软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(12):3130−3139 [doi: 10.13328/j.cnki.jos.004744]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于查询子主题分类的多样性搜索评价方法*

陈 飞[1,2,3], 刘奕群[1,2,3], 张 敏[1,2,3], 马少平[1,2,3]

[1](智能技术与系统国家重点实验室(清华大学),北京 100084)
[2](清华信息科学与技术国家实验室(筹)(清华大学),北京 100084)
[3](清华大学 计算机科学与技术系,北京 100084)
通讯作者: 陈飞, E-mail: chenfei27@gmail.com, http://www.csai.tsinghua.edu.cn/

摘 要: 多样化检索结果的评测通常假设一个查询词包含多个权重各不相同的用户子意图,并在此假设的基础上对检索结果进行评测.虽然大多数已经存在的多样化检索评测方法利用了这些特性对检索结果进行评测,但在评测过程中,它们都忽略了查询子意图的类型信息;而不同类型的查询子意图对信息需求具有不同的特点.首先,通过引入衰减函数对这种特点进行描述,进而对用户子意图的分类方法进行抽象;在此基础上,提出了利用查询子意图类型信息进行多样化检索结果评测的框架,该框架定义了利用查询子意图类型信息进行多样化检索评测的方法应该具有的结构;然后,讨论了在用信息类和导航类作为子意图分类方法的前提下,其对应的衰减函数的形式;最后,在TREC 与 NTCIR 测试集上的实验结果表明了所提出方法的有效性.

关键词: 多样化检索评测;分类;检索结果多样化
中图法分类号: TP18

中文引用格式: 陈飞,刘奕群,张敏,马少平.基于查询子主题分类的多样性搜索评价方法.软件学报,2015,26(12):3130−3139. http:// www.jos.org.cn/1000-9825/4744.htm

英文引用格式: Chen F, Liu YQ, Zhang M, Ma SP. Subtopic taxonomy-aware framework for diversity evaluation. Ruan Jian Xue Bao/Journal of Software, 2015,26(12):3130−3139 (in Chinese). http://www.jos.org.cn/1000-9825/4744.htm

Subtopic Taxonomy-Aware Framework for Diversity Evaluation

CHEN Fei[1,2,3], LIU Yi-Qun[1,2,3], ZHANG Min[1,2,3], MA Shao-Ping[1,2,3]

[1](State Key Laboratory of Intelligent Technology and Systems (Tsinghua University), Beijing 100084, China)
[2](Tsinghua National Laboratory for Information Science and Technology (Tsinghua University), Beijing 100084, China)
[3](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: To evaluate search result diversification, which is supposed to meet different needs behind a same query, a number of evaluation frameworks are proposed. Although most of these frameworks take the probability distribution of information needs (usually called subtopics) underlying a query topic into account, they usually do not consider the subtopic taxonomy information. In this paper, the decay function is first introduced to take the subtopic taxonomy information into account. And then based on the decay function, a novel framework called the subtopic taxonomy-aware (STA) framework is proposed to define the structure that the taxonomy-aware diversity evaluation metrics would have. The decay functions used for the informational and navigational subtopics are also discussed. Experiments based on TREC and NTCIR test collections show the effectiveness of the proposed method.

Key words: diversity evaluation; taxonomy; diversified search

* 基金项目: 国家自然科学基金(61532011, 61472206); 国家重点基础研究发展计划(973)(2015CB358700)
Foundation item: National Natural Science Foundation of China (61532011, 61472206); National Program on Key Basic Research Project of China (973 Program) (2015CB358700)
收稿时间: 2013-12-13; 修改时间: 2014-09-26; 定稿时间: 2014-10-17

搜索引擎是获取互联网海量信息的主要工具.用户的信息需求促使其构造查询词并提交给搜索引擎,且期望搜索引擎能够在尽可能靠前的位置返回能够满足其信息需求的网页.然而,数量巨大的互联网用户意味着搜索引擎需要处理不同用户的查询意图(即信息需求);并且即使是同一个查询词,不同用户可能都意味着不同的意图.比如用户向搜索引擎提交了查询词“苹果”,有的用户可能需要获取与苹果公司产品相关的网页,有的用户可能需要获取介绍苹果这种水果的网页,而有的用户则可能需要观看名为“苹果”的电影.另一方面,由于用户的惰性,其提交给搜索引擎的查询词往往比较简短,且呈越来越短的趋势[1],以至于搜索引擎很难从查询词本身获取有关用户查询意图的更多信息.为合理地解决这样的问题,研究者们提出了多样化检索的方法,并有越来越多的研究者、搜索引擎公司参与到这样的研究中[2−5].多样化检索的目的在于:对用户提交的一个查询词,该方法返回给用户的检索结果能够在尽可能靠前的位置满足尽可能多的不同用户的信息需求,且最大程度的去除检索结果中的冗余信息[6,7].在多样化检索方法的研究中,一个查询词可能包含的用户查询意图被称作查询的子意图(subtopic).

另一方面,合理的评测方法不仅作为衡量搜索引擎检索结果好坏的评价标准,其在搜索引擎的参数调优方面也占据着重要作用[8].传统检索结果(即,非多样化检索)的评测方法,诸如 P(precision),MAP[9],MRR[10],*n*DCG[11]等是建立在每个查询词对应于一个用户意图的假设下对检索结果进行评测,因此其已不再胜任多样化检索结果的评测.为此,研究者们相继提出α-*n*DCG[12],Intent Aware Metrics(IA 类评测方法,如 *n*DCG-IA,ERR-IA 等)[13,14]和 D#-measures[8,15]等评测方法,以解决多样化检索结果的评测问题.这些评价方法假设一个查询词可能包含多个不同权重的子意图,且检索结果中的网页对这些子意图有着不同程度的相关性,并以此来评测一个多样化检索结果的好坏.

然而,上述评测方法都忽略了一个重要信息,即,不同类型的查询子意图需要不同的策略来满足.而上述方法在进行多样化评测时,并未考虑这样的不同.如果用 Broder 等人[16]提出的按照信息类、导航类和事务类来区分同一个查询的不同子意图,则对于导航类子意图,用户正在搜寻的某个特定网页(比如官网)即可满足用户的信息需求,且其他的与此子意图相关的网页在多样化检索任务中则会被认为是冗余信息,应当在评测中得到惩罚;而对于信息类或事务类子意图,多个相关网页能更加全面的满足用户的信息需求.在关于多样化检索结果评测的研究中已初步证明[17]:即使仅按照信息类和导航类对查询词的子意图进行分类,并将分类信息用于多样化检索的评测,即可得到更加合理的评测结果.但在文献[17]中,作者仅简单地利用了查询子意图的类型信息,即:对信息类子意图的评测未做任何改变,而对导航类子意图的评测则只考虑检索结果中第 1 个相关的文档.在本文中,我们考虑如何更加充分地利用查询子意图的分类信息,并提出一个多样化检索的评测框架,该框架展示了合理利用查询子意图的分类信息进行多样化检索结果评测的方法所对应的结构.通过引入衰减函数来描述不同类型查询子意图在多样化检索评测中的特性,并泛化框架中查询子意图的分类方法,该框架可以对已有的多样化评测方法按照任何分类方法进行重新定义.重新定义后的评测方法则可以根据不同类型的查询子意图来评测多样化检索结果.为了与现有的多样化评测方法进行对比,我们还讨论了如何在考虑信息类和导航类作为查询子意图的分类方法的假设下,为此新框架定义对应的衰减函数.最后,用实验验证了新框架下的评测方法比现有的方法更加合理.

本文第 1 节介绍已有的多样化检索结果评测的相关工作.第 2 节介绍提出的多样化检索评测框架.第 3 节讨论在新框架下按照信息类和导航类对查询子意图进行分类时对应的衰减函数.第 4 节介绍如何评价多样化检索评测方法本身以及本文实验的数据集、实验过程、实验结果及分析.第 5 节给出总结与未来工作.

1 相关工作

在对传统检索结果进行评测时,我们往往假设每个查询词只对应着唯一的一个查询意图,而参与评测的文档会根据其与该查询意图的相关程度进行等级标注.文档的相关等级通常用 0~h 之间的数字表示.如果 h=1,则此标注被称为二级标注;若 h>1,则称为多级标注.根据文档标注的相关等级,可以计算出该文档在评测中的文档增益(document gain),通常记为 $G(d)$.根据评测方法的不同,$G(d)$的计算方式亦有所不同.如在 *n*DCG 中[10], $G(d)$

的计算公式为

$$G(d)=\frac{2^x-1}{2^h} \tag{1}$$

其中,x 为文档 d 的标注等级.注意到,$G(d)$的计算只和当前评测文档本身的相关性有关.当获得了所有参与评测文档的增益之后,则可以根据不同的评测方法,计算出一个检索结果在该评测方法中的得分.比如,P 和 nDCG 的计算式分别如公式(2)、公式(3)所示.

$$P@l=\frac{\sum_{r=1}^{l}J(d_r)}{l} \tag{2}$$

$$nDCG@l=\frac{\sum_{r=1}^{l}G(d_r)/\log(r+1)}{\sum_{r=1}^{l}G^*(d_r)/\log(r+1)} \tag{3}$$

其中,d_r 表示检索结果中的第 r 个网页,$J(d_r)$为指示网页 d_r 与查询是否相关的示性函数.即:当 d_r 与查询相关时,$J(d_r)=1$;否则,$J(d_r)=0$.$G^*(d_r)$表示与查询词相关的网页的理想排序评测得分,用于归一化检索结果的评测得分.

当对多样化后的检索结果进行评测时,由于查询词所包含的意图不再是传统评测方法中所认为的仅有一个,而可能是多个且具有不同权重,因此参与评测的文档需要按照与查询子意图的相关性分别进行标注.同样按照标注的相关等级,我们可以计算出文档关于某一个具体的查询子意图 s_i 的文档增益,记为 $G_i(d)$.$G_i(d)$在不同的多样化检索评测方法中具有不同的定义,下面将分别介绍.

1.1 α-nDCG

当利用α-nDCG[12]进行评测时,用户的信息需求被称为 Nugget.每个查询词具有多个 Nuggets 对应着多个查询子意图,而每个文档也可能与多个这样的 Nuggets 相关.α-nDCG 还考虑了文档标注出错的概率α,即,文档与某个 Nugget 不相关却被错误地标注为相关的概率(因为评测时只有被标注为相关的文档才参与计算,因此α-nDCG 没有考虑当文档与某个 Nugget 相关却被错误地标注为不相关的情况).α-nDCG 的文档增益 $G(d)$可按照公式(4)进行计算.

$$G(d)=\sum_{s_i\in C_d}(1-\alpha)^{N_i} \tag{4}$$

其中,C_d 表示文档 d 所包含的 Nuggets 集合,N_i 表示位于文档 d 之前的文档集中包含 s_i 的文档数量.将公式(4)替换公式(3)中的 $G(d_r)$,即可得到α-nDCG.

1.2 IA类评测方法

在对传统检索结果进行评测时,MAP,nDCG,ERR 等方法都假设每个查询词只包含一个查询意图,而多样化检索假设每个查询词包含了多个权重不等的子意图.基于这样的差异,Agrawal 等人[13]提出:可利用传统评测方法按照查询词的每个子意图依次对多样化检索结果进行评测,然后将评测结果按照对应子意图的权重线性加权,即可实现多样化检索结果评测.这样的评测方法统称为 IA 类评测方法(IA measures).如果其中的传统评测方法采用的是 nDCG,则其对应的 IA 类方法称为 nDCG-IA,其计算式见公式(5).

$$nDCG\text{-}IA@l=\sum_{s_i\in C}P(s_i\mid q)\times nDCG(s_i)@l \tag{5}$$

其中,C 为查询 q 的子意图集合,$nDCG(s_i)$表示按照子意图 s_i 对检索结果进行评测时的 nDCG 值,$P(s_i|q)$为子意图 s_i 的权重.

1.3 D#-Measure

在多样化检索的评测中,用 $G_i(d)$表示文档相对于查询词的某个子意图的增益.为能够衡量一个文档相对于这个查询词的总体增益,Sakai 等人[8,15]提出将文档增益 $G_i(d)$和其对应的查询子意图的权重线性合并,得到能够用于衡量文档与查询词相关性的全局增益(global gain),其形式化定义见公式(6).然后,再以此全局增益替换传统评测方法中的文档增益,得到评测方法 D-Measure.D-Measure 虽然考虑了文档与查询词的各个子意图之间的

相关性,但其并未考虑来自检索结果中位于当前文档之前的文档集合的影响.为实现这一点,Sakai 等人提出 I-rec,见公式(7),其中,N_l 表示检索结果的前 l 个网页中与至少一个查询子意图相关的网页数量,并将 D-Measure 和 I-rec 线性加权,即得到 D#-Measure,其形式化的定义见公式(8),其中,λ作为平衡因子一般设为 0.5.

$$GG(d)=\sum_{s_i\in C}P(s_i\mid q)\times G_i(d) \tag{6}$$

$$I\text{-}rec@l=\frac{N_l}{l} \tag{7}$$

$$D\#\text{-}Measure=\lambda D\text{-}Measure+(1-\lambda)I\text{-}rec \tag{8}$$

1.4 DIN#-Measure

从公式(6)可以看出,文档的全局增益并未考虑不同类型查询子意图的差异.比如:导航类查询子意图比较明确,因此只需要返回与之对应的目标网页即可满足用户的需求;而信息类子意图则需要更多的相关网页,以便能够全面地覆盖用户的信息需求.根据如上的观察,Sakai[17]提出在对多样化检索结果进行评测时,应当按照导航类与信息类区分查询子意图;并提出对于导航类子意图,仅评测第 1 个与之相关的网页,而认为后续与之相关的网页的增益为 0.在计算文档的全局增益时,Sakai 等人采用如公式(9)所示的示性函数来描述导航类子意图的这种特性,其中,N_i 表示已发现的与子意图 s_i 相关的网页数量.对于信息类需求,则按正常方法评测网页.这就是公式(10)的含义.其中,C_i 表示查询词的信息类子意图集合,而 C_n 表示导航类子意图集合.通过这样计算得到的文档增益替换 D#-Measure 中的全局增益,即得到 DIN#-Measure.

$$isnew_i=\begin{cases}1, & N_i=0\\ 0, & N_i>0\end{cases} \tag{9}$$

$$GG^{DIN}(d)=\sum_{s_i\in C_i}P(s_i\mid q)G_i(d)+\sum_{s_i\in C_n}isnew_iP(s_i\mid q)G_i(d) \tag{10}$$

2 利用查询子意图的类型信息对多样化检索结果进行评测的方法

在搜索引擎排序算法中,对查询词按导航类和信息类进行分类并针对不同的类型采用不同排序策略,已经被证明是提高搜索引擎检索质量的有效手段[5,16,18,].同样地,在多样化检索结果评测方面,Sakai 等人的工作[17]也初步证明:仅简单地用示性函数来描述导航类查询子意图的特性,即可得到更合理的评测方法.这些工作都启示我们,应当更加充分地利用查询子意图的类型信息进行多样化检索评测.本文从用户信息需求的角度入手,讨论了用户在浏览检索结果时,不同类型的信息需求被满足的程度随着相关网页的增多而具有不同的变化趋势,并引入衰减函数来描述这样的趋势;然后,通过扩展现有评测方法中的文档增益,重新定义一个考虑了查询子意图类型信息的文档增益(subtopic-aware gain,简称 STA-G);最后,以 STA-G 替换现有评测方法中的文档增益,得到利用查询子意图类型信息对多样化检索结果进行评测的方法.由于在此方法中引入了衰减函数,用以抽象子意图的不同分类方法所具有的特性,从而使得其不依赖于具体的子意图分类方法,因此,该方法从框架层面定义了利用查询子意图类型信息进行多样化检索结果评测的方法应该具有的结构,称为 Subtopic-Aware 框架,简称 STA 框架.其系统框图如图 1 所示,我们将 STA 框架下的评测方法称为 STA-Measure.

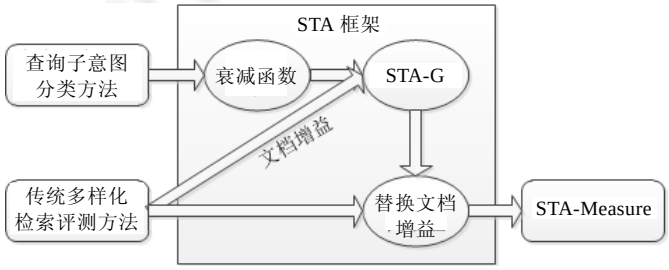

Fig.1 Procedure of the STA-Measure evaluation framework
图 1 STA 框架示例图

2.1 衰减函数

在 STA-Measure 中,衰减函数用于描述在某一特定的分类方法下不同类型子意图被满足的程度随着相关文档增加的变化趋势,并根据子意图被满足的程度来对后续文档的增益进行衰减.因此从总体上来看,不管何种分类方法对应的衰减函数,都应该是查询子意图的类型以及在用户浏览过程中已经浏览过的文档集合的函数.我们可以将其形式化地定义为$f_k(s_i,S_d)$,其中,k 为查询子意图 s_i 的类型,而 S_d 为位于文档 d 之前的文档集合.由于$f_k(s_i,S_d)$应当根据 S_d 中的文档满足查询子意图 s_i 的程度来对后续与 s_i 相关的文档的增益进行衰减,因此,$f_k(s_i,S_d)$应该具有单调不增的特性.

在 STA-Measure 框架的实际应用中,应当首先选择一种能够对查询子意图进行合理分类的方法,并针对该分类方法中每一种类型的特性,定义对应的衰减函数.

2.2 文档增益

当定义了衰减函数之后,STA-Measure 框架在计算文档的增益时,不仅仅计算文档自身与查询子意图的相关性(即 $G_i(d)$),还将来自集合 S_d 所带来的影响一并考虑,从而计算文档增益.这种影响主要体现在:如果 S_d 中的文档对查询子意图有所满足,则当前文档的文档增益会受到衰减.

结合第 2.1 节中对衰减函数的定义,STA-Measure 定义文档增益 STA-G 为

$$STA\text{-}G_i(d|S_d)=G_i(d)\times f_k(s_i,S_d).$$

注意到,衰减函数$f_k(s_i,S_d)$具有抽象的表示,因此,框架中的 STA-G 实际上只是定义了一种利用查询子意图类型信息来计算文档增益的模型所具有的结构,其具体形式依赖于评测时所采用的查询子意图的分类方法以及这种方法对应的衰减函数.

2.3 多样化评测框架

在对 STA-G 进行一般化的定义之后,我们可以将现有多样化检索评测方法中的文档增益 $G_i(d)$替换为 STA-G,并按照图 1 的步骤得到新的评测方法.新的方法在原始方法的基础上利用了查询子意图的类型信息来对多样化检索进行评测.

例如,我们对 D#-nDCG 中的 GG 进行扩展,得到:

$$STA\text{-}GG_i(d\mid S_d)=\sum_{s_i\in C}P(s_i\mid q)\times G_i(d)\times f_k(s_i,S_d) \tag{11}$$

然后,将公式(11)替换 D#-nDCG 中的 GG,即可得到 STA-Measure 框架下的评测方法 STA-D#-nDCG.

2.4 STA-Measure与现有评测方法的关系

很明显,如果$f_k(s_i,S_d)$退化为常数 1,则 STA-G 就会退化为对应的原始文档增益;而当$f_k(s_i,S_d)$变为如下形式时,则 STA-G 退化为公式(10):

$$f_k(s_i,S_d)=\begin{cases}1, & \text{当}k\text{为导航类},N_i=0\\ 0, & \text{当}k\text{为信息类,或者}k\text{为导航类且}N_i>0\end{cases}.$$

因此,现有的评测方法都可以看作是在 STA-Measure 框架下,其衰减函数为某种函数的一种特殊形式.

3 STA-Measure 的应用

STA-Measure 作为一个框架性方法而提出,定义了根据查询子意图的类型信息进行多样化检索结果评测的方法应当具有的结构.为了在后续实验中验证其有效性,本节我们采用与文献[17]一致的信息类与导航类对查询子意图进行分类.针对这样的分类,我们将提出新的衰减函数并进行讨论.

3.1 衰减函数的讨论

关于衰减函数,在信息检索评测方法的相关文献中已经有了非常多的研究.如:

- Jarvelin 等人在文献[11]中对 nDCG 进行研究时提出的按照文档所在位置进行衰减的函数:

$$f(r)=1/\log_2(r+1),$$

Clarke 等人亦在文献[12]中对其进行了讨论;

- Chapelle 等人[14]提出了关于文档所在位置的线性倒数衰减函数:$f(r)=1/r$;
- Clarke 等人后来又研究了指数衰减函数[19]:$f(r)=\beta^{r-1}(0\leqslant\beta\leqslant 1)$.

不同形式的衰减函数具有不同的衰减性质,比如对数衰减函数,随着文档位置 r 的增加,其衰减的变化程度比线性倒数函数要缓慢;且 r 越大,这种放缓的差距越大.而指数衰减函数具有 $f(r+1)/f(r)=\beta$为常数的特性.

由于我们在本文中研究的是随着已评测的相关文档的增加,用户需求得到满足的程度与后续文档增益之间的衰减关系,因此衰减函数的自变量不再是文档所在位置,而是文档集合 S_d.

为了合理量化这样的影响,我们采用与 I-rec 类似的方法,对 S_d集合中的相关文档进行计数,并用该计数作为衰减函数的变量对文档增益进行衰减.我们定义$|S_d|_i$为 S_d集合中与查询子意图 s_i 相关的文档数量,则将上述关于文档位置的衰减函数中的 r 替换为$|S_d|_i$ 后变为

$$\left.\begin{aligned} f_{\log}(s_i,S_d)&=1/\log_2(|S_d|_i+2)\\ f_r(s_i,S_d)&=1/(|S_d|_i+1)\\ f_\beta(s_i,S_d)&=\beta^{|S_d|_i}\end{aligned}\right\}\tag{12}$$

图 2 中标记为 log,r 和β的曲线分别表示 $f_{\log}$,f_r 和 f_β函数随着相关文档数量的增加,其取值的变化趋势.从图中我们可以看到:

- $f_{\log}$,f_r 和 f_β函数随着相关文档数量的增加,其衰减的程度依次递增;
- 当相关文档数量达到一定值后,衰减的趋势趋于平滑.

考虑到这样的平滑特性以及上述对各类型查询子意图的讨论,本文在实验中将分别用 $f_{\log}$,f_r 和 f_β函数作为信息类查询子意图的衰减函数,并取其中的参数β=0.5,表示一个被标注为相关的文档有 50%概率为误标注.由引言的讨论,对于导航类的查询子意图,仅需用户请求的目标网页即可满足,因此,导航类查询意图需要一个比上述衰减更加剧烈的函数.为此,我们采用如下形式的函数:

$$f_a(s_i,S_d)=\begin{cases}(c-|S_d|_i)/c, & 0\leqslant|S_d|_i\leqslant c\\ 0, & \text{else}\end{cases},$$

其中,c 表示当有 c 个文档与导航类子意图相关时,即可满足用户的导航类信息需求,而后续的文档的增益直接被衰减至 0.在实验中,我们取 c=2.这是因为如果假设被标注为相关的文档有 50%的概率被误标注,则 c=2 时,在平均情况下有 1 个文档能够满足用户导航类的信息需求.从图 2 中我们可以看到:$f_{\log}$,f_r,f_β和 f_a 曲线的衰减程度呈依次递增的趋势;且 f_a 在相关文档数大于等于 2 时,函数的取值为 0.因此,f_a 适用于描述导航类子意图的衰减.

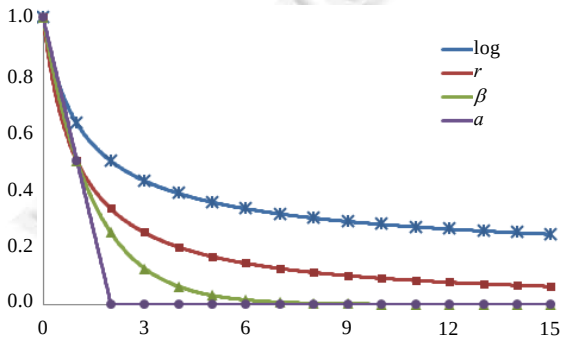

Fig.2 Decay functions of $f_{\log}$, f_r, f_β and f_a

图 2 衰减函数 $f_{\log}$,f_r,f_β和 f_a 曲线示例,分别对应图中的 log,r,β和 a 曲线

4 实 验

4.1 多样化检索评测方法的评测

我们需要对比不同多样化检索评测方法的好坏,这其实涉及到如何对评测方法本身进行评测.Sakai 等人提出了 Discriminative Power[20]来比较评测方法的区分能力,其原理是:用待对比的评测方法分别对多个已知不同的检索系统输出进行评测,并将评测结果两两组对;然后,通过 Bootstrap 的方法部分替换各个系统的评测结果;最终统计待评测方法能够显著区分不同系统输出结果对的比例,如果比例越高,则表明该评测方法越具有区分能力.但 Discriminative Power 有个严重的缺陷,即:它只能衡量评测方法所具有的对不同检索系统的区分能力,并不能给出这样的区分能力是有利还是有弊的.因此,本文中还采用了 Sakai 提出的 Intuitiveness Test[17]方法对评测方法进行评价.

为了更合理地评价多样化评测方法,Sakai 在文献[17]中提出,可以首先采用一些专注于评价我们所关注的某一方面特性(比如检索结果的相关性、多样性等)的评价方法来评测检索结果.这样的方法需要足够简单,被称为 Gold Standard Metric,即,黄金标准.并用待对比的评测方法分别对相同检索结果进行评测,然后比较待对比的方法与黄金标准的评测结果之间的一致性;一致性越好的评测方法,表明其越擅长于该黄金标准所评测的特性.例如文献[17]中,利用 I-rec 作为衡量检索结果多样性的黄金标准,而利用 Ef-P 作为评价检索结果相关性的黄金标准.Ef-P 为 P 评测方法的扩展:其在评测导航类查询子意图时,只考虑第 1 个相关文档;而在评测信息类查询子意图时,则按 P 的定义进行.之所以选择对 P 这种没有根据位置对文档增益进行衰减的方法进行扩展,是因为作为黄金标准,除了应当尽可能的简单,还需要能有效地衡量检索结果某一方面的特性(P 方法可评价检索结果的相关性).因此在本文中,我们采用 I-rec 作为评测检索结果多样化特性的黄金标准,而用 Ef-P 作为评测相关性特性的黄金标准.Intuitiveness Test 的计算过程见算法 1.

算法 1. Intuitiveness Test 算法.

$Disagreements=Correct_1=Corrent_2=0$

foreach pair of search system outputs (r_1,r_2) {

  foreach query q {

    $\Delta M_1=M_1(q,r_1)-M_1(t,r_2)$

    $\Delta M_2=M_2(q,r_1)-M_2(t,r_2)$

    $\Delta GM=GM(t,r_1)-GM(t,r_2)$

    If $(\Delta M_1\times\Delta M_2<0)$ { //M_1 给出对(r_1,r_2)之间大小的判断与 M_2 不同.

      $Disagreements$++

      If $(\Delta M_1\times\Delta GM\geqslant 0)$ //M_1 与 GM 的判断一致.

        $Correct_1$++

      If $(\Delta M_2\times\Delta GM\geqslant 0)$ //M_2 与 GM 的判断一致.

        $Corrent_2$++

      }

}

$I(M_1|M_2,GM)=Correct_1/Disagreements$

$I(M_2|M_1,GM)=Corrent_2/Disagreements$

其中,GM 为选择的黄金标准,M_1,M_2 为待比较的两个评测方法.如果 $I(M_1|M_2,GM)>I(M_2|M_1,GM)$,则表示 M_1 比 M_2 更接近于黄金标准,即,M_1 更擅长于评测黄金标准所评测的特性.

4.2 数据集

本实验采用在参与 TREC 10 Diversity Task[21]评测时,各参赛队所提交的检索结果作为数据集.TREC 10 的 Diversity Task 共向参赛者发布 50 个英文查询词,每个参赛队需要在 ClueWeb 09 所包含的 10 亿英文网页中针

对每个查询词返回包含 1 000 个网页的多样化检索结果.我们将每个参赛队提交的对这 50 个英文查询词的检索结果作为一组结果,这样共有 32 组不同结果.在本文的实验中,我们从这 32 组结果中随机抽取 20 组结果并将其两两组成一对,作为 Discriminative Power 和 Intuitiveness Test 中已知的不同系统输出对,因此,本实验的数据集一共包含 20×19/2=190 个不同系统输出对.

另一方面,为了进一步说明 *STA-Measure* 在不同数据集上的有效性,我们还采用 NTCIR 9 的 Document Ranking 任务[22]中提交的结果做了同样的实验.这是因为:NTCIR 9 的 Document Ranking 任务向参赛者发布 100 个中文查询词,并要求参赛者在 SogouT 所包含的超过 1 亿中文网页中,针对每个查询词返回多样化检索结果.这不同于 TREC 10 Diversity Task 所使用的英文查询词与英文网页,因此,由 NTCIR 9 Document Ranking 的参赛者提交的检索结果所组成的实验数据集与前述由 TREC 10 所组成的数据集有显著不同.同样,我们从 NTCIR 9 的 Document Ranking 任务提交的 24 组检索结果中随机选择 20 组,并两两组成 190 个不同系统输出对作为实验的数据集.

4.3 实验结果

由于在文献[15,17]中,作者已经用实验证明:在现有的多样化检索结果评测方法中,D#-*n*DCG,DIN#-*n*DCG 在 Discriminative Power 和 Intuitiveness Test 的评测中比其他方法更优异,因此在本文的实验中,我们采用这两种方法作为对比的评测方法.考虑到 DIN#-*n*DCG 作为 D#-*n*DCG 的一个扩展(见第 2 节),简单地利用了示性函数来描述不同类型查询子意图的特性,因此,我们也对 D#-*n*DCG 进行 STA-Measure 框架下的扩展.如第 3 节所述,我们用 f_a 作为导航类的衰减函数,而分别用 $f_{\log}$,f_r 和 f_β函数作为信息类的衰减函数.这样不仅可以和 D#-*n*DCG 进行对比说明我们所提出的衰减函数的有效性,而且还可以和使用示性函数作为衰减函数的 DIN#-*n*DCG 进行对比.扩展后的方法标记为 STA-D#-*n*DCG.实验结果见表 1,括号里的 log,β,r 表示扩展时信息类子意图分别采用对应的 $f_{\log}$,f_r 和 f_β衰减函数(导航类子意图都是 f_a).

Table 1 Experimental results in discriminative power

表 1 Discriminative Power 的实验结果

评测方法	TREC 10 数据集		NTCIR 9 数据集	
	Cutoff 10	Cutoff 20	Cutoff 10	Cutoff 20
STA-D#-nDCG(log)	**68.42**	**69.47**	**72.11**	**74.21**
STA-D#-nDCG(β)	68.42	68.95	71.05	74.21
STA-D#-nDCG(r)	68.42	69.47	71.05	74.21
D#-nDCG	67.37	68.42	68.95	73.16
DIN#-nDCG	66.84	66.32	70.00	72.63

从表 1 我们可以看到:通过 STA-Measure 框架扩展之后的 STA-D#-*n*DCG,无论采用 $f_{\log}$,f_r 或 f_β三者中的任何一种衰减函数,且无论在 TREC 10 数据集还是在 NTCIR 9 数据集上,其 Discriminative Power 的值都高于 D#-*n*DCG 和 DIN#-*n*DCG.这表明:扩展后的方法对于不同系统输出的区分度高于待对比的两种方法,且在不同数据集上具有稳定性.考虑到其实 D#-*n*DCG 和 DIN#-*n*DCG 是 STA-Measure 框架的特殊形式,即,分别采用了常数和示性函数作为衰减函数,结合上述实验结果,可说明本文提出的衰减函数有更好的结果.注意到,当信息类子意图采用 $f_{\log}$ 衰减函数时具有最佳区分度,因此在 Intuitiveness Test 实验中,我们仅按照以 $f_{\log}$ 作为信息类子意图的衰减函数进行实验.

从表 2 中我们可以看到:在 TREC 09 数据集和 NTCIR 9 数据集上,无论采用 I-rec 作为检索结果多样性评价的黄金标准,还是采用 Ef-P 作为检索结果相关性评价的黄金标准,STA-D#-*n*DCG 都比对比的两种评测方法表现得更加与黄金标准一致.为了综合考虑待对比方法的多样性评测和相关性评测能力,我们按照文献[17]中的方法,将 I-rec 与 Ef-P 按照等权重线性加权得到被称为 *Both* 的黄金标准,并以此进行 Intuitiveness Test 实验.从表 2 中我们可以看出:当同时考虑 I-rec 和 Ef-P 时,STA-D#-*n*DCG 都具有最佳的结果.这同样说明,第 3 节中我们所提出的衰减函数在应用于 STA-measure 框架进行多样化检索结果评测时,要优于分别以常数和示性函数作为衰减函数的评测方法 D#-*n*DCG 与 DIN#-*n*DCG.

Table 2 Experimental results in Intuitiveness Test

表 2 Intuitiveness Test 的实验结果

TREC 10 数据集					
P(*STA-D#-nDCG*| *D#-nDCG,I-rec*)	*P*(*D#-nDCG*| *STA-D#-nDCG,I-rec*)	*P*(*STA-D#-nDCG*| *D#-nDCG,Ef-P*)	*P*(*D#-nDCG*| *STA-D#-nDCG,Ef-P*)	*P*(*STA-D#-nDCG*| *D#-nDCG,Both*)	*P*(*D#-nDCG*| *STA-D#-nDCG,Both*)
1	0.29	0.57	0.43	0.57	0.14
P(*STA-D#-nDCG*| *DIN#-nDCG,I-rec*)	*P*(*DIN#-nDCG*| *STA-D#-nDCG,I-rec*)	*P*(*STA-D#-nDCG*| *DIN#-nDCG,Ef-P*)	*P*(*DIN#-nDCG*| *STA-D#-nDCG,Ef-P*)	*P*(*STA-D#-nDCG*| *DIN#-nDCG,Both*)	*P*(*DIN#-nDCG*| *STA-D#-nDCG,Both*)
0.875	0.5	0.5	0.5	0.5	0.25
P(*DIN#-nDCG*| *D#-nDCG,I-rec*)	*P*(*D#-nDCG*| *DIN#-nDCG,I-rec*)	*P*(*DIN#-nDCG*| *D#-nDCG,Ef-P*)	*P*(*D#-nDCG*| *DIN#-nDCG,Ef-P*)	*P*(*DIN#-nDCG*| *D#-nDCG,Both*)	*P*(*D#-nDCG*| *DIN#-nDCG,Both*)
1	0.33	0.67	0.33	0.67	0.33
NTCIR 9 数据集					
P(*STA-D#-nDCG*| *D#-nDCG,I-rec*)	*P*(*D#-nDCG*| *STA-D#-nDCG,I-rec*)	*P*(*STA-D#-nDCG*| *D#-nDCG,Ef-P*)	*P*(*D#-nDCG*| *STA-D#-nDCG,Ef-P*)	*P*(*STA-D#-nDCG*| *D#-nDCG,Both*)	*P*(*D#-nDCG*| *STA-D#-nDCG,Both*)
0.848	0.608	0.846	0.329	0.483	0.238
P(*STA-D#-nDCG*| *DIN#-nDCG,I-rec*)	*P*(*DIN#-nDCG*| *STA-D#-nDCG,I-rec*)	*P*(*STA-D#-nDCG*| *DIN#-nDCG,Ef-P*)	*P*(*DIN#-nDCG*| *STA-D#-nDCG,Ef-P*)	*P*(*STA-D#-nDCG*| *DIN#-nDCG,Both*)	*P*(*DIN#-nDCG*| *STA-D#-nDCG,Both*)
0.807	0.635	0.849	0.316	0.501	0.202
P(*DIN#-nDCG*| *D#-nDCG,I-rec*)	*P*(*D#-nDCG*| *DIN#-nDCG,I-rec*)	*P*(*DIN#-nDCG*| *D#-nDCG,Ef-P*)	*P*(*D#-nDCG*| *DIN#-nDCG,Ef-P*)	*P*(*DIN#-nDCG*| *D#-nDCG,Both*)	*P*(*D#-nDCG*| *DIN#-nDCG,Both*)
0.898	0.555	0.648	0.631	0.568	0.347

5 结 论

本文首先讨论了现有的多样化检索结果评测方法,并介绍了这些方法在评测中并未考虑到不同类型查询子意图对信息需求的不同.为了解决这样的问题,本文引入了衰减函数,用以描述不同类型查询子意图的信息需求被满足的程度随着相关文档集合的增加而变化的趋势.通过抽象定义衰减函数,实现了对查询子意图分类方法的抽象,进而使得本文提出的方法成为一个多样化检索评测框架.该框架从抽象的层次定义了根据查询子意图的类型进行多样化检索结果测评的方法具有的结构,并指出当前的多样化检索评测方法都可以看做是这个框架下的一种特例.最后,我们按照信息类和导航类的分类方法,在此框架下讨论了其衰减函数的形式,并通过 Discriminative Power 和 Intuitiveness Test 实验在两个不同数据集上验证了本文提出的衰减函数所构成的多样化检索评测方法优于现有的方法.

在未来的工作中,我们将尝试提出针对查询子意图进行分类的新方法,并讨论在新的分类方法下其衰减函数应当具有的形式,以及验证按照这样的衰减函数组成的新评测方法的效果.

References:

[1] Yu HJ, Liu YQ, Zhang M, Ru LY, Ma SP. Research in search engine user behavior based on log analysis. Journal of Chinese Information Processing, 2007,21(1). [doi: 10.3969/j.issn.1003-0077.2007.01.018]

[2] Dou Z, Hu S, Chen K, Song R, Wen JR. Multi-Dimensional search result diversification. In: Proc. of the ACM WSDM 2011. Hong Kong: ACM Press, 2011. 475−484. [doi: 10.1145/1935826.1935897]

[3] Rafiei D, Bharat K, Shukla A. Diversifying Web search results. In: Proc. of the ACM WWW 2010. Raleigh: ACM Press, 2010. 781−790. [doi: 10.1145/1772690.1772770]

[4] Santos RL, Macdonald C, Ounis I. Selectively diversifying Web search results. In: Proc. of the ACM CIKM 2010. Toronto: ACM Press, 2010. 1179−1188. [doi: 10.1145/1871437.1871586]

[5] Santos RL, Macdonald C, Ounis I. Intent-Aware search result diversification. In: Proc. of the ACM SIGIR 2011. Beijing: ACM Press, 2011. 595−604. [doi: 10.1145/2009916.2009997]

[6] Clarke CL, Craswell N, Soboroff I. Overview of the TREC 2009 Web track. In: Proc. of the TREC 2009. 2010.

[7] Clarke CL, Craswell N, Soboroff I, Voorhees EM. Overview of the TREC 2011 Web track. In: Proc. of the TREC 2011. 2012.

[8] Sakai T, Craswell N, Song RH, Robertson S, Dou Z, Lin CY. Simple evaluation metrics for diversified search results. In: Proc. of the 3rd Int'l Workshop on Evaluating Information Access. Tokyo, 2010. 42−50.

[9] Andrew T, Falk S. User performance versus precision measures for simple search tasks. In: Proc. of the 29th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. 2006. [doi: 10.1145/1148170.1148176]

[10] Voorhees EM. Overview of the 9th text retrieval conference. In: Proc. of the 8th Text Retrieval Conf. TREC-8 Question Answering Track Report. 1999. 77–82.

[11] Jarvelin K, Kekalainen J. Cumulated gain-based evaluation of IR techniques. ACM Trans. on Information Systems (TOIS), 2002, 20(4):422–446. [doi: 10.1145/582415.582418]

[12] Clarke CL, Kolla M, Cormack GV, Vechtomova O, Ashkan A, Buttcher S, MacKinnon I. Novelty and diversity in information retrieval evaluation. In: Proc. of the ACM SIGIR 2008. Singapore: ACM Press, 2008. 659–666. [doi: 10.1145/1390334.1390446]

[13] Agrawal R, Gollapudi S, Halverson A, Ieong S. Diversifying search results. In: Proc. of the 2nd ACM Int'l Conf. on Web Search and Data Mining. ACM, Barcelona, 2009. 5–14. [doi: 10.1145/1498759.1498766]

[14] Chapelle O, Metzler D, Zhang Y, Grinspan P. Expected reciprocal rank for graded relevance. In: Proc. of the ACM CIKM 2009. Hong Kong: ACM Press, 2009. 621–630. [doi: 10.1145/1645953.1646033]

[15] Sakai T, Song RH. Evaluating diversified search results using per-intent graded relevance. In: Proc. of the ACM SIGIR 2011. Beijing: ACM Press, 2010. 1043–1052. [doi: 10.1145/2009916.2010055]

[16] Broder A. A taxonomy of Web search. SIGIR Forum, 2002,36(2):3–10. [doi: 10.1145/792550.792552]

[17] Sakai T. Evaluation with informational and navigational intents. In: Proc. of the ACM WWW 2012. Lyon: ACM Press, 2012. 499–508. [doi: 10.1145/2187836.2187904]

[18] Rose DE, Levinson D. Understanding user goals in Web search. In: Proc. of the ACM WWW 2004. Manhattan: ACM Press, 2004. 13–19. [doi: 10.1145/988672.988675]

[19] Clarke CL, Kolla M, Vechtomova O. An effectiveness measure for ambiguous and underspecified queries. In: Proc. of the 2nd Int'l Conf. on the Theory of Information Retrieval. Cambridge, 2009. 188–199. [doi: 10.1007/978-3-642-04417-5_17]

[20] Sakai T. Evaluating evaluation metrics based on the bootstrap. In: Proc. of the ACM SIGIR 2006. Seattle: ACM Press, 2006. 525–532. [doi: 10.1145/1148170.1148261]

[21] Clarke CL, Craswell N, Soboroff I, Cormack GV. Overview of the TREC 2010 Web track. In: Proc. of the TREC 2010. 2011.

[22] Song RH, Zhang M, Sakai T, Kato MP, Liu YQ, Sugimoto M, Orii N. Overview of the NTCIR-9 INTENT task. In: Proc. of the NTCIR-9. 2011.

陈飞(1987－),男,重庆人,博士,主要研究领域为信息检索.

张敏(1977－),女,博士,副教授,主要研究领域为信息检索.

刘奕群(1981－),男,博士,副教授,主要研究领域为信息检索.

马少平(1961－),男,博士,教授,博士生导师,主要研究领域为知识工程,信息检索,汉字识别后处理,中文古籍数字化.