ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.2, February 2008, pp.291−300
DOI: 10.3724/SP.J.1001.2008.00291
© 2008 by *Journal of Software*. All rights reserved.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

文本检索的查询性能预测*

郎 皓+, 王 斌, 李锦涛, 丁 凡

(中国科学院 计算技术研究所,北京 100080)

Predicting Query Performance for Text Retrieval

LANG Hao+, WANG Bin, LI Jin-Tao, DING Fan

(Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100080, China)

+ Corresponding author: Phn: +86-10-62600635, Fax: +86-10-62600602, E-mail: langhao@ict.ac.cn, http://www.ict.ac.cn

Lang H, Wang B, Li JT, Ding F. Predicting query performance for text retrieval. *Journal of Software*, 2008,19(2):291−300. http://www.jos.org.cn/1000-9825/19/291.htm

Abstract: Predicting query performance (PQP) has recently been recognized by the IR (information retrieval) community as an important capability for IR systems. In recent years, research work carried out by many groups has confirmed that predicting query performance is a good method to figure out the robustness problem of the IR system and useful to give feedback to users, search engines and database creators. In this paper, the basic predicting query performance approaches for text retrieval are surveyed. The data for experiments and the methods for evaluation are introduced, the contributions of different factors to overall retrieval variability across queries are presented, the main PQP approaches are described from Pre-Retrieval to Post-Retrieval aspects, and some applications of PQP are presented. Finally, several primary challenges and open issues in PQP are summarized.
Key words: information retrieval; query performance prediction

摘 要: 目前,查询性能预测(predicting query performance,简称 PQP)已经被认为是检索系统最重要的功能之一.近几年的研究和实验表明,PQP 技术在文本检索领域有着广阔的发展前景和拓展空间.对文本检索中的 PQP 进行综述,重点论述其主要方法和关键技术.首先介绍了常用的实验语料和评价体系;然后介绍了影响查询性能的各方面因素;之后,按照基于检索前和检索后的分类体系概述了目前主要的 PQP 方法;简介了 PQP 在几个方面的应用;最后讨论了 PQP 所面临的一些挑战.
关键词: 信息检索;查询性能预测
中图法分类号: TP311 **文献标识码**: A

信息检索(information retrieval,简称 IR)研究如何从海量的信息资源中找出满足用户信息需求(information need)的信息子集,它涉及到信息的获取、表示、组织、存储及访问等问题[1].文本检索的任务主要是研究如何从给定的无结构或半结构化文档集中找到与用户查询相关的文档子集,并依据相关度排序把检索结果返回给

* Supported by the National Natural Science Foundation of China under Grant No.60603094 (国家自然科学基金); the National Basic Research Program of China under Grant No.2004CB318109 (国家重点基础研究发展计划(973)); the Beijing Science and Technology Planning Program of China under Grant No.D0106008040291 (北京市科技计划)

Received 2007-06-20; Accepted 2007-08-24

© 中国科学院软件研究所 http://www.jos.org.cn

用户.近几十年来,文本检索的研究取得了很大的进展,许多检索模型被相继提出来并在实践中不断得到检验、改进和验证.但是,大多数检索系统还存在着严重的鲁棒性问题,即针对不同的查询或文档集其性能往往存在较大的差异;即使检索系统的平均性能很好,但对某些查询而言,它的检索结果也不能令人满意[2,3].因此,我们迫切希望检索系统能够自动识别那些检索返回低质量文档集的查询,并对它们做相应的处理.

查询性能预测(predicting query performance,简称 PQP)也称为查询难易度预测(predicting query difficulty),它试图在没有相关信息(即该查询在给定文档集合中的正确答案)的情况下,评估检索系统对于某一查询其返回结果的好坏程度[4].人们通常用 Precision@10 或平均准确率(average precision,简称 AP)作为评价指标,对检索系统针对某一查询的检索结果进行质量评估[3].相对于长期的为提高检索系统的性能在检索模型上的探索,查询性能预测的研究还处于初级阶段.目前,信息检索界已经认识到这一问题的重要性,并把查询性能预测认定为检索系统最重要的功能之一[4].TREC 于 2003 年在 Ad hoc 检索任务的基础上提出了 Robust 任务,旨在关注检索系统的鲁棒性问题.为解决此问题,Robust 任务要求参赛单位预测每个查询的性能,并据此将查询排序[3].信息检索的顶级会议 ACM SIGIR 也于 2005 年引入了 Predicting Query Difficulty Workshop (http://www.haifa.ibm.com/sigir05-qp/index.html).本次 Workshop 的主要议题是,如何通过事先判定某一查询的性能来有针对性地提高检索的性能.这次 Workshop 取得了一定的阶段性成果,产生了一些从查询分析角度来处理鲁棒性问题的思路.同时,马萨诸塞州大学的 CIIR 研究中心、IBM 的 Haifa 研究院和微软剑桥研究院等国际著名信息检索研究机构都对此问题开展了研究,发表了多篇高水平的学术论文[4−7].

查询性能预测能够同时对检索用户和检索系统产生有益的影响.从用户的角度讲,查询性能预测可以提供有价值的反馈信息,这些信息可以指导一个用户的信息查询过程.比如,当检索系统预测检索返回结果的质量较差并将该信息反馈给用户时,用户可以重构他的查询或者更好地与检索系统互动以获得更好的检索结果,比如与检索系统进行相关反馈(relevance feedback).在查询性能预测的帮助下,用户可以很快地构造恰当的查询以获得他所需要的信息.否则,在一个查询检索结果不能令人满意的情况下,用户将花费很多时间读那些不相关的返回文档,并重新构造新的查询以期获得满意的检索结果.同时,查询性能预测能够对检索系统的管理员提供有帮助的信息:管理员可以搜集到目前对于该检索系统比较“难”(查询性能低)并且用户很关注的一类查询话题,并通过加大关于该主题的文档集来达到提高检索系统性能的目的.

从另一个角度来说,在理想情况下,如果一个检索系统能够预测某一查询的性能,那么它就可以自动地调整其参数或者算法来更好地适应不同的查询,从而获得更好的检索性能.特别地,查询性能预测是提高检索系统鲁棒性的重要的第一步[3].目前,我们通常是以检索系统在一个待测试的查询集合上表现的平均性能作为评价某一检索系统好坏的依据.虽然检索系统在为数不多的查询上表现得很差并不会很明显地影响对其平均性能的评价,但是对那些查询感兴趣的用户并不能容忍检索系统返回与其需要不相关的文档.比起那些在大多数查询上表现很好但是偶尔会表现很差的检索系统而言,用户更加倾向于使用那些总是返回可接受相关文档的稳定的检索系统.为了提高检索系统的稳定性,我们首先需要利用查询性能预测技术识别那些返回结果质量低的查询,然后对它们作有针对性的处理.查询性能预测在提高检索系统性能稳定性上的重要作用已经得到信息检索界的认可[3].

本文对文本检索中的查询性能预测方法及其关键技术进行综述,为便于讨论,我们将其称为 PQP 方法.

本文第 1 节介绍 PQP 研究中用到的语料及其评价体系.第 2 节阐述影响查询性能的各个方面的因素.第 3 节总结目前提出的主要的 PQP 方法.第 4 节简要介绍 PQP 的应用.第 5 节对本文进行总结,并讨论 PQP 的下一步研究方向.

1 语料和评价体系

1.1 语 料

为了对不同的 PQP 算法进行比较,必须有一个可供比较的平台,包括公共的评测语料和统一的评价方法.本节主要介绍目前该项研究工作中用到的公共语料.目前,一般都利用 TREC(Text Retrieval Conference)会议

© 中国科学院软件研究所 http://www.jos.org.cn

(http://trec.nist.gov/)提供的文档集作为实验用的语料.主要使用的测试集见表 1.

Table 1 Summary of test collections

表 1 测试集摘要

TREC	Collection	Topic number	Number of document
1+2+3	Disk 1+2	51~150	741 856
4	Disk2+3	201~250	567 529
5	Disk2+4	251~300	524 939
Robust 2004	Disk 4+5 minus CR	301~450; 601~700**	528 155
Terabyte 2004	GOV2	701~750	25 205 197
Terabyte 2005	GOV2	751~800	25 205 197

在 Robust 2004 测试集中包含有最多的 249 个查询,因此,它可以被用来更加可靠地评价一个系统的性能. Terabyte 2004 和 Terabyte 2005 测试集利用了大规模的真实网页作为测试文档集,与传统的 TREC 测试集相比,它们含有更多的噪音信息.因此,PQP 需要具有一定的抗噪音能力,以获得在大规模网页测试集上稳定的预测效果.

通常,TREC 查询含有 3 个域:Title,Description 和 Narrative.在实验中,可以有选择地使用这 3 个域的不同组合:短查询,利用 Title 域;长查询,利用 Description 域.

1.2 评价体系

我们将 PQP 预测查询 Q 性能的结果标记为 $P(Q)$,查询 Q 在检索系统中实际的表现标记为 $R(Q)$,变量 $P(Q)$ 和变量 $R(Q)$的关系标记为 $Relation(P(Q),R(Q))$.这里,我们利用 $P(Q)$和 $R(Q)$这两个变量关系的强弱来评价 PQP 方法的好坏,$Relation(P(Q),R(Q))$越大,则 PQP 预测的查询性能越准;否则,PQP 表现较差.一般利用 Precision@10 或 *AP* 值来衡量一个查询 Q 的实际表现 $R(Q)$.

由于 $P(Q)$和 $R(Q)$这两个变量的分布(distribution)是未知的,因此,为了衡量 $Relation(P(Q),R(Q))$,通常利用 Spearman 等级相关检验(Spearman's rank correlation test)[8]、Kendall 等级相关检验(the Kendall's rank correlation test)[8]和 Pearson 等级相关检验(Pearson's correlation test)[9].Spearman 等级相关检验和 Kendall 等级相关检验是非参数关系检验(non-parametric test)的典型方法.在实验中,Spearman 等级相关性(Spearman's rank correlation)和 Kendall 等级相关性(Kendall's rank correlation)被用来比较基于 $P(Q)$排序的查询序列和基于 $R(Q)$排序的查询序列的相关性.Pearson 等级相关性(Pearson's correlation)则反映了两个变量之间线性相关性的强弱.

这 3 种相关系数的值(correlation coefficient)都是从–1.0 变化到 1.0,–1.0 表示两个变量具有最佳负相关性(perfect negative correlation),1.0 表示两个变量具有最佳正相关性(perfect positive correlation),0 表示两个变量没有相关性.

另外,在评价一个真实 PQP 系统时,还要考虑算法实现的时空效率.

2 查询性能差异性相关因素

一个查询检索结果的好坏与多方面因素相关,包括查询本身的因素、文档集的因素和检索系统的算法及其实现机制[10].在一个典型的信息检索过程中,一个用户根据他的信息需求(information need)构造一个查询提交给检索系统,然后检索系统返回给用户一个相关文档序列(a ranked list).面对互联网上浩瀚的信息,有时用户都不确定自己的需求是什么,因此也就无法构造一个有效的查询;同时,从用户真正的信息需求到最后构造的查询中间存在一个转换过程,这依赖于人的经验知识,可能会出现如下一些情况:查询不能准确、完整地表达用户的信息需求;查询因为具有很大的歧义性(ambiguity)而不能很好地刻画查询的相关文档,故将其从整个文档集合中区分出来等.

查询的性能不仅与查询本身的描述有关,考虑到检索的过程实际上主要是将查询与文档集中的文档依次

** Topic 672 is removed because it has no relevant documents.

© 中国科学院软件研究所 http://www.jos.org.cn

匹配的过程,因此,它还与整个文档集能够提供的信息紧密相关.一种极端的情况是,如果文档集中没有任何与查询相关的文档,查询的性能必然会急剧下降.通常情况下,一个查询的性能与文档集中和查询相关的文档子集(relevance information)的性质息息相关:相关子集的大小;相关子集含有多少个不同的子主题(subtopic),即对相关文档子集进行聚类操作后形成多少个表征不同子主题的团(子主题越多,则相关文档子集越难以完全地被检索系统返回给用户);文档子集与整个文档集中其他文档的区分度(区分度越大,则相关文档子集越容易被返回给用户).

同时,查询的性能与检索系统的算法及其实现机制有关.当检索系统采用不同检索模型时,对同样查询的检索结果会有一定的差异性,这也就影响到了对查询性能的预测.2003 年,由 NIST 举办了 The Reliable Information Access (RIA) Workshop[2],探究不同检索系统机制的因素和不同查询的因素导致检索性能差异的根源性问题.通过对检索失败的查询进行分析,共总结出 10 种检索失败的模式,见表 2,每一种模式都给出了一个相应的例子.10 种失败类别中的 5 类都是因为检索系统没有完全识别查询的所有主题(all aspects).该 Workshop 的结论之一就是,不同检索系统失败的根源是相似的.通常,不同检索系统返回不同的文档,但是在大多数类别上,所有系统都是由同样的原因导致失败.

Table 2 RIA topic failure categorization

表 2 RIA 查询失败模式

Category	Topic example
1. General success—Present systems worked well	Identify documents that discuss in vitro fertilization
2. General technical failure (stemming, tokenization)	Identify systematic explorations and scientific investigations of antarctica, current or planned
3. All systems emphasize one aspect missing another required term	What incidents have there been of stolen or forged art?
4. All systems emphasize one aspect missing another aspect	Identify documents discussing the development and application of spaceborne ocean remote sensing
5. Some systems emphasize one aspect some another; need both	What disasters have occurred in tunnels used for transportation?
6. All systems emphasize one irrelevant aspect missing point of topic	The spotted owl episode in America
7. Need outside expansion of “general” term (Europe for example)	Identify documents that discuss the European conventional arms cut as it relates to the dismantling of Europes arsenal
8. Need QA query analysis and relationships	How much sugar does Cuba export and which countries import it?
9. Systems missed difficult aspect that would need human help	What are new methods of producing steel?
10. Need proximity relationship between two aspects	What countries are experiencing an increase in tourism?

3 查询性能预测方法

本文暂不考虑不同文档集和不同检索系统的实现因素对查询性能的影响,即在给定文档集和检索系统的情况下,探究查询性能变化的相关因素及其预测方法.

基于所利用的特征信息,我们可以把查询性能预测方法分为两大类:基于检索前(pre-retrieval)的方法和基于检索后(post-retrieval)的方法.基于检索前的方法只利用在检索前即可计算获得的静态信息(static information);除了静态信息,基于检索后的方法还可以利用检索结束后才可获得的动态信息(dynamic information).如图 1 所示,静态信息和动态信息是可以被用来预测查询性能的两大类特征.静态特征是可以在检索前计算获得的查询本身的一些特性,比如查询中词项在文档集的统计信息(倒转文档频率 IDF 等)和查询的语言特性(linguistic feature).动态信息只有在检索后才可获得,它主要是分析返回的文档集,并抓住其中的重要特征信息,比如查询中词项在返回文档集中的分布和返回文档集中文档间的相互关系等.通常,静态信息可以利用词项在整个文档集索引中的统计信息快速、简单地计算获得.由于动态信息很好地利用了检索返回文档集中的信息,所以它可以更加可靠地被用于查询性能的预测,其中需要考虑到的是,动态信息的获取需要大量的计算时间.

© 中国科学院软件研究所 http://www.jos.org.cn

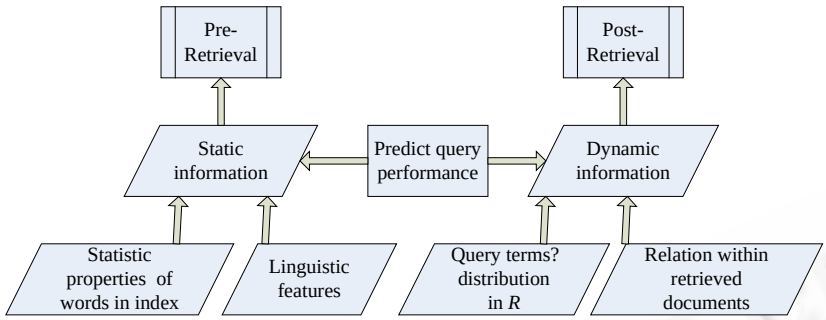

Fig.1 Static and dynamic information of a query to predict query performance

图 1 查询性能预测的静态和动态信息

3.1 基于检索前的查询性能预测方法(pre-retrieval)

基于检索前的方法试图利用查询本身的性质来预测查询的性能.比如,特异性(specificity)和一般性(generality)是一个查询最典型的性质.一个查询具有多强的区分力可以将相关文档与不相关文档区分开来是预测查询性能的最重要的查询性质之一.经研究发现,这些重要的查询性质与查询在整个文档集倒排索引的统计信息有关,也与查询的语言特性相关.

3.1.1 IDF-related方法(倒转文档频率,inverse document frequency)

IDF 是信息检索中常用到的一个因子,它是词项所出现的文档数目的倒数,常常用于反映该词项的区分能力,IDF 越高,区分能力也越高.一些研究者提出利用与 IDF 相关的特征来预测查询的性能.比如,Tomlinson 等人提出用查询词项 IDF 的平均值作为预测的方法[11].He 等人提出了利用查询词项 IDF 的标准方差(standard deviation)来预测查询的性能[12].Plachouras 等人利用 Kwok's inverse collection term frequency(IDF 的一种变形)来表征一个查询词项的质量,然后用查询词项的平均质量预测查询性能[13].以上方法认为,一个查询词项的 IDF 反映了它的质量,检索结果差的查询词项具有相似的 IDF 值.这些方法可以利用查询词项在倒排索引的统计信息,很快、很方便地计算出结果.但是,由于它们没有考虑到检索算法的因素,所以不能很好地预测查询的性能.表 3 给出了不同查询性能预测方法的性能评价.在表 3 中,*越多,表示该方法预测查询性能越准,相同个数的*表示两种方法预测查询性能的能力处于同一水平.由于测试集合与测试条件的差异,指标(*的个数)仅作为方法效果的参考,不能完全作为方法效果之间比较的依据.

3.1.2 Specificity-related方法

一个查询的特异性(specificity)反映了该查询在多大程度上精确地刻画了用户的信息需求并将相关文档与不相关文档区分开来的能力.目前,一些方法利用该性质来预测查询的性能.He 等人提出了一种简化版本的 clarity score 的方法,为了提高计算的速度,查询的语言模型利用查询词项在查询中的相对频率来代替[12].Clarity score 表征了一个查询的特异性,并利用查询的语言模型与整个文档集的语言模型的差异性来衡量,下文还会有具体的介绍.同时,He 等人提出了 query scope 的概念,它表示整个文档集中有百分之多少的文档至少含有 1 个查询词项,从另一个角度刻画了查询的特异性.

3.1.3 基于语言学的方法

不同于以上的从查询词项在文档集中的统计特性的角度来预测查询的性能,Mothe 等人探究了查询的 16 种不同语言特征(linguistic feature)与查询性能的关系[14].这些特征分别从语言的不同角度——词法(morphological)、句法(syntactical)和语义(semantic)出发,考察查询内容与查询性能的关系.其中的两个特征(syntactic links span 和 polysemy value)经实验表明与其查询的性能具有一定的相关性.虽然相关程度有限,但其仍然证明了查询的语言特性是与查询的性能相关的.吕学强等人根据 WordNet 及其附带的 Brown 语料库构造了单词义项分布词典,再把查询中的词项按歧异性大小分为 7 类,通过计算平均词项歧义度来预测查询的性能.该方法经实验表明具有一定的预测能力[15].

© 中国科学院软件研究所 http://www.jos.org.cn

Table 3 Properties and effectiveness for most of query performance predictors

表 3 主要查询性能预测方法的性质和效果

Category		Method	Main features	Speed	Prediction precision	Remark
Pre-Retrieval predictors	IDF related	Mean of query term's IDF[11]	Mean of query term's IDF	Fast	**	Easy and fast to compute, do not predict query performance well
		Variance of query term's IDF[12]	Standard deviation of query term's IDF	Fast	**	
		avICTF[13]	Mean of query term's Kwok's inverse collection term frequency	Fast	**	
	Specificity related	Simplified clarity score[12]	KL divergence between query language model and collection language model, query language model is computed by the relative frequencies of query terms in the query	Fast	***	Easy and fast to compute, predict query performance to some extent, have better performance for short quires
		Query scope[12]	The percentage of documents that contain at least one query term in the collection	Fast	**	
	Linguistic feature	16 kinds of linguistic features[14]	Morphological, syntactical, semantic features	Fast	*	Weak predictors, indicate a promising link between some linguistic characteristics and query performance
		Total ambiguity of query terms[15]	Mean of ambiguity of query terms	Fast	*	
Post-Retrieval predictors	Homogeneity	Document similarity[16]	Similarity between pairs of retrieved documents	Slow	**	State of arts effectiveness, high computational complexity
		Document perturbation[5]	Whether the retrieved documents are a random set of points or a clustered set of points	Slow	******	
		Robustness score[6]	How stable the ranked lists is in the presence of uncertainty in the ranked documents	Slow	******	
	Separation	Clarity score[7]	KL divergence between query language model and collection language model	Normal	****	Clarity score and DFR are similar, time features can increase prediction power
		DFR[17]	KL divergence between a query term's frequency in retrieved documents and the frequency in the whole collection	Normal	****	
		Clarity score+ time features[18]	Time features are integrated to the clarity score	Normal	****	
	Others	A histogram-based predictor and a decision tree based predictor[4]	Document frequency of query terms, the overlap of top retrieval results between using the full query and the individual query terms	Fast	****	The main advantage is its simplicity and that it can be computed efficiently during query execution, the performance is relied on the limited training data

3.2 基于检索后的查询性能预测方法(post-retrieval)

基于检索后的方法试图分析检索系统返回的,并被其认为与查询相关的文档子集的性质,包括返回文档子集中文档间的相互关系以及返回文档子集与整个文档集的关系等.

3.2.1 对检索结果同质性分析的方法

聚团性假设(cluster hypothesis)[19]指出,与一个查询相关的文档,彼此间会具有一定的潜在关系.因此,与那些不相关的文档相比,与查询相关的文档会具有一定的相似性或者称为同质性.相关文档会形成一个团(form a group),并与那些不相关文档区分开来.在实际中,我们将聚团性假设用另一种形式来刻画:如果检索返回的文档间比较相似,也就是说具有较高的同质性,我们就认为返回的文档子集与查询主题比较相关;从另一个角度来说,如果检索返回的文档之间没有任何相似性,即检索返回的结果具有很高的随机性(randomness),那么我们就认为返回文档集中可能含有很多不相关的文档,这一查询的检索返回结果质量较差.

从分析检索结果的同质性或者说相似性出发,Kwok 等人提出了利用检索返回文档之间的相似度来预测查询性能的方法[16].该方法认为,如果检索返回的文档都是与查询相关的,那么那些文档之间应该具有很高的相似

© 中国科学院软件研究所 http://www.jos.org.cn

性.该方法的实验结果表明它预测查询性能的准确率较低.

Vinay 等人进一步研究了返回文档的内聚性(cluster tendency),提出了 4 种预测查询性能的方法[5].这些方法关注于返回文档的几何学特性(即返回文档间的相互关系,比如内聚性等),探究返回文档是形成随机分布的点集(a random set of points)还是具有内聚性的点集(a clustered set of points).从聚类的观点来看,这些方法着重评价返回文档间的紧密程度(compactness).Document perturbation 是这 4 种方法中预测效果最好的,该方法的主要思路是,考虑检索返回的一个文档子集,取其中任意一个文档作为伪查询(pseudo query),将文档子集中的文档按照与伪查询的相似度进行排序,那个作为伪查询的文档应该在新的返回序列中排第一位.但是,在给该伪查询加入噪音的情况下(即按照一定的规则,改变该伪查询中不同词项的频率),那个文档在新的返回序列中可能将不再排在第一位,而具有一个新的序列值.Document perturbation 方法就是根据随机分布的点集和具有内聚性的点集在一定噪音(扰动)影响下的不同表现来预测查询的性能.该方法在 TREC disk4 和 TREC disk5 的 200 个长查询中得到了目前公布的最好的预测效果.这些方法获得了较大的成功,能够较好地预测查询的性能.但值得注意的是,它们都需要计算返回文档的几何学特性,其中包含了大量的文档间两两相似度的计算.因此,这些方法计算比较耗时,并不能直接应用于实际的应用场景中.

Zhou 等人提出了一个新的度量返回文档集稳定性的概念 ranking robustness,并提出了一种基于统计的方法 robustness score 来计算 ranking robustness 的值[6].该方法的大体思路是,针对一个查询检索返回的文档列表(ranked list),对该列表中的每个文档加入一定的噪音后,再次用相同的查询和相同的检索系统对加入噪音后的文档按相关度进行排序得到新的文档列表,前后两个列表的相似性反映了初始返回文档集的稳定性,即可得到 robustness score.实验表明,robustness score 与查询的性能有着很稳定的相关性,在 TREC 的多个测试集中,对于短查询,robustness score 都比较准确地预测了查询的性能.

3.2.2 对检索结果突显性分析的方法

聚团性假设中指出,与查询相关的文档存在相似性,它们彼此之间紧密地形成一个团,同时会与文档集中的不相关文档区分开来.检索返回的结果与整个文档集的区分度越大,我们就认为检索结果的突显性(突显性强调了检索返回文档集与整个文档集合的区分度)越强,检索的结果较好;相反地,如果检索结果的突显性较差,即检索的结果与整个文档集很相似,那么我们就认为检索的结果里面含有很多不相关的文档.

Cronen-Townsend 等人提出了 clarity score 的方法,该方法是首先成功预测查询性能的方法之一[7].Clarity score 实际上是查询的语言模型和整个文档集语言模型的 Kullback-Leibler 距离.Clarity score 刻画了查询的特异性,如果两个语言模型差异很大,我们就认为查询能够识别出整个文档集中小部分相关的文档,因此,查询具有很高的特异性并获得较高的 clarity score 值.

Amati 等人提出了一种类似于 clarity score 的方法.该方法以查询中词项在检索返回文档集中的分布与在整个文档集中的分布的 KL-divergence 来预测查询的性能[17].Diaz 等人扩展了 clarity score 方法,加入了时间特征(time features).实验表明,加入时间特征可以提高 clarity score 预测查询性能的准确度[18].

3.2.3 其他基于检索后的方法

除了对检索结果的同质性和突显性进行分析以外,一些学者还尝试从其他角度来预测查询的性能.Elad Tom-Tov 等人提出了基于柱状图的预测方法(histogram-based predictor)和基于决策树的预测方法(decision tree based predictor)[4].这两种预测方法主要利用了两类特征:查询词项的文档频率(document frequency),基于整个查询检索返回的文档列表与基于查询中一个词项检索返回的文档列表的耦合度.这些方法认为,如果查询中的词项都赞同初始检索返回的文档,那么检索的性能较好.这些方法的优点在于其简单性并易于实施,它们可以在查询处理过程中高效地计算出结果.但是,这些方法包含一个学习过程,正如该方法的作者所称,它们的性能非常依赖于有限的训练数据集.

3.3 小 结

由于基于检索前的查询性能预测方法只利用了查询中词项在文档集的统计信息或者查询的语言特性,它们都可以很简单、很快地计算出来,因此,它们非常满足实际应用场景中对速度和效率的要求.但是,它们都没有

© 中国科学院软件研究所 http://www.jos.org.cn

利用任何检索返回的文档信息,所以,它们都不能很好地预测出查询的性能.现在,所有基于检索前的查询性能预测方法都只取得了比较有限的成功,即它们的计算速度很快而预测的准确度却不够高.

基于检索后的方法深入分析了检索返回的文档集,并提取其中重要的信息用于查询性能预测,因此它们取得了较好的预测效果.总的来说,目前提出的该类方法主要是分析检索返回文档集的几何特性,返回文档之间的相似度如何?是不是一个凝聚成很紧密的团?是不是与其他不相关的文档具有很好的区分度?按照这个思路,一些学者着重分析了检索返回文档的同质性和突显性,并提出了一些高效的查询性能预测方法(见表 3).其中,robustness score 方法在多个 TREC 测试集中对于短查询都取得了很好的性能预测效果,document perturbation 方法在 TREC disk4 和 TREC disk5 的 200 个长查询中获得了目前公布的最好的预测准确率.尽管基于检索后的方法取得了一定的成功,但有一点需要注意:由于这些方法主要是考察检索返回文档的几何学特性,需要计算返回文档两两间的相似度,这就要求具备大量的计算资源,因此,它们还不能直接应用到实际的应用场景中.

4 查询性能预测的应用

4.1 有选择的查询扩展(selective automatic query expansion)

查询扩展(query expansion,简称 QE)是为了解决查询词的不匹配问题,利用用户初始查询返回的结果,通过一定的策略在初始查询中加入一些与主题相关的词,以达到提高检索性能的目的[20].实际上,查询扩展技术并没有应用在大多数实际的系统当中,这是因为该技术会导致一些查询失效甚至于损害它们的检索结果.注意到查询扩展技术是基于初始检索返回文档与查询相关的假设的,即只有在初始返回文档都与查询相关或者说初始检索结果较好的情况下,查询扩展技术才会产生积极的作用.因此,我们可以利用查询性能预测技术识别那些性能较差的查询,只是利用它们的初始检索结果;对于其他性能较好的查询,对它们进行查询扩展,以获得更好的检索结果.因此,我们可以利用查询性能预测技术进行有选择的查询扩展,这样,一方面解决了查询扩展对部分查询的失效的问题,同时也提高了检索系统的鲁棒性和稳定性.

4.2 缺失文档发现(detecting missing content)

有这样一类查询,文档集中没有任何文档与其相关,因此检索返回的文档都将是与查询无关的,我们将这类查询定义为缺少相关文档的查询(missing content query,简称 MCQ).查询性能预测技术的一个很好的应用就是识别 MCQ,这样对于用户和检索系统都有积极的作用.前者可以知道该检索系统是否含有他所需要的信息,后者可以搜集到用户目前关注但是检索系统缺少的一类话题.文献[4]中提到,可以利用查询性能预测的结果作为特征,利用机器学习的方法从已标注的查询中学习得到一定的规则,并利用这些规则识别 MCQ.实验表明[4],基于查询性能预测的方法对于识别 MCQ 具有很好的效果.

4.3 分布式信息检索的结果融合(merging the results in a distributed IR system)

一个典型的分布式信息检索的场景是,将一个查询在多个不同的数据集中进行检索,每个数据集返回一个相关文档列表,然后将这些文档列表按一定的策略进行融合,形成一个新的独立的文档列表返回并呈现给用户.在将多个文档列表进行融合时,我们需要将来自不同数据集的文档列表根据其重要性设置不同的权重,然后根据相应的权重将结果融合.这里,我们可以利用查询性能预测技术,对每一个数据集返回的文档列表按照其与查询的相关度(也就是该列表的质量)进行打分,并以该打分作为返回列表的权重.文献[4]中提到,利用该权重计算方法的分布式信息检索的效果优于传统的基于 COIR(collection retrieval inference network)[21]的效果.

5 总 结

为了解决检索系统的鲁棒性问题,查询性能预测的研究受到了人们的日益关注.查询性能预测已经被认为是检索系统最重要的功能之一.本文对查询性能预测中的主要方法和关键技术进行综述,介绍了目前进行研究时主要采用的语料和评价体系,阐述了影响查询性能的若干因素,并按照基于检索前和检索后的分类方法概述

© 中国科学院软件研究所 http://www.jos.org.cn

了目前主要的查询性能预测方法,最后介绍了该项技术在 3 个方面的应用.

查询性能预测这个问题的提出,为信息检索研究领域带来了新的机遇,同时也遇到了一些新的挑战.就目前来看,我们认为仍有以下问题有待进一步探讨和研究:

查询性能存在差异性的根源是什么?关于这一点,目前还没有一个明确的答案,只是有人提出了一个查询的性能是与查询、文档集以及检索系统相关的,它们之间的关系则还没有形式化的分析.

进一步研究高性能的查询性能预测算法.目前,许多学者已经提出了很多有效的查询性能预测方法,比如 robustness score 和 document perturbation,它们都能比较准确地预测出查询的性能,但是它们需要大量的计算时间,还不能直接应用于实际的应用场景当中.所以,我们迫切需要进一步研究出高速、高准确率的查询性能预测方法.

目前,我们提出的方法都是在一个数据集中针对某一个检索系统预测不同查询性能的相对高低,这只是对查询性能的一个方面的刻画.同时,我们希望能够预测在一个数据集中某一查询在不同检索系统下性能的相对高低.这样,我们就可以利用多特征信息(multi-evidence)来提高检索系统的性能及其稳定性.比如,目前已经提出了很多经典的检索模型,有的模型对某些查询会获得最优的结果,其他模型可能会对其他查询获得最优的效果,如果我们对某一查询能够自动地选择最合适的检索模型,那么我们就可以获得稳定且鲁棒的检索性能.利用查询性能预测技术预测某一查询在不同模型下的性能,然后选择最合适的模型,就可以实现此目的.

最后,我们希望将查询性能预测技术应用于基于查询分析(query analysis)的自适应信息检索系统(adaptive information retrieval)的研究当中.检索系统的性能依赖于其应用的特征信息及其检索模型,其中,检索模型的参数往往会对检索性能产生重要的影响.通常,我们会花费大量的人力去人为地调节参数,或者根据有限的查询集合训练出参数.这里,无论是人为调节的参数还是根据一个固定查询训练集学习出来的参数,它们都是试图将检索系统的平均性能最优化,而没有考虑到不同查询的差异性,即不同类型的查询也许需要不一样的参数.因此,我们可以利用查询性能预测的结果相应地调节检索系统利用的特征信息及其参数,这样我们就可能得到更加稳定且鲁棒的检索性能.第 4.1 节介绍的有选择的查询扩展即是查询性能预测技术在自适应信息检索系统研究中的一次有意义的尝试.除此之外,我们相信查询性能预测技术在自适应信息检索系统中的研究将有更为广阔的应用前景.

References:

[1] Yates B, Neto R. Modern Information Retrieval. New York: ACM Press, 1999.

[2] Harman D, Buckley C. The NRRC reliable information access (RIA) workshop. In: Sanderson M, Jarveln K, Allan J, Bruza P, eds. Proc. of the 27th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Sheffield: ACM Press, 2004. 528−529.

[3] Voorhees EM. Overview of the TREC 2004 robust retrieval track. In: Online Proc. of the 2004 Text Retrieval Conf. (TREC 2004). 2004. http://trec.nist.gov/pubs/trec13/t13_proceedings.html

[4] Yom-Tov E, Fine S, Carmel D, Darlow A. Learning to estimate query difficulty: Including applications to missing content detection and distributed information retrieval. In: Proc. of the 28th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Salvador: ACM Press, 2005. 512−519.

[5] Vinay V, Cox IJ, Milic-Frayling N, Wood K. On ranking the effectiveness of searches. In: Proc. of the 29th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2006. 398−404.

[6] Zhou Y, Croft WB. Ranking robustness: A novel framework to predict query performance. In: Proc. of the 15th ACM Int'l Conf. on Information and Knowledge Management. Arlington: ACM Press, 2006. 567−574.

[7] Cronen-Townsend S, Zhou Y, Croft WB. Predicting query performance. In: Proc. of the 25th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Tampere: ACM Press, 2002. 299−306.

[8] Gibbons JD, Chakraborty S. Nonparametric Statistical Inference. 3rd ed., New York: Marcel Dekker, 1992.

[9] Kreyszig E. Advanced Engineering Mathematics. John Wiley & Sons, Inc., 1997.

[10] Carmel D, Yom-Tov E, Darlow A, Pelleg D. What makes a query difficult? In: Proc. of the 29th Annual Int'l ACM SIGIR Conf. on

© 中国科学院软件研究所 http://www.jos.org.cn

Research and Development in Information Retrieval. New York: ACM Press, 2006. 390−397.

[11] Tomlinson S. Robust, Web and terabyte retrieval with hummingbird search server at TREC 2004. In: Online Proc. of the 2004 Text Retrieval Conf. (TREC 2004). 2004. http://trec.nist.gov/pubs/trec13/papers/humingbird.robust.web.tera.pdf

[12] He B, Ounis I. Inferring query performance using pre-retrieval predictors. In: Apostolico A, Melucci M, eds. String Processing and Information Retrieval, 11th Int'l Conf., SPIRE 2004. LNCS 3246, 2004. 43−54.

[13] Plachouras V, He B, Ounis I. University of Glasgow at TREC 2004: Experiment in Web, robust, and terabyte tracks with terrier. In: Online Proc. of the 2004 Text Retrieval Conf. (TREC 2004). 2004. http://ir.dcs.gla.ac.uk/terrier/publications/glasgowTrec2004.pdf

[14] Mothe J, Tanguy L. Linguistic features to predict query difficulty. In: Proc. of the 29th Annual Int'l ACM SIGIR 2005 Workshop on Predicting Query Difficulty—Methods and Applications. http://www.haifa.il.ibm.com/sigir05-qp/index.html

[15] Lü XQ, Lai ZG, Sun B, Yu SW. Evaluation of topic difficulty. Journal of Tsinghua University (Science and Technology), 2005, 45(S1):1833−1837 (in Chinese with English abstract).

[16] Kwok KL, Grunfeld L, Dinstl N, Deng P. TREC 2005 robust track experiments using PIRCS. In: Online Proc. of the 2005 Text Retrieval Conf. (TREC 2005). 2005. http://trec.nist.gov/pubs/trec14/papers/queensc-kwok.robust.pdf

[17] Amati G, Carpineto C, Romano G. Query difficulty, robustness and selective application of query expansion. In: Proc. of the ECIR 2004. 127−137.

[18] Diaz F, Jones R. Using temporal profiles of queries for precision prediction. In: Proc. of the 27th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Sheffield: ACM Press, 2004. 18−24.

[19] van Rijsbergen CJ. Information Retrieval. 2nd ed., London: Butterworths, 1979.

[20] Xu JX, Croft WB. Query expansion using local and global document analysis. In: Proc. of the 19th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Zürich: ACM Press, 1996. 4−11.

[21] Callan JP, Lu ZH, Croft WB. Searching distributed collections with inference networks. In: Proc. of the 18th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Seattle: ACM Press, 1995. 21−28.

附中文参考文献:

[15] 吕学强,赖治国,孙斌,俞士汶.检索主题难易度评价.清华大学学报(自然科学版),2005,45(S1):1833−1837.

郎皓(1982－),男,江苏南京人,博士生,主要研究领域为信息检索,查询分析.

李锦涛(1962－),男,博士,研究员,博士生导师,CCF 高级会员,主要研究领域为数字媒体处理技术,虚拟现实技术,普适计算技术.

王斌(1972－),男,博士,副研究员,主要研究领域为信息检索.

丁凡(1980－),男,博士生,主要研究领域为信息检索.

© 中国科学院软件研究所 http://www.jos.org.cn