基于话题模型的专家发现方法*

刘 健,李 绮,刘宝宏,张 云
(国防科技大学 机电工程与自动化学院,湖南 长沙 410073)

摘 要:专家发现是实体检索的一个重要方面。经典的专家发现模型建立在专家与词项的条件独立性假设基础上。在实际应用中该假设通常不成立,使得专家发现的效果不够理想。本文提出了一种基于话题模型的专家发现方法,该方法无需依赖候选专家与词项的条件独立性假设,且其可操作性比经典模型更强。同时,使用了一种排序截断技术,该技术极大地降低了模型的计算复杂度。使用CERC(CSIRO Enterprise Research Collection)数据集对模型的性能进行评估。实验结果表明,基于话题模型的专家发现方法在各个评价指标上均优于经典的专家发现模型,能够有效地提高专家发现的效能。

关键词:实体检索;专家发现;基于话题的模型;排序截断
中图分类号:TP391 **文献标志码**:A **文章编号**:1001-2486(2013)02-0127-05

An expert finding method based on topic model

LIU Jian, LI Qi, LIU Baohong, ZHANG Yun
(College of Mechatronics Engineering and Automation, National University of Defense Technology, Changsha 410073, China)

Abstract: Expert finding is an important part of entity retrieval. Classical expert finding models rest upon the conditional independence assumption between the candidate and term-given document. However, this assumption is usually invalid in real world applications, which makes the performances of classical expert finding models not ideal. In this research, an expert finding method is proposed based on the topic model (EFTM). This method discards the conditional independence assumption in classical models and is more maneuverable. In addition, a ranking truncation approach which largely decreases the computational complexity of the model was used. Finally, the performances of the new model were evaluated using the CSIRO Enterprise Research Collection. The results shows that the EFTM model outperformed the classical model significantly on all the metrics and can effectively improve the performances of the expert finding system.

Key words: entity retrieval; expert finding; topic-based model; ranking truncation

“信息检索”(Information Retrieval,IR)一词由Calvin Mooers于1948年到1950年间提出[1]。在上世纪70年代到80年代,其主要的检索对象是具有特定内容的文档。近年来,随着人们对于信息需求的不断变化,信息检索的内涵也不再局限于传统的文档检索,而是扩展到各个领域。其中一个重要的分支是实体检索(Entity Retrieval,ER)[2-4]。专家发现作为ER的一个重要方面近年来已经得到了越来越多的关注[2-13]。

专家发现(Expert Finding)也称专家搜索,它研究的问题是:给定特定的查询,如何找到具有相关领域知识的专家,同时将这些专家按照其专业水平进行排序[3,5,10]。针对这类问题,目前主要有两种模型:基于Profile的模型和基于Document的模型[3,5-6,9-10]。这两种模型虽然能够在一定程度上满足专家发现任务的基本要求,但是仍然存在一些固有的缺陷。其中最重要的不足之处在于这两种模型都是建立在候选专家与查询词项的条件独立性假设基础上。该假设在实际应用中往往无法成立。本文针对这一问题,提出了一种基于话题模型的专家发现方法(an Expert Finding Method on the Topic Model),简称为EFTM模型。该方法无需依赖于候选专家和查询词项之间的条件独立性假设,实验结果表明该方法的各项性能指标都高于传统的专家发现模型。

1 经典专家发现模型

1.1 基于Profile的模型

基于Profile的模型也被称作基于候选人的模型(Candidate-based Model)或者查询独立模型(Query-independent Model),该模型的基本思想是:首先综合每一位专家各方面信息为其建立一

* **收稿日期**:2012-09-12
基金项目:国家自然科学基金资助项目(60704038)
作者简介:刘健(1983—),男,辽宁锦州人,博士研究生,E-mail:supakito@163.com;
刘宝宏(通信作者),男,副教授,博士,硕士生导师,E-mail:lbh_nudt@163.com

个 profile,然后使用概率[3, 5, 10]模型来计算该专家的 profile 与查询之间的相关性,并按照该相关性对专家进行排序。其中最为典型的是 Balog 等人提出的 Model 1[10]。

Model 1 是由 Balog 等人在 2006 年提出的一种专家发现模型[10]。记给定查询 *query* 的条件下,候选专家 *ca* 出现的概率为 $P(ca|query)$,则根据贝叶斯公式有

$$P(ca|query) = \frac{P(query|ca)P(ca)}{P(query)} \tag{1}$$

其中 $P(ca)$ 表示候选专家出现的先验概率,一般情况下认为其服从均匀分布。$P(query)$ 是查询 *query* 出现的概率,给定 *query* 的情况下,$P(query)$ 为常数。因此

$$P(ca|query) \propto P(query|ca) \tag{2}$$

根据语言模型有

$$P(query|\theta_{ca}) = \prod_{term \in query} [(1-\lambda_{ca})P(term|ca) + \lambda_{ca}P(term)]^{n(term,query)} \tag{3}$$

$$P(term|ca) = \sum_{doc \in D_{ca}} P(term|doc,ca)P(doc|ca) \tag{4}$$

其中 θ_{ca} 是候选专家 *ca* 的模型,*term* 是 *query* 中的一个词项,$n(term,query)$ 表示 *term* 在 *query* 中出现的次数,D_{ca} 为与 *ca* 相关的文档集合。

假设给定文档 *doc* 的条件下,词项 *term* 和候选专家 *ca* 是条件独立的。则 Model 1 可表示为

$$P(query|\theta_{ca}) = \prod_{term \in query} \{(1-\lambda_{ca})(\sum_{doc \in D_{ca}} P(term|doc)P(doc|ca)) + \lambda_{ca}P(term)\}^{n(term,query)} \tag{5}$$

1.2 基于 Document 的模型

基于 Document 的专家选择方法也被称为查询依赖模型(Query-dependent Model),其基本思想是:首先使用文档检索方法获得与查询相关的文档,然后按照专家与这些文档的相关程度对专家进行排序。一般而言,基于 Document 的模型其效果要好于基于 Profile 的模型[3, 5, 10, 14]。其中最经典的是 Balog 等人提出的 Model 2。该模型可视为一种生成模型,具体生成过程如下[3]:

(1) 给定一个候选专家 *ca*

(2) 选择同 *ca* 关联的文档 *doc*

(3) 根据该文档和候选专家,用给定概率 $P(query|ca,doc)$ 生成一个查询 *query*。对所有与 *ca* 相关联的文档进行加权求和,即可获得

$$P(query|ca) = \sum_{doc \in D_{ca}} P(query|doc,ca)P(doc|ca) \tag{6}$$

假定查询 *query* 中,各个词项是独立同分布的,则

$$P(query|doc,ca) = \prod_{term \in query} P(term|doc,ca)^{n(term,query)} \tag{7}$$

根据文献[3, 5, 8 – 10],假设 *term* 与 *ca* 之间是条件独立的,则

$$P(term|doc,ca) = P(term|\theta_{doc}) \tag{8}$$

对式(8) 进行平滑化处理,得

$$P(term|\theta_{doc}) = (1 - \frac{\beta}{\beta + n(doc)})P(term|doc) + \frac{\beta}{\beta + n(doc)}P(term) \tag{9}$$

其中,$n(doc)$ 为文档 *doc* 中所有的词项总数,β 为一常数。$P(term|doc)$ 为 *term* 在文档 *doc* 中出现的频率。$P(term)$ 是指 *term* 在整个文档集中出现的概率。

综合式(6) ~ (9),得

$$P(query|ca) = \sum_{doc \in D_{ca}} \prod_{term \in query} [(1 - \frac{\beta}{\beta + n(doc)})P(term|doc) + \frac{\beta}{\beta + n(doc)}P(term)]^{n(term,query)} P(doc|ca) \tag{10}$$

2 基于话题模型的专家发现方法

上述两类模型都依赖于条件独立性假设,即认为在给定文档的条件下,候选专家和查询词项是条件独立的。然而,在实际应用中,这一假设往往是不成立的。为了解决这个问题,Balog 等人相继提出了两种改进模型,分别称为 Model 1B 和 Model 2B[3, 5, 10]。这两种改进模型虽然能够从某种程度上解决该问题,但又需要引入另一种独立性假设,即认为文档、候选专家与窗口的大小是独立的。同时,在这两种模型中还需要计算不同尺度的窗口出现的先验概率。该方法计算量大、可操作性相对较差。本文提出了一种基于话题模型的专家发现方法。新模型建立在潜在狄利克莱分配(Latent Dirichlet Allocation,LDA) 方法的基础上,无需依赖 *ca* 和 *term* 之间的条件独立性假设。

2.1 话题生成模型

LDA 是 Blei 等人在 2003 年提出的一种话题

模型[15]。该模型是一种文档建模方法,认为文档是由一系列代表不同话题(Topic)的词组成的[15-17]。同时,整个语料集是由若干个包含不同话题的文档构成的。语料集的话题服从参数为 α 的狄利克莱分布,各个文档的话题分布为多项分布,此外,在每个话题中词项的分布也服从多项分布。建立了模型之后,通过参数估计,即可得到整个语料集以及各个文档的话题分布。

本文借助 LDA 方法的基本思想,提出如图 1 所示的专家发现模型。该模型为一个生成模型,具体生成过程如下:

(1) 首先,对于每一个文档 doc,从参数为 α 的 Dirichlet 分布中抽取一个多项式分布 θ_d,该分布代表了文档 doc 的话题分布,即 $P(z \mid doc)$。

(2) 对于文档 doc 中的每个词项 $term$,从多项式分布 θ_d 抽取出一个话题 z 作为该词项的话题。

(3) 对于每个话题 z,从参数为 β 的 Dirichlet 分布中抽取一个多项式分布 ϕ_z,该分布表示给定话题 z 的情况下,词项 $term$ 的分布,即 $P(term \mid z)$;

(4) 对于每个话题 z,从参数为 γ 的 Dirichlet 分布中抽取一个多项式分布 φ_z,该分布表示给定话题 z 的情况下,候选专家 ca 的分布,即 $P(ca \mid z)$

(5) 从多项分布 ϕ_z 抽取出一个词项 $term$

(6) 从多项分布 φ_z 抽取出一个候选专家 ca

重复上述过程,即可生成整个语料集。

假设 Z 为语料集的话题集合,z_n 为任意话题,则

$$P(term \mid doc,ca) = \sum_{z_n \in Z} P(term \mid doc,ca,z_n) P(z_n \mid doc,ca) \tag{11}$$

同时

$$P(term \mid doc,ca,z_i) = \frac{P(term,doc,ca,z_i)}{P(doc,ca,z_i)} \tag{12}$$

$$P(z_i \mid doc,ca) = \frac{P(z_i,doc,ca)}{P(doc,ca)} \tag{13}$$

将式(12)、(13)代入式(11)中得

$$\begin{aligned} P(term \mid doc,ca) &= \sum_{z_n \in Z} P(term \mid doc,ca,z_n) P(z_n \mid doc,ca) \\ &= \sum_{z_n \in Z} \frac{P(term,doc,ca,z_n)}{P(doc,ca)} \\ &= \sum_{z_n \in Z} \frac{P(term,ca,z_n \mid doc)}{P(ca \mid doc)} \end{aligned} \tag{14}$$

由图 1 可知,在给定 z 的条件下,ca 与 $term$、doc,ca、doc 与 $term$ 之间均是条件独立的,所以

$$\begin{aligned} &P(term,ca,z_n \mid doc) \\ &= P(term \mid ca,z_n,doc) P(ca \mid z_n,doc) P(z_n \mid doc) \\ &= P(term \mid z_n) P(ca \mid z_n) P(z_n \mid doc) \end{aligned} \tag{15}$$

式(15)中,$P(term \mid z_n)$、$P(ca \mid z_n)$、$P(z_n \mid doc)$ 都可以通过对图 1 所示的模型进行参数估计得到。

2.2　参数估计

本文使用 Gibbs 采样方法对 EFTM 模型进行参数估计。Gibbs 采样是马尔科夫蒙特卡罗(MCMC,Markov Chain Monte Carlo)方法的一种特殊情况。假设给定 $term$ 和 ca 的条件下话题的分布为 $P(z \mid term,ca)$,其中 $term$ 和 ca 分别表示词项和候选专家的集合,则根据图 1 可知

$$\begin{aligned} &P(z_i = j \mid z_{-i},term,ca) \\ &\propto P(z_i = j \mid z_{-i}) P(term_i \mid z,term_{-i}) P(ca_i \mid z,ca_{-i}) \\ &\propto \frac{H_{dj}^{N_d N_z} + \alpha}{\sum_{j'} H_{dj'}^{N_d N_z} + T\alpha} \frac{H_{mj}^{N_t N_z} + \beta}{\sum_{m'} H_{m'j}^{N_t N_z} + N_t\beta} \frac{H_{nj}^{N_{ca} N_z} + \gamma}{\sum_{n'} H_{n'j}^{N_{ca} N_z} + N_{ca}\gamma} \end{aligned} \tag{16}$$

其中,$z_i = j$ 表示第 i 个词项/候选专家属于话题 j,z_{-i} 表示其余词项/候选专家的话题,α、β 和 γ 为事先设定的超参数,N_d 为语料集中文档的数量,N_z 为话题的数量,N_{ca} 为候选专家的数量,N_t 为词表的大小,$H^{N_d N_z}$、$H^{N_t N_z}$ 和 $H^{N_{ca} N_z}$ 分别表示文档 - 话题、词项 - 话题以及候选专家 - 话题矩阵,其元素分别表示将文档(词项或候选专家)指派给某一话题的概率。此时,需要估计的变量为 θ、ϕ 和 φ。依照式(16)进行 Gibbs 采样,可以得到 $P(z \mid term,ca)$ 的一系列样本,则 θ、ϕ 和 φ 可以表示为如下形式:

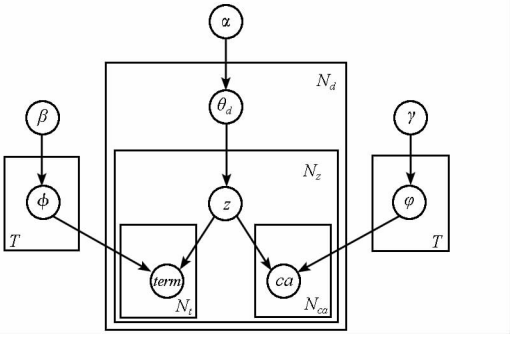

图 1　基于话题模型的专家发现模型

Fig. 1　Expert finding method based on topic model

$$\hat{\theta}_j^{(d)} = \frac{H_{dj}^{N_d N_z} + \alpha}{\sum_{j'} H_{dj'}^{N_d N_z} + N_z\alpha} \tag{17}$$

$$\hat{\phi}_j^{(t)} = \frac{H_{mj}^{N_t N_z} + \beta}{\sum_{m'} H_{m'j}^{N_d N_z} + N_t\beta} \tag{18}$$

$$\hat{\varphi}_j^{(ca)} = \frac{H_{nj}^{N_{ca}N_z} + \gamma}{\sum_{n'} H_{n'j}^{N_{ca}N_z} + N_{ca}\gamma} \quad (19)$$

因此,式(15)变为

$$\begin{aligned} & P(term, ca, z_i \mid doc) \\ &= P(term \mid z_n) P(ca \mid z_n) P(z_n \mid doc) \quad (20) \\ &= \hat{\theta}_j^{(d)} \hat{\phi}_j^{(term)} \hat{\varphi}_j^{(ca)} \end{aligned}$$

3 实验分析与对比

3.1 实验数据说明

本文采用的实验数据是由文本检索会议(Text Retrieval Conference, TREC)2007提供的CERC(CSIRO Enterprise Research Collection)数据集,包含从CSIRO网站上获取的文档共计370715篇,大小约为4.2G。该数据集包含Web页面、源代码以及数据文件等多种类型的文档。与TREC 2005和TREC 2006提供的W3C数据集不同的是,CREC并未提供候选专家列表,本文对Balog等人提供的列表[3]进行了一定的完善,形成新的候选专家列表,该候选专家列表中共包含专家3559名。本文使用"名/名首字母+姓"的匹配方式将文档集合中的候选专家姓名替换为候选专家列表中的email,并利用Lucene建立索引。

3.2 实验设计

为了降低LDA模型的计算复杂度,本文采用了一种排序截断技术[18]。首先,对于给定的查询,使用经典的文档检索方法获得与查询相关的前N篇文档。将这前N篇文档作为新的语料集。之后,分别使用经典的专家发现方法和EFTM方法进行专家检索。使用排序截断技术后,大大降低了$H^{N_tN_z}$、$H^{N_dN_z}$和$H^{N_{ca}N_z}$的规模,从而显著降低了(约5~6个数量级)计算的空间复杂度和时间复杂度。

3.3 实验结果

表1给出了Model 1、Model 2与本文提出EFTM模型的性能对比。可以看出,在所有的模型中,EFTM Model 2的性能最优。相对于Model 2而言,EFTM Model 2的各项性能指标均有大幅度的提升,其中MAP(Mean Average Precision)值提高了约40%,MRR(Mean Reciprocal Rank)值提高了约65%,P@5值提高了约20%。相比之下,对于Model 1,新算法和经典算法的各项指标则相差不大。这是由Model 1的特点决定的。在建立候选专家的profile过程中,没有考虑与候选专家相关联的文档的话题分布,这样会导致profile的话题分布均一化。因此,在profile上使用话题模型时,就可能会导致专家发现的性能提升不显著。

表1 截断值$N=200$的实验结果对比

Tab. 1 Comparison of the performance of the baseline model and he EFTM model when the parameter N is 200

	MAP	MRR	P@5	P@10
Model 1	0.2118	0.2746	0.1240	0.0980
EFTM Model 1	0.1929	0.2835	0.1000	0.0740
Model 2	0.2239	0.2865	0.1400	0.1060
EFTM Model 2	0.3126	0.4720	0.1680	0.1060

LDA的迭代次数为4000,话题数为20,$\alpha=0.1$,$\beta=0.4$,$\gamma=0.4$。

4 总结与展望

本文针对专家发现问题提出了一种基于话题模型的专家发现方法,该方法能够有效地去除经典专家发现模型中存在的条件独立性假设。同时相对于基于窗口的专家发现模型,新方法的可操作性更强。另外,本文应用了一种文档排序截断技术,该技术在实验中能够有效地降低模型的训练时间。实验表明基于话题模型的专家发现方法其性能较经典专家发现模型有大幅度的提升。

参考文献(References)

[1] Manning C D, Raghavan P, Schutze H. 信息检索导论[M]. 王斌,译. 北京: 人民邮电出版社, 2010.
Manning C D, Raghavan P, Schutze H. Introduction to information retrieval[M]. Beijing: Posts & Telecom Press.

[2] Adafre S F, de Rijke M, Sang E T K. Entity retrieval[C]// Proceedings of RANLP, Bulgaria, September, 2007.

[3] Balog K. People search in the enterprise[D]. University of Amsterdam, 2008.

[4] Rode H. From document to entity retrieval: improving precision and performance of focused text search [D]. University of Twente, 2008.

[5] Balog K, Azzopardi L, de Rijke M. A language modeling framework for expert finding[J]. Information Processing & Management, 2009, 45(1): 1 - 19.

[6] Serdyukov P, Hiemstra D. Advances in information retrieval [M]. Springer Berlin / Heidelberg, 2008.

[7] Balog K, Weerkamp W, de Rijke M. A few examples go a long way: constructing query models from elaborate query formulations [C]// Proceedings of the 31th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR 2008). 2008: ACM.

[8] Petkova D, Croft W B. Proximity-based document representation for named entity retrieval[C]// Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, New York: ACM, 2007.

[9] Fang H, Zhai C. Probabilistic models for expert finding[C]// Proceedings of the 29th European Conference on IR Research, Rome, Italy: Springer-Verlag, 2007.

[10] Balog K, Azzopardi L, Rijke M D. Formal models for expert finding in enterprise corpora[C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA: ACM, 2006.

[11] Guan Z, Miao G, McLoughlin R, et al. Co-occurrence based diffusion for expert search on the web[J]. Knowledge and Data Engineering, IEEE Transactions on, 2012 (99): 1 - 16.

[12] Smirnova E, Balog K. A user-oriented model for expert finding[C]// Proceedings of the 33rd European Conference on Advances in Information Retrieval, Dublin, Ireland: Springer-Verlag, 2011.

[13] Macdonald C, White R W. Usefulness of click-through data in expert search[C]// Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, MA, USA: ACM, 2009.

[14] Craswell N, de Vries A P, Soboroff I. Overview of the trec-2005 enterprise track [C]// Proceedings of the FourteenthText REtrieval Conference (TREC 2005), 2005.

[15] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. J. Mach. Learn. Res., 2003, 3: 993 - 1022.

[16] Griffiths T L, Steyvers M. Finding scientific topics[C]// Proceedings of the National Academy of Science, 2004.

[17] Blei D M, Griffiths T L, Jordan M I, et al. Hierarchical topic models and the nested Chinese restaurant process [J]. Advances in Neural Information Processing Systems, 2003.

[18] Liu J, Zhang Y, Liu B, et al. A DocRank-based document priors model for expert search [C]// Proceedings of Intelligence Science and Information Engineering, Lushan: Atlantis Press, 2012.