查询歧义性程度自动标注指标的替代性验证研究*

桂思思[1,2]　张晓娟[3]　王　鑫[1,2]
[1](武汉大学信息管理学院　武汉 430072)
[2](武汉大学信息检索与知识挖掘研究所　武汉 430072)
[3](西南大学计算机与信息科学学院　重庆 400715)

摘要:【目的】针对查询歧义性程度的标注问题, 通过分析自动标注指标间的相关性及自动标注指标与人工标注指标的一致性, 以期获得在一定程度上能替代其他自动标注指标和人工标注的自动标注指标。**【方法】**分别选取基于文档、用户以及查询词项特征的自动标注指标, 依据查询词项对应类目的频率改进一种基于查询词项特征的自动标注指标; 利用皮尔逊相关系数与对称 AP 相关系数分析自动标注结果之间的相关性, 利用宏平均 F1 与宏平均准确率分析自动标注指标与人工标注结果的一致性。**【结果】**自动标注指标之间相关性较弱; 本文改进的自动标注指标与人工标注指标之间一致性最高: 宏平均 F1 值与宏平均准确率分别为 0.623 与 0.707。**【局限】**限于目录型网站的查询词项覆盖率, 部分自动标注指标无法用于所有歧义性查询, 导致用于检验替代性的歧义查询数量较少。**【结论】**自动标注指标之间的替代性较弱; 查询词项对应类目的频率能提高基于查询词项特征的自动标注指标间一致性; 与已有自动标注指标相比, 本文改进的自动标注指标与人工标注结果一致性最高, 在一定程度上可替代人工标注。

关键词: 查询歧义性程度　自动标注　人工标注　替代性　相关性　一致性
分类号: G354
DOI: 10.11925/infotech.2096-3467.2018.0449

1　引　言

查询是用户需求的文字表达, 常由多个关键词组成。然而, 搜索引擎允许用户输入的查询关键词个数有限制, 使得用户提交的查询简短, 从而容易产生歧义。以查询“mustang (野马)”与查询“travel(旅行)”为例, 前者可指代多个不同的事物(如车或动物), 后者虽指代一个明确的事物, 但仍可指代关于该明确事物的多个方面(如旅游注意事项、旅游景点推荐或者旅行社等)。为提高检索系统处理歧义性查询的性能, 部分研究从用户角度入手, 分析用户查询意图中的歧义属性[1-3]以及查询的歧义性对用户使用检索系统的影响[4-5]等; 部分研究从技术角度入手, 利用查询自动分类识别具有歧义性的查询[3,6-12], 利用信息挖掘技术进行查询消歧[13-14], 结合多样化检索技术优化歧义性查询的检索结果列表[15-18]等。其中, 识别查询歧义性程度是上述研究展开的基础。

当前查询歧义性程度识别研究大多转化为查询自动分类问题, 即基于机器学习思想, 根据查询的歧义性程度, 构建查询分类体系(例如 Baeza-Yates 等[6]的分类体系: 信息类查询、非信息类查询与模糊类查询; 或 Song 等[9-10]的分类体系: 歧义查询、宽泛查询、明确查询), 在标注数据集上选取分类特征, 利用分类器训练分类模型, 以此实现歧义性查询的自动识别。其中,

通讯作者: 桂思思, ORCID: 0000-0001-7562-7447, E-mail: sgui0229@whu.edu.cn。
*本文系国家社会科学基金青年项目“融合用户个性化与实时性意图的查询推荐模型研究”(项目编号: 15CTQ019)的研究成果之一。

如何对查询歧义性进行标注是歧义性查询识别的重要工作。已有查询歧义性程度标注方法主要分为人工标注法[2,5,6,9-10]与自动标注法[19-23]。人工标注法的主要思想是请标注者, 依据某分类体系及预先制定的标注规则对数据进行标注。该方法的标注结果具有一定准确性, 但只能对小数据集进行标注, 且在标注规则制定不完备的情况下, 人工标注一致性检验不理想。自动标注法指在不需要人工参与的情况下, 根据查询词项特征[20], 查询所对应检索文档特征[21]或者用户特征[22-23]等, 提出一个自动标注指标计算某查询的具体歧义性数值, 以此完成查询歧义程度的标注。相对人工标注法, 自动标注法虽减少了人力, 可节约标注成本且适用于对大量数据的标注, 但是如何验证自动标注与人工标注结果的一致性还需进一步探讨; 另因部分特征较难获取, 使得某些自动标注指标难以实现, 故一种理想方法是利用简单、易于实现的指标完成查询歧义性自动标注, 但这种方法是否可行取决于自动指标之间是否能相互替代。

基于此, 本文从已有研究中分别选取基于文档特征、基于用户特征和基于查询词项特征三类自动标注指标, 利用查询词项对应类目的频率改进基于查询词项特征的自动指标, 最后针对上述自动标注指标的替代性进行分析, 包括: 分析自动标注指标之间相关性来验证自动指标之间的替代性, 分析自动标注结果与人工标注结果间的一致性来验证自动指标与人工标注结果的替代性。

2 相关研究

2.1 查询歧义性程度的人工标注法

依据分类体系定义的标注规则是人工标注法实施的依据, 因标注规则不同, 标注过程存在差异。根据查询歧义性程度标注中所遵循的不同分类体系, 人工标注法的研究包括以下三类。

(1) 基于查询含义(Meaning)定义的分类体系。Aurelio 等[5]将查询歧义性的程度定义为三类: 查询只有一种含义; 查询包含两种含义; 查询包含三种含义。然而他们没有给出明确的数据、标注过程以及标注结果描述。

(2) 基于用户目标定义的分类体系。Baeza-Yates 等[6]将查询歧义性的程度定义为三类:

①信息类(Informational): 获取与查询相关的信息资源;

②非信息类(Not Informational): 获取与查询相关的其他资源, 或获取特定的网络交互(如购物、下载、保存等);

③模糊(Ambiguous): 无法推测用户目标。该分类体系被多位学者采用, 用于人工标注 TodoCL 搜索引擎的 6 042 个样本查询歧义性程度的依据, 例如 Calderón-Benavides 等[1]、González-Caro 等[3]、Baeza-Yates 等[6]、Mendoza 等[7]。

(3) 基于查询含义及查询子主题共同定义的分类体系。Nguyen 等[2]与 Song 等[9-10]均从查询含义及查询包含子主题两个方面共同定义了三类查询歧义性的程度, 只是使用的类别标签不同。Nguyen 等[2]使用“一词多义(Polysemous)”、“宽泛(General)”、“专指(Specific)”三个标签; Song 等[9-10]使用“歧义(Ambiguous)”、“宽泛(Broad)”、“明确(Clear)”三个标签, 具体含义如下:

①一词多义/歧义: 查询有多个含义;

②宽泛: 查询有一个明确的含义, 却有多个子主题, 用户通过提交新的查询词获取与子主题相关的信息;

③专指/明确: 查询有一个明确含义、且该含义对应一个范围窄的主题。

Nguyen 等[2]根据所提分类体系, 聘请 5 名标注者标注了 AllTheWeb 查询日志中 75 个样本查询的歧义程度, ANOVA 检验表明, 不同标注者的标注结果间无显著差异(F value=0.4297, p=0.7871)。Song 等[9-10]的分类体系被多位学者采用, 如 Yano 等[22]。Song 等[9-11]雇请 5 名标注者分别从词典、检索结果以及用户行为(点击日志数据)三个角度对 MSN 查询日志中 60 条样本查询标注查询歧义的程度, 发现根据检索结果及用户行为, 被至少 4 名标注者标注相同结果的查询占 90%。随后, 他们雇请两名标注者根据用户行为标注对 MSN 查询日志中 400 条样本查询, 发现 63%查询的标注结果相同。Yano 等[22]雇请三名网站员工对网站的 600 条样本查询标注歧义程度, 研究表明, 2/3 标注者对于 582 条查询的标注结果相同(Krippendorff's alpha=0.42)。

总体来说, 一方面, 因人力、时间成本高, 致使参与实验的标注者数量较少(一般不超过 5 人), 因此人工标注法只适于小量数据, 难以用于大量数据; 另一方面, 因个人对标注规则理解的差异, 不同标注者标注结果之间的一致性检验通常不理想。

2.2 查询歧义性程度的自动标注法

自动标注法的本质是提出自动标注指标, 将特征转化为具体的查询歧义性值。根据采用的特征, 自动

标注法研究可归为以下类别。

(1) 基于查询词项特征的自动标注。该类指标根据查询词项之间的相似程度衡量查询歧义性程度, 认为查询词项之间相似度越高, 查询歧义性程度越低[20]。文献[20]还表明, 该指标与检索结果的准确率正相关。尽管如此, 此类指标只考虑任意两个词项之间是否存在相同类目, 忽略了该类目出现次数对词项间相似度的作用。

(2) 基于查询所对应检索文档特征的自动标注。该类指标根据查询所对应检索文档的清晰度度量查询歧义性的程度, 认为文档清晰度越高, 查询歧义性越低[21]。文献[21]也表明该指标与检索的平均准确率正相关。然而此类指标需获取检索文档全文, 计算比较耗时。

(3) 基于用户特征的自动标注。该类指标通过用户点击行为或点击文档内容的差异性衡量查询的歧义性程度[22-23]。Yano 等[22]以查询对应的用户点击文档涵盖主题数目来度量查询歧义性的程度, 认为用户点击文档所涵盖的主题数目越少, 查询歧义性程度越低。实验表明, 自动标注结果与 600 条人工标注结果的相关性不高。Teevan 等[23]总结了两类查询歧义性的度量方式:

①基于用户相关性评价数据的显式度量法: Fleiss Kappa 系数[24]或者潜在个性化曲线(Potential for Personalization Curve)[25];

②基于用户点击数据的隐式度量法: 基于点击数据的潜在个性化曲线和点击熵[26]。

其研究表明, 显式度量法与隐式度量法的度量结果相关, 特别是基于点击数据的潜在个性化曲线度量结果与两个显式度量公式的结果显著相关。该类指标的不足之处在于难以获取真实且实时的用户数据。

相对人工标注来说, 自动标注法对标注人员依赖较小, 可用于大量数据, 从而克服人工标注法的成本消耗问题, 但如何验证自动标注结果还需进一步探讨。多数研究没有直接验证自动标注指标[20-21]; 其他研究也只从某一个方面验证自动标注指标, 例如, Yano 等[22]只考虑自动标注指标与人工标注结果间一致性; Teevan 等[23]只考虑自动指标之间的相关性。但是以上研究均为本文验证自动标注指标的替代性提供了一定启发。

3　查询歧义性程度的自动标注指标

依据自动标注指标度量特征的不同, 本文分别选取三个自动标注指标: 类 I: 基于查询所对应检索文档特征的 $Clarity_{CTC}(q)$指标[21]; 类II: 基于用户特征的 $TopicEntropy(q)$指标[22]; 类 III: 基于查询词项特征的 $Clarity_Q(q)$指标[20]。另外, 本文在 $Clarity_Q(q)$基础上, 提出一个新的查询歧义自动标注指标 $VClarity_Q(q)$。

3.1　已有的自动标注指标

(1) 类 I: 基于检索文档特征的 $Clarity_{CTC}(q)$指标

指标利用查询所返回检索文档的清晰度度量查询歧义性程度。给定一个查询 q, 一个文档集合 D 以及该文档集合中的所有词项集合 V, $Clarity_{CTC}(q)$的定义如公式(1)所示。

$$Clarity_{CTC}(q) = D_{KL}(P(w|\theta_q) \| P(w|\theta_C)) \quad (1)$$

其中, D_{KL} 表示基于查询的语言模型与基于文档集合的语言模型之间的 KL 散度(Kullback-Leibler Divergence), $P(w|\theta_C)$ 表示基于文档集合的语言模型 θ_C 生成 $w \in V$ 的概率, 主要利用最大似然估计(Maximum Likelihood Estimate, MLE)估算词 w 在文档集合 D 中的相对频率; $P(w|\theta_q)$ 为基于查询的语言模型 θ_q 生成词 $w \in V$ 的概率, 由 Lavrenko 等[27]的相关模型(Relevance Model)方式 1 估算: 假设一个查询 q 中包含 n 个查询词项, $q = (w_q^1, \cdots, w_q^i, \cdots, w_q^n), (n \geqslant 1)$, 则 $P(w|\theta_q)$ 计算如公式(2)所示。

$$P(w|\theta_q) = \frac{\sum_{m \in M} P(m)P(w|m)\prod_{i=1}^{n} P(w_q^i|m)}{P(q)} \quad (2)$$

其中, M 表示词项在文档 $d \in D$ 中一元分布 m 的有限集合, $P(m)$ 是一元分布 m 的先验概率, $P(q)$ 是查询 q 的先验概率; $P(w|m)$ 的定义如公式(3)所示。

$$P(w|m) = \lambda P(w|\theta_d) + (1-\lambda)P(w|\theta_C) \quad (3)$$

其中, $P(w|\theta_d)$ 表示词 w 在文档 d 中的相对频率, 也通过 MLE 计算; λ 为调节参数, 用于控制 $P(w|\theta_d)$ 与 $P(w|\theta_C)$ 的权重, 取值为 0 至 1 之间。

(2) 类 II: 基于用户特征的 $TopicEntropy(q)$指标

利用查询对应用户点击文档所涵盖的主题数目度量查询歧义性的程度。给定一个查询 q, 其相关文档集合 R 以及一个主题集合 T, $TopicEntropy(q)$指标的定义如公式(4)所示。

$$TopicEntropy(q) = \frac{1}{|R|}\sum_{d \in R} D_{KL}(P(t|d) \| P(t|q)) \quad (4)$$

其中，$|R|$是相关文档的个数；$P(t|d)$是文档$d \in R$关于主题$t \in T$的分布，可通过主题模型或词袋模型估算；文献[22]表明，估算 $TopicEntropy(q)$中的$P(t|d)$时，使用主题模型 Latent Dirichlet Allocation (LDA)优于词袋模型，故本文使用 LDA 估算$P(t|d)$；$P(t|q)$是查询q关于主题$t \in T$的分布，以R中所有文档的平均主题分布估算。

(3) 类 III: 基于查询词项特征的 $Clarity_Q(q)$指标

利用查询中词项对应类目的交叉程度度量查询歧义性。给定查询q，首先定义W_q为查询q的查询词项集合的子集，使得$\forall w \in W_q$均对应一个非空类目集$C_w = \{c_1, \cdots, c_i, \cdots, c_k\}, (k \geqslant 1)$，$c_i$为一个类目。本文通过目录型网站(Directory Web Site)获取C_w：将查询词项提交至目录型网站进行检索，将返回结果所属类目作为该查询词项对应的类目。在此基础上，$Clarity_Q(q)$指标的定义如公式(5)所示。

$$Clarity_Q(q) = 1/(interScore_q + 1) \tag{5}$$

其中，$interScore_q$表示某查询q中任意两个查询词项所对应相同类目的个数，即公式(6)中集合C_{inter}包含元素的数目。

$$C_{inter} = \{c \,|\, c \in C_{w_i}, c \in C_{w_j}, w_i \in W_q, w_j \in W_q, i \neq j\} \tag{6}$$

3.2 改进的 $VClarity_Q(q)$指标

一般而言，自动标注指标所能依据特征的选择有限，如查询词项特征、检索文档特征、用户特征等，且后两种特征均属于后检索特征，获取、计算均有一定难度，因此，如何设计自动标注指标将有限的特征转化为一个表示查询歧义性的数值，有待进一步研究。其中，公式(5)中定义的$Clarity_Q(q)$只考虑两个词项之间是否存在相同的类目，忽略了该类目频率对不同词项的作用。为解决此问题，本文利用查询词项对应类目的频率重新定义$interScore_q$，将改进的自动标注指标称为$VClarity_Q(q)$，重新定义的$interScore_q$如公式(7)所示。

$$interScore_q = \frac{1}{\#(i,j)} \sum \overrightarrow{w_i} \bullet \overrightarrow{w_j} \tag{7}$$

其中，$\#(i,j)$为查询词项两两无序配对的对数，$\overrightarrow{w_i}$与$\overrightarrow{w_j}$分别为查询词项w_i与w_j的l维向量表示。词项w的向量表示$\vec{w}$如公式(8)所示。

$$\vec{w} = (c_1, Weight_{c_1}; \cdots; c_i, Weight_{c_i}; \cdots; c_l, Weight_{c_l}) \tag{8}$$

其中，l为目录型网站的一级类目个数，c_i是第i个类目，$Weight_{c_i}$为类目c_i对应的权重。根据类目出现次数，$Weight_c$有$CF(c)$、$CFIQF(c)$以及$CFIQFN(c)$三种计算方式，定义如公式(9)至公式(11)所示。

$$CF(c) = \frac{cf_c}{\left\|\vec{w}\right\|} \tag{9}$$

$$CFIQF(c) = cf_c \times log_{10}\frac{N}{n_c} \tag{10}$$

$$CFIQFN(c) = \frac{CFIQF(c)}{\left\|\vec{w}\right\|} \tag{11}$$

其中，$\left\|\vec{w}\right\|$表示向量$\vec{w}$模的长度，cf_c为类目c在C_w的频率；n_c为属于类目c的所有查询词项的数目；N为所有查询词项的数目。

4 实验设计

4.1 数据获取与预处理

本文获取 2009 年至 2012 年 TREC Web Track 中 Adhoc 任务提供的测评数据集①，并在此基础上自主构建 ClueWebRel 数据集，所有数据均使用 Indri 停用词表②去除停用词。实验数据包括以下部分:

(1) 查询任务描述(Topic Full Statement File)数据集。该数据集以 XML 格式记录每个查询任务(topic)的查询条目(query)、查询任务描述(description)以及查询子主题(subtopic)③，具体数据样例如图 1 所示。其中，TREC 主办方标注了每个查询的类别(type): 模糊性(ambiguous)(见图1(a))或者多面性(faceted(见图1(b))。模糊性查询定义为包含多个不同且互不相关含义(Interpretation)的查询，单个用户只对此类查询中包含的某一种含义感兴趣，如查询“windows”，可能的含义为“计算机系统windows”或者“建筑windows”；多面性查询仅包含一个主要含义，但是包含多个关于该含义的多个子主题，单个用户同时对一个或多个子主题感

①http://trec.nist.gov/data/webmain.html.

②http://www.lemurproject.org/stopwords/stoplist.dft.

③查询任务与查询条目一一对应，如无特殊说明，本文将查询任务、查询条目统称为查询。

兴趣[28-31]，如查询“建筑 windows”，可能的子主题为“窗户安装”、“窗户设计”等。TREC 每年发布 50 个不重复的查询，本文共获得 200 个查询，其中包含 58 个模糊性查询，142 个多面性查询。

```
<topic number="58" type="ambiguous">
  <query>penguins</query>
  <description>Find information about penguins.</description>
  <subtopic number="1" type="nav">Find the homepage of the
  Pittsburgh Penguins</subtopic>
  <subtopic number="2" type="nav">
    Find Pittsburgh Penguins merchandise such as hockey
    jerseys.
  </subtopic>
  <subtopic number="3" type="nav">Find information about
  penguins.</subtopic>
  <subtopic number="4" type="inf">Find penguin photos.
  </subtopic>
  <subtopic number="5" type="nav">
    Find pictures of the penguins from the animated movie,
    "Madagascar".
  </subtopic>
</topic>
```

(a) 模糊性查询

```
<topic number="69" type="faceted">
  <query>sewing instructions</query>
  <description>
    Find beginners instructions to sewing, both by hand and by
    machine.
  </description>
  <subtopic number="1" type="nav">Find sewing sites for
  beginners.</subtopic>
  <subtopic number="2" type="inf">Find instructions for using a
  sewing machine.</subtopic>
  <subtopic number="3" type="nav">Find downloadable sewing
  patterns.</subtopic>
  <subtopic number="4" type="inf">Find helpful sewing tips for
  beginners.</subtopic>
  <subtopic number="5" type="inf">Find materials for teaching
  sewing to children.</subtopic>
</topic>
```

(b) 多面性查询

图 1　查询任务描述数据样例

(2) ClueWeb09 Category B 数据集①: Lemur 项目组于 2009 年 1 月至 2009 年 2 月从网络上采集的前 5 亿条英文网页全文数据。

(3) 上述数据集对应的相关性评分标准结果集(qrel)，如图 2 所示。该结果集中每一条记录包含一个任务编号(第 1 列)、一个文档编号(第 3 列)以及该文档与该任务的相关性得分(第 4 列)。相关性得分是 TREC 每年发布的 50 个查询与其在 ClueWeb09 数据集上采用 Pooling 方法获得的检索集合的人工相关性判断。虽然 TREC 每年相关性评分标准的量表不同②，但是评分标准相同：正值表示相关，零值表示无关，负值表示垃圾网页。

(4) ClueWebRel 为 ClueWeb09 Category B 数据集的子集，是笔者根据 qrel 从 ClueWeb09 Category B 中抽取的查询相关文档(相关性得分为正值)集合，如图 2 所示。剔除 6 个无法获得相关文档全文的查询③，ClueWebRel 包含 194 个查询的全部相关文档全文，共计 11 037 篇，其中不重复文档共计 11 022 篇，平均每个查询的相关文档为 56.89 篇。

```
58 0 clueweb09-en0008-01-12108 3
58 0 clueweb09-en0008-05-26937 0
58 0 clueweb09-en0008-07-14334 1
58 0 clueweb09-en0008-07-15284 2
58 0 clueweb09-en0008-07-15589 0
58 0 clueweb09-en0008-07-15591 0
58 0 clueweb09-en0008-07-16526 1
```

图 2　相关性评分标准结果集数据样例

4.2　人工标注方法实现

笔者发现，基于 Nguyen 等[2]及 Song 等[9-10]定义的查询歧义程度分类体系(一词多义/歧义、宽泛、专指/明确)与 TREC 查询歧义性程度标注所采用的分类体系(模糊、多面)之间具有相似性，他们均考虑了查询的含义及查询所包含的子主题。因此，没有单独设计人工标注实验，而直接使用 TREC 标注的查询歧义性程度数据进行相关研究。本文相关研究部分已说明，Nguyen 等[2]和 Song 等[9-10]的差异仅在于两个分类体系中描述歧义程度的标签文字不同，因此，统一使用“完全歧义”、“中度歧义”及“略微歧义”分别指代两个体系中的“一词多义/歧义”、“宽泛”和“专指/明确”，并利用子主题数目这一指标将 TREC 的两类歧义性(模糊、多面)转换为 Nguyen 等[2]和 Song 等[9-10]的三类歧义性(略微歧义、中度歧义、完全歧义)，具体操作如下：TREC 的“模糊”对应“完全歧义”，TREC 的

①https://lemurproject.org/clueweb09.php/.

②2010 年、2012 年的相关性评分标准采用 6 点量表，2011 年采用 5 点量表，2009 年采用 3 点量表。具体可参见文献[28-31]。

③根据 qrel 中，有：编号为 95,100 的查询无相关性文档评分；编号为 20 的查询无相关文档。根据 ClueWeb09 Category B 数据集，有：编号为 112, 143, 152 的查询的全部相关文档缺失。

“多面”根据“多面”查询的子主题个数对应“略微歧义”或者“中度歧义”。

笔者分析发现, 142 个 TREC “多面”查询中, 查询子主题数目的范围为 2-8, 且多数查询的子主题个数为 3 或 4, 分别占 28.17%与 43.66%。因此以 3 作为阈值, 若该“多面”查询的子主题数不超过 3(≤3), 则该“多面”查询对应于“略微歧义”; 否则为“中度歧义”。最终, 本文 200 个查询的查询歧义性程度的统计信息如表 1 所示。

表 1　200 个查询的歧义程度统计表

比较项	全部查询	完全歧义	中度歧义	略微歧义
查询个数	200	58	100	42
占比	100%	29%	50%	21%

4.3　自动标注的实现

(1) 类 I: 指标 $Clarity_{CTC}(q)$的实现

计算 200 个查询的 $Clarity_{CTC}(q)$值, 实现细节与文献[21]相同, 即:

①针对每一个查询, 利用 Indri 5.7 语言模型(Dirichlet 平滑, μ=2500)从 ClueWebRel 检索 500 个文档作为其文档集合 D;

②公式(3)中 λ 设置为 0.6;

③公式(1)中, 只考虑文档集合 D 中的不重复词项。

该实验组简写为“CTC”。

(2) 类 II: 指标 $TopicEntropy(q)$的实现

只对 194 个存在相关文档的查询(参见 4.1 节)计算 $TopicEntropy(q)$值。其中, 该指标中的 LDA 模型训练与主题推演由 MALLET[①]实现, 最佳主题数目由 MLE 确定。根据模型训练所需要的步骤与数据, 本文涉及三个 LDA 实验组。

①LDA1: 仅包含模型训练步骤, 即针对每个查询的相关文档全文训练一个最优主题模型。在模型训练时, 主题数目取值范围为 1-70, 取值间隔为 1。最终, 194 个主题模型的最优主题数取值范围为 1-67, 其中 2 与 19 为最常见的最佳主题数, 分别占 6.2%与 5.2%。

②LDA2: 包含模型训练以及主题推演两个步骤。模型训练时, 仅在 ClueWebRel 上训练一个最优主题模型, 主题数目取值范围为 100-1 000, 取值间隔为 5, 最终确定最优主题数目为 885; 主题推演时, 在每个查询的相关文档上, 均只使用这一个最优主题模型。

③LDA3: 与 LDA2 相似, 但训练数据不同, 训练数据采用 Pooling 方式构建。对于全部 200 个查询, 针对每个查询, 利用 Indri 的三个检索模型在 ClueWeb09 Category B 分别各自检索前 100 个文档, 最终获得 36 591 篇文档。三个检索模型分别为: 语言模型(Dirichlet 平滑, μ=2500)、语言模型(Jelinek-Mercer 平滑, λ=0.4)及 OKAPI BM25(k_1=1.2, b=075, k_3=7)。模型训练时, 主题数目取值范围为 100-3 000, 取值间隔为 100。最终确定最优主题数目为 2 200。

本文将使用不同 LDA 实验组的 $TopicEntropy(q)$实验组分别简写为“TE1”、“TE2”以及“TE3”。

(3) 类 III: 指标 $Clarity_Q(q)$与指标 $VClarity_Q(q)$的实现

本文向目录型网站提交单个查询词项, 获得(至多)前 50 条原始检索结果(网站或网页), 并记录检索结果所属的一级类目。因 Open Directory Project (http://www.dmoz.org/)自 2017 年 3 月 17 起停止服务, 笔者于 2017 年 11 月 14 日选用以下 5 个目录型网站: BWD (Best of the Web, https://botw.org/)、DLIVE(DMOZ, http://dmozlive.com/)、JANT(JoeAnt, http://www.joeant.com/)、HVN(Hot vs Not, http://www.hotvsnot.com/) 、MID (Marketing Internet Directory, http://www. marketing internetdirectory.com)。受目录型网站收录网站、网页数据量的影响, 并非 200 条查询词的所有 400 个不重复查询词项均可获取类目的对应关系, 不同目录型网站上查询词项、查询的覆盖率等统计信息如表 2 所示。最终本文只针对能被任一目录型网站覆盖的 141 个查询计算 $Clarity_Q(q)$值和 $VClarity_Q(q)$值。

表 2　查询词项与类目数目统计表

目录型网站	一级类目数目	覆盖查询词项数目(%)	查询词项对应一级类目平均数目	覆盖查询数目(%)
BWD	16	394 (98.5%)	6.77	140 (70.0%)
DLIVE	15	386 (96.5%)	6.68	137 (68.5%)
JANT	18	335 (83.8%)	8.76	117 (58.5%)
HVN	16	314 (78.5%)	6.77	111 (55.5%)
MID	13	254 (63.5%)	4.81	91 (45.5%)
M	22	398 (99.5%)	13.30	141 (70.5%)

①http://mallet.cs.umass.edu/.

从表 2 可知, 5 个目录型网站预设的一级类目数目范围为 13-18, 查询词项对应一级类目的平均数目范围为 4.81-8.76; 5 个目录型网站涵盖的查询词比例范围为 63.5%-98.5%, 涵盖的查询比例范围仅为 45.5%-70.0%。因此, 计算 $Clarity_Q(q)$或 $VClarity_Q(q)$时, 本文使用平均式(A)或融合式(M)两种方法综合从 5 个目录型网站获取的查询词项与类目之间的对应关系。其中, 平均式综合基于不同目录型网站的查询歧义值, 具体实现过程如下:

①分别依据不同目录型网站获取的对应关系计算 $Clarity_Q(q)$或 $VClarity_Q(q)$;

②利用最大最小归一化法, 将基于单个目录型网站计算的 $Clarity_Q(q)$或 $VClarity_Q(q)$数据分别线性化映射至[0,1], 以减小基于不同目录型网站计算的 $Clarity_Q(q)$或 $VClarity_Q(q)$取值范围的差异;

③取归一化后的 $Clarity_Q(q)$或 $VClarity_Q(q)$平均值为最终结果。

融合式综合不同目录型网站的类目体系, 具体实现过程如下:

①针对 5 个目录型网站预设的一级类目, 采用字符串匹配方式, 人工将类目名称含有相同字符子串的类目合并为一个类目(如"Health"与"Health and Fitness"), 最终将 5 个目录型网站预设的一级类目体系合并为一个类目体系;

②分别将从 5 个目录型网站获取的结果映射到合并后的类目体系(类目合并后查询词项与类目对应关系的统计信息见表 2 最后一行);

③计算 $Clarity_Q(q)$或 $VClarity_Q(q)$值。

一共涉及 8 组实验(CLA、CLM、V1A、V1M、V2A、V2M、V3A、V3M)。前两个字符表明自动标注指标: CL 为 $Clarity_Q(q)$; V1、V2、V3 为基于公式(9)–公式(11)计算的 $VClarity_Q(q)$; 最后一个字符表明 5 个目录型网站结果的融合方式(A 或 M)。

5　实验结果分析

三类 6 种自动标注方法实现的 12 组实验如下:

(1) 类 I: CLC;

(2) 类 II: TE1、TE2、TE3;

(3) 类 III: CLA、CLM、V1A、V1M、V2A、V2M、V3A、V3M。

5.1　自动标注指标之间相关性分析

本文将自动标注指标间的对比问题转化为相同长度的列表间相关性检测问题, 每个自动标注实验组的结果均可视为一列得分值, 对于两个列表中包含的相同查询, 使用皮尔逊相关系数(ρ)与对称 AP 相关性系数($symm\tau_{ap}$)[32]测量其相关性。其中, ρ 与 $symm\tau_{ap}$ 的取值区间均为[-1,1]; 其系数值的正负表明正相关或者负相关; ρ 或 $symm\tau_{ap}$ 的绝对值越大, 表明相关性越强, 反之, 相关性越弱。以 Cohen[33]提出的相关性强度的对应关系为参考, 具体如表 3 所示。

表 3　相关性强度与相关系数对应表

系数绝对值	0.00-0.09	0.10-0.29	0.30-0.49	0.50-1.0
强度	无	弱	适中	强

查询歧义性程度自动标注指标间 ρ 检验与 $symm\tau_{ap}$ 检验结果分别如表 4 与表 5 所示, 单元格颜色的深与浅表明相关性的强与弱。由表 4 与表 5 数据可得如下结论:

(1) ρ 数值比 $symm\tau_{ap}$ 数值略高;

(2) 在 12 个自动标注实验组中, 以 ρ 而言, TE2 与 TE3 的相关性最强(ρ=0.896); 以 $symm\tau_{ap}$ 而言, V1A 与 V3A 的相关性最强($symm\tau_{ap}$=0.573);

表 4　查询歧义性程度自动标注指标 ρ 检验

	CLC	TE1	TE2	TE3	CLM	CLA	V1M	V1A	V2M	V2A	V3M	V3A
TE1	0.255											
TE2	0.206	0.776										
TE3	0.158	0.766	0.896									
CLM	–0.085	0.090	0.054	0.057								
CLA	–0.237	0.119	0.045	0.099	0.660							
V1M	–0.030	0.069	0.046	0.033	0.317	0.274						
V1A	–0.088	0.092	0.089	0.105	0.414	0.503	0.778					
V2M	0.072	–0.107	–0.099	–0.111	0.683	0.430	0.245	0.345				
V2A	–0.030	–0.047	–0.036	–0.025	0.517	0.607	0.306	0.536	0.678			
V3M	0.042	–0.146	–0.103	–0.097	–0.100	0.023	0.516	0.436	0.256	0.308		
V3A	–0.049	–0.053	0.006	–0.004	0.141	0.264	0.652	0.809	0.304	0.536	0.694	

(3) 多数自动标注实验组间适中正相关, 只有少数类 III 实验组(CL, V1, V2, V3)与类 I 实验组(CLC)、与少数类 II 实验组(TE)之间弱负相关;

(4) 相同自动标注指标的不同实验组之间适中或强正相关: 以 ρ 而言, TE 所有实验组之间相关性最好, 平均为 0.813; 以 $symm\tau_{ap}$ 而言, V1 所有实验组之间相关性最好, 平均为 0.508;

(5) 不同自动标注指标之间相关性差, 只存在弱负相关或者弱正相关, 说明不同自动指标之间的替代性较弱。

表 5 查询歧义性程度自动标注指标 $symm\tau_{ap}$ 检验

	CLC	TE1	TE2	TE3	CLM	CLA	V1M	V1A	V2M	V2A	V3M	V3A
TE1	–0.075											
TE2	–0.042	0.448										
TE3	–0.062	0.427	0.556									
CLM	–0.275	0.050	–0.038	–0.021								
CLA	–0.229	0.052	–0.018	0.004	0.495							
V1M	–0.057	0.050	0.048	0.016	0.105	0.124						
V1A	–0.126	0.022	0.039	0.043	0.172	0.282	0.508					
V2M	0.038	–0.097	–0.096	–0.089	0.248	0.233	0.084	0.137				
V2A	–0.053	–0.046	–0.042	–0.018	0.244	0.397	0.130	0.286	0.389			
V3M	0.030	–0.056	–0.018	–0.025	0.010	0.074	0.325	0.257	0.199	0.190		
V3A	–0.077	–0.014	0.017	0.022	0.103	0.230	0.438	0.573	0.173	0.330	0.413	

5.2 自动标注指标与人工标注之间的一致性分析

本文将查询歧义性自动标注法与人工标注法一致性检验问题转为一个多类别分类问题, 使用宏平均 F1 和宏平均准确率[34]度量, 得分越高表明一致性越好。将人工标注的结果作为正确类别(Gold Label), 而自动标注法则分别根据两个得分阈值 k_1 和 k_2 预测查询歧义性程度类别: 对于 $Clarity_{CTC}(q)$、$Clarity_Q(q)$ 及 $VClarity_Q(q)$指标, 若分值大于等于阈值 $k_2(\geq k_2)$, 则归类为“略微歧义”; 若分值小于等于阈值 $k_1(\leq k_1)$, 则归类为“完全歧义”。对于 $TopicEntropy(q)$指标, 若其分值大于等于阈值 $k_2(\geq k_2)$, 则归类为“完全歧义”; 若分值小于等于阈值 $k_1(\leq k_1)$, 则归类为“略微歧义”。本文以 min, max, median 以及 std 分别表示一系列数据中的最小值、最大值、中位数以及标准差, 并测试 S1、S2 以及 S3 三种阈值选择方法。

①S1: 采用 grid 方法, 选择使得宏平均 F1、宏平均准确率取值最优的 k_1 及 k_2 组合。k_1 取值范围为 min 至 median; k_2 取值范围为 k_1 至 median; 间隔为(max–min)/100。

②S2: k_1 及 k_2 取值分别如公式(12)与公式(13)所示。

$$k_1=\text{argmax}\{\text{min},\text{median}-\text{std}\} \quad (12)$$

$$k_2=\text{argmin}\{\text{max},\text{median}+\text{std}\} \quad (13)$$

③S3: 针对每个实验组的得分, 按照升序对数据进行排序, 取位于 1/3 处的数值为 k_1, 位于 2/3 处的数值为 k_2。

因为 S2 与 S3 中 k_1 及 k_2 的设定依赖于标签均匀分布的数据集, 故笔者在尽可能多涵盖查询的基础之上, 构建一个均匀分布的查询集合, 用以 S1、S2 以及 S3 三种阈值选择方法。该集合共包含 66 个查询, 均可被 12 组实验组计算, 每个歧义程度的查询为 22 个。在该查询集合上, 12 个自动标注实验组与人工标注的检验结果如表 6 所示。

从表 6 可知, 在所有阈值设定下, 若不考虑本文新提出的 $VClarity_Q(q)$, 三类自动标注指标中, 类 II 实验组(TE)效果最好; 若考虑本文提出的 $VClarity_Q(q)$, V1A 效果最好, 特别是使用 S1 时, 宏平均 F1 最高可达 0.623, 宏平均准确率最高可达 0.707。

在所有类 III 自动标注指标的实验组中, 使用平均式(A)结果融合方法的实验组效果优于使用融合式(M), 且 $VClarity_Q(q)$的三种方法效果略优于 $Clarity_Q(q)$, 在三种 $VClarity_Q(q)$中, V1 的宏平均 F1 与宏平均准确率最高。

表 6　查询歧义性程度的自动标注与人工标注一致性检验

类别	方法	S1		S2		S3	
		宏平均 F1	宏平均准确率	宏平均 F1	宏平均准确率	宏平均 F1	宏平均准确率
I	CLC	0.561	0.657	0.404	0.545	0.379	0.586
II	TE1	0.594	0.687	0.446	0.606	0.303	0.535
	TE2	0.543	0.647	0.413	0.596	0.364	0.575
	TE3	0.519	0.677	0.251	0.556	0.425	0.616
III	CLA	0.532	0.637	0.374	0.576	0.318	0.546
	CLM	0.455	0.606	0.421	0.586	0.345	0.566
	V1A	**0.623**	**0.707**	**0.477**	**0.616**	**0.485**	**0.657**
	V1M	0.564	0.697	0.319	0.546	0.424	0.616
	V2A	0.511	0.657	0.303	0.535	0.424	0.616
	V2M	0.394	0.657	0.344	0.586	0.440	0.626
	V3A	0.501	0.667	0.359	0.566	0.455	0.636
	V3M	0.563	0.667	0.350	0.566	0.349	0.566

6　结　语

针对查询歧义性程度标注, 本文从已有研究中分别选取基于不同特征(查询词项、检索文档、用户)的自动标注指标, 利用查询词项对应类目的频率改进了一种基于查询词项特征的自动标注指标。在此基础上, 从两方面对自动标注指标进行替代性检验: 以皮尔逊相关系数与对称 AP 相关系数验证自动标注指标之间的相关性; 以宏平均 F1 与宏平均准确率验证自动标注指标与人工标注结果之间的一致性。研究结果表明: 不同类型的自动标注指标之间相关性弱, 说明自动标注指标之间替代性较弱; 在所有自动标注指标中, 本文提出的自动标注指标 $VClarity_Q(q)$与人工标注的一致性最高(宏平均 F1 为 0.623, 宏平均准确率为 0.707), 说明该方法在一定程度上可替代人工标注。尽管如此, 本文仍存在一些不足, 也是笔者在未来工作中进一步深入探讨的内容:

(1) 计算 $Clarity_Q(q)$指标以及 $VClarity_Q(q)$指标时, 限于目录型网站的查询词项覆盖率, 部分自动标注指标无法用于查询歧义性程度标注, 导致用于检验有效性的歧义查询数量较少, 需要寻找查询词项覆盖率更高的目录型网站或在更大歧义查询数据集上, 进一步验证自动标注指标的有效性;

(2) 现有自动标注方法只使用单一自动标注指标, 可探讨融合不同特征的自动标注法的有效性;

(3) 在自动标注歧义查询的基础上, 探讨如何提高歧义性查询的检索性能。

参考文献:

[1] Calderón-Benavides L, González-Caro C, Baeza-Yates R. Towards a Deeper Understanding of the User's Query Intent[C]// Proceedings of the SIGIR 2010 Workshop on Query Representation and Understanding. 2010: 21-24.

[2] Nguyen B V, Kan M Y. Functional Faceted Web Query Analysis[C]// Proceedings of the 16th International World Wide Web Conference. 2007.

[3] González-Caro C, Baeza-Yates R. A Multi-faceted Approach to Query Intent Classification[C]// Proceedings of the 18th International Conference on String Processing and Information Retrieval. 2011: 368-379.

[4] Clough P, Sanderson M, Abouammoh M, et al. Multiple Approaches to Analysing Query Diversity[C]// Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2009: 734-735.

[5] Aurelio D N, Mourant R R. The Effects of Web Search Engine Query Ambiguity and Results Sorting Method on User Performance and Preference[J]. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2002, 46(12): 1271-1275.

[6] Baeza-Yates R, Calderón-Benavides L, González-Caro C. The Intention Behind Web Queries[C]// Proceedings of the 13th International Conference on String Processing and Information Retrieval. 2006: 98-109.

[7] Mendoza M, Baeza-Yates R. A Web Search Analysis Considering the Intention Behind Queries[C]// Proceedings of the 2008 Latin American Web Conference. 2008: 66-74.

[8] Wang Y, Agichtein E. Query Ambiguity Revisited: Clickthrough Measures for Distinguishing Informational and Ambiguous Queries[C]// Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010: 361-364.

[9] Song R, Luo Z, Wen J R, et al. Identifying Ambiguous Queries in Web Search[C]// Proceedings of the 16th International Conference on World Wide Web. ACM, 2007: 1169-1170.

[10] Song R, Luo Z, Nie J Y, et al. Identification of Ambiguous Queries in Web Search[J]. Information Processing and Management, 2009, 45(2): 216-229.

[11] Song R, Dou Z, Hon H W, et al. Learning Query Ambiguity Models by Using Search Logs[J]. Journal of Computer Science and Technology, 2010, 25(4): 728-738.

[12] Pradhan N, Deolalikar V, Li K. Atypical Queries in eCommerce[C]// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, 2015: 1767-1770.

[13] Lioma C, Blanco R, Moens M. A Logical Inference Approach to Query Expansion with Social Tags[C]// Proceedings of the 2nd ACM International Conference on the Theory of Information Retrieval. 2009: 358-361.

[14] Lioma C, Ounis I. A Syntactically-based Query Reformulation Technique for Information Retrieval[J]. Information Processing and Management, 2008, 44(1): 143-162.

[15] Welch M J, Cho J, Olston C. Search Result Diversity for Informational Queries[C]// Proceedings of the 20th International Conference on World Wide Web. ACM, 2011: 237-246.

[16] Santos R L T, Macdonald C, Ounis I. Intent-aware Search Result Diversification[C]// Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2011: 595-604.

[17] Ashkan A, Clarke C L A. On the Informativeness of Cascade and Intent-aware Effectiveness Measures[C]// Proceedings of the 20th International Conference on World Wide Web. ACM, 2011: 407-416.

[18] Zhou K, Whiting S, Jose J, et al. The Impact of Temporal Intent Variability on Diversity Evaluation[C]// Proceedings of the 35th European Conference on Advances in Information Retrieval. 2013: 820-823.

[19] Stojanovic N. On Analysing Query Ambiguity for Query Refinement: The Librarian Agent Approach[C]// Proceedings of the 22nd International Conference on Conceptual Modeling. 2003: 490-505.

[20] Qiu G, Liu K, Bu J, et al. Quantify Query Ambiguity Using ODP Metadata[C]// Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2007: 697-698.

[21] Cronen-Townsend S, Croft W B. Quantifying Query Ambiguity[C]// Proceedings of the 2nd International Conference on Human Language Technology Research. 2002: 104-109.

[22] Yano Y, Tagami Y, Tajima A. Quantifying Query Ambiguity with Topic Distributions[C]// Proceedings of the 25th ACM Conference on Information and Knowledge Management. 2016: 1877-1880.

[23] Teevan J, Dumais S T, Liebling D J. To Personalize or Not to Personalize: Modeling Queries with Variation in User Intent[C]// Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval. 2008: 163-170.

[24] Fleiss J L. Measuring Nominal Scale Agreement Among Many Raters[J]. Psychological Bulletin, 1971, 76(5): 378-382.

[25] Teevan J, Dumais S T, Horvitz E. Potential for Personalization[J]. ACM Transactions on Computer-Human Interaction, 2010, 17(1): Article No.4.

[26] Dou Z, Song R, Wen J R. A Large-scale Evaluation and Analysis of Personalized Search Strategies[C]// Proceedings of the 16th International Conference on World Wide Web. 2007: 581-590.

[27] Lavrenko V, Croft W B. Relevance Based Language Models[C]// Proceedings of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2001:120-127.

[28] Clarke C L, Craswell N, Soboroff I. Overview of the TREC 2009 Web Track[C]// Proceedings of the Text Retrieval Conference. 2009.

[29] Clarke C L, Craswell N, Soboroff I, et al. Overview of the TREC 2010 Web Track[C]// Proceedings of the Text Retrieval Conference. 2010.

[30] Clarke C L, Craswell N, Soboroff I, et al. Overview of the TREC 2011 Web Track[C]// Proceedings of the Text Retrieval Conference. 2011.

[31] Clarke C L, Craswell N, Voorhees E M. Overview of the TREC 2012 Web Track[C]// Proceedings of the Text Retrieval Conference. 2012.

[32] Yilmaz E, Aslam J A, Robertson S. A New Rank Correlation

Coefficient for Information Retrieval[C]// Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval. 2008: 587-594.

[33] Cohen J. Statistical Power Analysis for the Behavioral Sciences[M]. L. Erlbaum Associates, 1988.

[34] Sokolova M, Lapalme G. A Systematic Analysis of Performance Measures for Classification Tasks[J]. Information Processing and Management, 2009, 45(4): 427-437.

作者贡献声明:

桂思思: 提出研究思路, 设计研究方案, 实现算法, 完成数据分析, 撰写论文;
张晓娟: 确定研究思路, 论文修改及最终版本修订;
王鑫: 抓取 5 个目录型网站数据, 训练 LDA 主题模型。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储，E-mail: sgui0229@whu.edu.cn。
[1] 王鑫, 桂思思. webCrawl.rar. 5 个目录型网站中查询词项所对应的类目数据.
[2] 王鑫, 桂思思. C1.rar. ClueWebRel 数据.
[3] 桂思思. auto-result.rar. 12 个自动标注法实验组 200 个查询的歧义数值.
[4] 桂思思. hum-result.txt. 人工标注 200 个查询的歧义程度.

收稿日期: 2018-04-23
收修改稿日期: 2018-05-14

Automatically Rating Query Ambiguity with Alt-Metrics

Gui Sisi[1,2] Zhang Xiaojuan[3] Wang Xin[1,2]
[1](School of Information Management, Wuhan University, Wuhan 430072, China)
[2](Institute for Information Retrieval and Knowledge Mining, Wuhan University, Wuhan 430072, China)
[3](School of Computer and Information Science, Southwest University, Chongqing 400715, China)

Abstract: [Objective] This paper aims to find better alt-metrics for automatically rating query ambiguity. **[Methods]** First, we chose several existing auto-metrics based on documents, users and queries. Then, we modified one of them with query category occurences. Finally, we examined the relationship between the modified alt-metrics and other automatic or human rating metrics. Their correlations were tested with Pearson and symmetric AP correlation coefficients. Their degrees of agreement were tested with macro average accuracy and macro average F1. **[Results]** The proposed method showed significant relationship with human rating, and achieved F1 of 0.623 and accuracy of 0.707. **[Limitations]** Only examined the proposed model with data from online directories.**[Conclusions]** Automatic rating metrics for query ambuiguity can hardly be replaced by other automatic counterparts. Considering the occurences of top-level categories for each query could improve the degrees of agreement for automatic metrics. Compared to the exisiting automatic metrics, the proposed method can be used to replace the human metrics for query ambiguity.

Keywords: Query Ambiguity Rating Automatic Rating Human Rating Alternativeness Correlation Agreement