWeb 2013 数据集上的实验结果类似, 大文档片段策略要优于小文档策略。不同的是, 获得最好结果 (即最大 nDCG@20) 的主题数目 (K) 在 FedWeb 2014 数据集上为 50, 而在 FedWeb 2013 数据集上是 100。如何为不同的数据集设置最优参数, 如主题数目, 是使用主题模型进行数据源选择的一个待研究的问题。

和 FedWeb 2014 的所有其他参赛队相比, 本方法提交的最好结果仅次于华东师范大学的最好结果, 名列第二。表 10 列出了所有参赛队的最好结果。值得注意的是, Drexel University 队实验了所有文献中已经提出的基于关键词匹配使用小文档策略的数据源选择方法, 例如 ReDDE, ReDDE.top, CRCSLinear, CRCSEXP, CiSS, CiSSAprox 和 SUSHI^[29]。最好的结果是使用 SUSHI 模型得到的, nDCG@20 为 0.422。而 University of Illinois 队提交的两组结果使用的是经典的基于关键词匹配使用大文档策略的数据源选择算法, 如语言模型、CORI 等, 见第 1.2 节中的讨论。其最好结果为 0.361。因而, 可以看出基于主题匹配进行数据源选择的方法要优于所有传统的基于关键词匹配 (无论是使用小文档还是大文档策略) 的数据源选择方法。另外, 华东师范大学的算法中对结果贡献最大的部分是其基于 FedWeb 2013 人工评估结果估算出来的搜索引擎影响因子, 因为其只按数据源的搜索引擎影响因子排序而得到的结果是 0.651, 而其他没有考虑搜索引擎影响因子得到的结果都低于 0.3。而搜索引擎影响因子是一个查询无关的因素, 可以和本文提出的主题模型结合起来使用, 两者具有互补性, 很有可能会产生更好的结果。

表 10 与其他 TREC FedWeb 2014 参赛算法的结果比较

参赛组	Run ID	nDCG@20	所用的资源
East China Normal University	ecomsvz	0.712	snippets, Google search, KDD 2005
Renmin University of China	FW14Search50	0.517	snippets, Google search
Chinese Academy of Sciences	ICTNETRS05	0.436	documents, Google search, NLTK, GENSIM
Drexel University	drexelRS7	0.422	documents
University of Illinois	uiucGSLIS2	0.361	documents
University of Delaware	udeltrsb	0.355	documents
University of Stavanger	NTNUIsrs2	0.348	snippets, documents
University of Twente	UTTailyG2000	0.323	documents
University of Padova	UPDFW14tipsm	0.311	documents
University of Lugano	ULuganoDFR	0.304	documents

4 结语

现有的各种深层网数据源选择算法对样本文档集信息缺

失问题处理不足。本文针对此问题提出使用主题模型的方法,如 LDA,从数据源的样本文档集中发掘出数据源所涵盖内容的主题分布;并将用户的查询也映射到同一主题空间,表示成主题分布向量;最后通过比较数据源和查询主题分布之间的相近程度,来选择和查询高度相关的数据源。由于用户提出的关键词查询一般很短,所以在推断其主题分布之前,本文先用 Google Search API 扩展查询。在 TREC FedWeb 2013 和 2014 两个标准测试集上进行的实验结果显示这是一个非常有前景的方法。基于主题模型的数据源选择方法优于传统的基于关键词匹配的数据源选择方法。另外,不同于以往基于关键词匹配的数据源选择方法的发现,即小文档策略优于大文档策略,在本文提出的基于主题模型的数据源选择算法中,大文档策略要优于小文档策略,而且使用文档片段的大文档策略不仅代价很小并且同样可以获得较好的效果,所以更实用。

关于未来工作,本项目计划实验更多更复杂的主题模型,如层次结构的 LDA^[18]或显式语义分析(Explicit Semantic Analysis, ESA)^[30]等。另外还将深入研究如何将基于关键词匹配,主题匹配,以及数据源的先验有用性(如搜索引擎影响因子、数据源所包含的文档数目)等因素结合起来预测数据源对给定查询的相关性。

参考文献:

- [1] BERGMAN M K. The deep Web: surfacing hidden value [J]. *Journal of Electronic Publishing*, 2001, 7(1): 113–153.
- [2] HE B, PATEL M, ZHANG Z, *et al.* Accessing the deep Web: a survey [J]. *Communications of ACM*, 2007, 50(5): 94–101.
- [3] MADHAVAN J, JEFFERY S, COHEN S, *et al.* Web-scale data integration: you can only afford to pay as you go [EB/OL]. [2015-01-04]. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.66.9358&rep=rep1&type=pdf>.
- [4] CAFARELLA M J, HALEVY A, MADHAVAN J. Structured data on the Web [J]. *Communications of ACM*, 2011, 54(2): 72–79.
- [5] MADHAVAN J, KO D, KOT L, *et al.* Google's deep Web crawl [J]. *Proceedings of the Very Large Data Base Endowment*, 2008, 1(2): 1241–1252.
- [6] ARGUELLO J, CALLAN J, DIAZ F. Classification-based resource selection [C]// *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. New York: ACM, 2009: 1277–1286.
- [7] SHAN J, MAN L. Simple may be best — a simple and effective method for federated Web search via search engine impact factor estimation [EB/OL]. [2015-01-06]. http://trec.nist.gov/pubs/trec23/papers/pro-ECNU_federated.pdf.
- [8] CALLAN J, CONNELL M. Query-based sampling of text databases [J]. *ACM Transactions on Information Systems*, 2011, 19(2): 97–130.
- [9] HIEMSTRA D, DEMEESTER T, TRIESCHNIGG D. TREC federated Web search track [EB/OL]. [2015-01-03]. <https://sites.google.com/site/trecfedweb/>.
- [10] CALLAN J P, LU Z, CROFT W B. Searching distributed collections with inference networks [C]// *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM, 1995: 21–28.
- [11] SI L, JIN R, CALLAN J, *et al.* A language modeling framework for resource selection and results merging [C]// *Proceedings of the 11th International Conference on Information and Knowledge Management*. New York: ACM, 2002: 391–397.
- [12] SEO J, CROFT W B. Blog site search using resource selection [C]// *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. New York: ACM, 2008: 1053–1062.
- [13] SI L, CALLAN J. Relevant document distribution estimation method for resource selection [C]// *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM, 2003: 298–305.
- [14] SHOKOUHI M. Central-rank-based collection selection in uncooperative distributed information retrieval [C]// *Proceedings of the 29th European Conference on Information Retrieval*. Berlin: Springer, 2007: 160–172.
- [15] IPEIROTTIS P G, GRAVANO L. Classification-aware hidden-Web text database selection [EB/OL]. [2015-01-08]. <http://128.59.11.212/~gravano/Papers/2008/tois08.pdf>.
- [16] BELLOGIN A, GEBREMESKEL G G, HE J, *et al.* CWI and TU delft at TREC 2013: contextual suggestion, federated Web search, KBA, and Web tracks [EB/OL]. [2015-01-08]. <http://ir.ii.uam.es/~alejandro/2013/trec.pdf>.
- [17] XU J, CROFT W B. Cluster-based language models for distributed retrieval [C]// *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM, 1999: 254–261.
- [18] BAILLIE M, CARMEN M, CRESTANI F. A multiple-collection latent topic model for federated search [J]. *Information Retrieval*, 2011, 14(4): 390–412.
- [19] DEMEESTER T, NGUYEN D, TRIESCHNIGG D, *et al.* What snippets say about pages in federated Web search [C]// *Proceedings of the 8th Asia Information Retrieval Societies Conference*. Berlin: Springer, 2012: 250–261.
- [20] DEMEESTER T, NGUYEN D, TRIESCHNIGG D, *et al.* Snippet-based relevance predictions for federated Web search [C]// *Proceedings of the 35th European Conference on Advances in Information Retrieval*. Berlin: Springer, 2013: 697–700.
- [21] CALLAN J. Distributed IR testbed definitions [EB/OL]. [2015-01-08]. <http://boston.lti.cs.cmu.edu/callan/Data/#DIR>.
- [22] NGUYEN D, DEMEESTER T, TRIESCHNIGG D, *et al.* Federated search in the wild: the combined power of over a hundred search engines [C]// *Proceedings of the 21st ACM Conference on Information and Knowledge Management*. New York: ACM, 2012: 1874–1878.
- [23] DEMEESTER T, TRIESCHNIGG D, NGUYEN D, *et al.* Overview of the TREC 2013 federated Web search track [EB/OL]. [2015-01-02]. <https://biblio.ugent.be/input/download?func=downloadFile&recordId=4402037&fileId=4402038>.
- [24] DEMEESTER T, TRIESCHNIGG D, NGUYEN D, *et al.* Overview of the TREC 2014 Federated Web Search Track [EB/OL]. [2015-01-02]. <http://www.dcs.gla.ac.uk/~zhouke/papers/trec2014fedweb-draft.pdf>.
- [25] DEMEESTER T, ALY R, HIEMSTRA D, *et al.* Exploiting user disagreement for Web search evaluation: an experimental approach [C]// *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*. New York: ACM, 2014: 33–42.

地提高解的精度和加快算法的收敛。

4 结语

在利用群体智能优化算法求解约束优化问题时,其性能由智能算法和约束处理技术共同决定。采用非固定多段映射罚函数法处理约束条件,将原约束优化问题转化为无约束优化问题。为了提高 GWO 算法的求解精度与收敛速度并避免陷入局部最优,引入佳点集理论初始化种群个体,结合 Powell 局部搜索方法,本文提出了一种改进灰狼优化(IGWO)算法,利用 IGWO 算法对转换后的无约束优化问题进行求解。对 6 个标准测试问题进行了仿真实验,结果表明,与其他群体智能优化算法相比,本文算法具有较好的寻优性能和较快的收敛速度。

参考文献:

- [1] QU B, SUGANTHAN P N, DAS S. A distance-based locally informed particle swarm model for multimodal optimization [J]. IEEE Transactions on Evolutionary Computation, 2013, 17(3): 387 - 402.
- [2] XU S, LONG W. Differential evolution algorithm with dynamically adjusting number of subpopulation individuals [J]. Journal of Computer Applications, 2011, 31(11): 3101 - 3103. (徐松金, 龙文. 动态调整子种群个体的差分进化算法[J]. 计算机应用, 2011, 31(11): 3101 - 3103.)
- [3] MIRJALILI S, MIRJALILI S M, LEWIS A. Grey wolf optimization [J]. Advances in Engineering Software, 2014, 69(7): 46 - 61.
- [4] MADADI A, MOTLAGH M M. Optimal control of DC motor using grey wolf optimizer algorithm [J]. Technical Journal of Engineering and Applied Science, 2014, 4(4): 373 - 379.
- [5] EMARY E, ZAWBAA H M, GROSAN C, *et al.* Feature subset selection approach by gray-wolf optimization [C]// Proceedings of the International Afro-European Conference on Industrial Advancement. Berlin: Springer, 2014: 1 - 13.
- [6] MIRJALILI S. How effective is the grey wolf optimizer in training multilayer perceptrons [J]. Applied Intelligence, 2015, 42(4): 608 - 619.
- [7] EI-GAAFARY A A M, MOHAMED Y S, HEMEIDA A M, *et al.* Grey wolf optimization for multi input multi output system [J]. Universal Journal of Communications and Networks, 2015, 3(1): 1 - 6.
- [8] SONG H M, SULAIMAN M H, MOHAMED M R. An application of grey wolf optimizer for solving combined economic emission dispatch problems [J]. International Review on Modelling and Simulations, 2014, 7(5): 838 - 844.
- [9] WU L, WANG Y, ZHOU S, *et al.* Differential evolution for nonlinear constrained optimization using non-stationary multi-stage assignment penalty function [J]. Systems Engineering Theory and Practice, 2007, 27(3): 128 - 133. (吴亮红, 王耀南, 周少武, 等. 采用非固定多段映射罚函数的非线性约束优化差分进化算法[J]. 系统工程理论与实践, 2007, 27(3): 127 - 133.)
- [10] PARSOPOULOS K E, VRAHATIS M N. Particle swarm optimization method for constrained optimization problems [EB/OL]. [2015-01-03]. http://www.researchgate.net/publication/2527227_Particle_swarm_optimization_method_for_constrained_optimization_problems.
- [11] HAUPT R, HAUPT S. Practical genetic algorithm [M]. New York: John Wiley & Sons, 2004.
- [12] ZHANG L, ZHANG B. Good point set based genetic algorithm [J]. Chinese Journal of Computers, 2001, 24(9): 917 - 922. (张铃, 张钹. 佳点集遗传算法[J]. 计算机学报, 2001, 24(9): 917 - 922.)
- [13] POWELL M J D. A fast algorithm for nonlinearly constrained optimization calculations [C]// Numerical Analysis, Lecture Notes in Mathematics 630. Berlin: Springer, 1978: 144 - 157.
- [14] WU J, ZHANG J, CHEN H. Particle swarm optimization algorithm combination with Powell search method [J]. Control and Decision, 2012, 27(3): 343 - 348. (吴建辉, 章兢, 陈红安. 融合 Powell 搜索法的粒子群优化算法[J]. 控制与决策, 2012, 27(3): 343 - 348.)
- [15] AMIRJANOV A. The development of a changing range genetic algorithm [J]. Computer Methods in Applied Mechanics and Engineering, 2006, 195(19/20/21/22): 2495 - 2508.
- [16] BOUSSAID I, CHATTERJEE A, SIARRY P, *et al.* Biogeography-based optimization for constrained optimization problems [J]. Computers and Operations Research, 2012, 39(12): 3293 - 3304.
- [17] LU H Y, CHEN W Q. Self-adaptive velocity particle swarm optimization for solving constrained optimization problems [J]. Journal of Global Optimization, 2008, 41(3): 427 - 445.
- [18] GANDOMI A H, YANG X S, ALAVI A H, *et al.* Bat algorithm for constrained optimization tasks [J]. Neural Computing and Applications, 2013, 22(6): 1239 - 1255.
- [19] MEZURA M E, COELLO C A, MORALES E. Simple feasibility rules and differential evolution for constrained optimization [M]. Berlin: Springer, 2004: 707 - 716.
- [20] KARABOGA D, AKAY B. A modified artificial bee colony algorithm for constrained optimization problems [J]. Applied Soft Computing, 2011, 11(3): 3021 - 3031.
- [21] LONG W, LIANG X M, HUANG Y F, *et al.* An effective hybrid cuckoo search algorithm for global constrained optimization [J]. Neural Computing and Applications, 2014, 25(3/4): 911 - 926.

(上接第 2559 页)

- [26] KEKÄLÄINEN J, JÄRVELIN K. Using graded relevance assessments in IR evaluation [J]. Journal of the American Society for Information Science and Technology, 2002, 53(13): 1120 - 1129.
- [27] MCCALLUM A K. MALLET: a machine learning for language toolkit [EB/OL]. [2015-01-02]. <http://mallet.cs.umass.edu>.
- [28] LIU Z, ZHANG Y, CHANG E Y, *et al.* PLDA+: parallel latent Dirichlet allocation with data placement and pipeline processing [J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): Article No. 26.
- [29] SHOKOUHI M, SI L. Federated search [J]. Foundations and Trends in Information Retrieval, 2011, 5(1): 1 - 102.
- [30] GABRILOVICH E, MARKOVITCH S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis [C]// Proceedings of the 20th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 2007: 1606 - 1611.