



基于子主题选择与三级分层结构的 Web 文本挖掘方法

史玉珍, 单冬红

(平顶山学院软件学院, 河南 平顶山 467000)

摘要: 针对用户和查询之间的意图差距导致的查询模糊宽泛和数据稀疏问题, 根据流行性和多样性返回可能子主题的排名列表, 利用子主题选择与排序的分层结构进行 Web 文本挖掘。首先, 在名词性短语和可替代部分查询的基础上, 使用简单模式提取各种相关的短语作为候选子主题; 然后, 使用网页文档集中的相关文档构建候选子主题的三级层次结构; 最后, 综合考虑流行性和多样性, 利用该结构和估计的流行度进行排序。实验使用了 NTCIR-9 库的 100 个日文查询和来自 TREC 2009 库的 100 个英文查询以及网络跟踪多样性任务, 实验结果验证了本文方法可有效应用于各种搜索, 对于高排名的子主题挖掘优于外部资源。

关键词: 数据稀疏; 文本挖掘; 层次结构; 多样性; 流行性

中图分类号: TP391

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2016142

Web text mining method based on subtopic selection and three-level stratified structure

SHI Yuzhen, SHAN Donghong

School of Software, Pingdingshan University, Pingdingshan 467000, China

Abstract: As the problem of fuzzy inquiry and data sparseness caused by intention gap between users and queries, according to the ranking list of possible subtopic from popularity and diversity, subtopic selection and sorting of stratified structure were used for web text mining. Firstly, on the basic of noun phrase and substitute of part query, a simple model was used to extract a variety of related phrases as candidate subtopic. Then, related documents of a web document collection were used to build three-level stratified structure of candidate subtopic. Finally, considering popularity and diversity, the stratified structure and estimated popularity were applied for sorting. Based on 100 Japanese queries from NTCIR-9 library, 100 English queries from TREC 2009 library and network tracking diversity task, experiments verify that the proposed method can be effectively applied to a variety of search, and the proposed mining is better than external resources for high ranking subtopics.

Key words: data sparseness, text mining, stratified structure, diversity, popularity

收稿日期: 2016-03-03; 修回日期: 2016-05-08

基金项目: 河南省科技厅科技重点攻关项目 (No. 142102210226)

Foundation Item: Key Project of Science and Technology Department in Henan Province (No. 142102210226)

1 引言

智能设备的出现大大影响了网络搜索环境,不同于PC时代,现在需要加强搜索服务,以获得准确的搜索结果,因为当代用户更倾向于在个人环境中简化查询。事实上,很多网络查询模糊不清,一些用户无法选择合适的关键词进行搜索,也有些用户省略了搜索所需的关键信息^[1]。用户意图和查询之间的差距导致了查询结果模糊宽泛^[2]。在模糊查询^[2]中,用户获取的结果可能与意图完全不同;而对于宽泛查询^[3]来说,用户获取的结果不如预期的具体。虽然网络搜索引擎已经提供查询建议服务,帮助用户探索和表达自己查询所需的信息^[4]。但是,查询建议并没有明确考虑建议查询的流行性和多样性,因此这方面的研究很有现实意义。

子主题挖掘与查询建议紧密相关,这是因为子主题和查询建议获取的结果相似。通常情况下,建议使用查询日志的查询方法。将经常共同出现在相同搜索场景的查询看作相关查询^[5],通过历史点击数据找到相似的查询,这些相似的查询通常共享大量的点击网址^[6]。并利用<查询,点击链接>二分图内建议查询之间的相似性分析来提高多样性^[7]。

除了查询日志外,也使用链接文本和外部资源^[8],并将<链接文本,点击链接>看作<查询,点击链接>,来弥补数据稀疏^[9]。参考文献[4]使用网络文档语料库中的共生词语,该方法建议包含原始查询中靠前词语的新查询及靠后词语的短语,但是,因为 n 元语法词语,这些建议查询可能是不完整的短语。

NTCIR-9子主题挖掘任务推动了中日两种语言不同方法的发展。给中文(SougouT)和日文(ClueWeb09-JA)提供了网络文档集,但是只给中文查询(SougouQ)提供了日志。为了获取高水平性能,参与者使用网络搜索引擎(如百度、谷歌和雅虎)的建议查询和查询日志^[10,11]。然而,日文的子主题挖掘任务在只使用外部资源网络文档的情况下,才能获得最佳性能。参考文献[12]使用从网络文档中提取出来的链接和链接文本,并不依赖于任何其他资源。但是由于这些候选子主题必须与查询匹配,会出现数据稀疏问题。

之前大多数的研究都依靠查询日志来发现子主题,这样会因为查询日志中很少或不存在罕见的查询主题而出现数据稀疏。此外,流行性和多样性并不成比例。而

好的子主题必须与给定查询相关,同时需满足高流行性和高多样性。

为了解决这些问题,本文在相关文档的基础上通过候选子主题的简单模式和分层结构来挖掘子主题。其主要贡献如下:

- 只使用网络文档集,而不是查询日志和外部资源;
- 为了找到各种各样的相关候选子主题,尽可能多地从网络文档中提取“可理解”语句,这些语句通过简单模式与原始查询完全或部分匹配;
- 本文的分层结构可以保持流行性和多样性之间的平衡。

2 提出的方法

本文方法包括两个步骤,如下所述。

(1)子主题提取

子主题提取是从网络文档中尽可能多地提取各式各样的相关候选子主题。根据名词短语和可部分替换的查询创建简单模式,然后用这些模式找到候选子主题,最后通过过滤,降低相似候选子主题的冗余度。

(2)子主题排序

子主题排序是在考虑多样性和流行性之间平衡的基础上将子主题排序。在相关文档的基础上,给提取出来的候选子主题创建分层结构,并根据结构和流行性对子主题进行排序。

2.1 简单模式下的子主题提取

在开始阶段中,最重要的事情是尽可能正确地提取各式各样的候选子主题,一旦提取出不相关或不完整的短语作为候选子主题,在下一阶段中将会受到这些错误的影响,即错误传播。相反,如果在较高标准下提取候选子主题就会提高准确性,但也会出现一些问题,如数据稀疏、多样性低等。因此本文在相对宽松的标准下提取候选子主题,假设这些子主题都包含原始查询和至少一个使原始查询更具体的名词短语。根据这一假设,只考虑包含原始查询的名词短语,这些短语必须比原始查询表述得更具体。由于名词和真实查询词语之间的比例远高于其他词类^[13],从名词中提取出来的各式各样的候选子主题可变为新查询,对发现给定查询隐藏的查询意图很有帮助。

从语法角度来说,短语的主导词是决定句法类型的词语,短语中其他词称为修饰词。主导词和修饰词之间的关系对句法分析起到重要作用。将子主题的结构定义如下:



- 名词短语+可选的其他词+查询+可选的其他词+名词短语;
- 查询+可选的其他词+名词短语;
- 名词短语+可选的其他词+查询。

这些结构都包含距查询最近的名词短语,因此它们可以在不使用句法分析的情况下尽可能多地满足主导词—修饰词之间的关系。为了提取合适的能满足该结构的候选子主题,设计的简单模式如下。

模式 1:((形容词)?(名词)(非名词))*?(查询)((非名词)*(形容词)?(名词))*?

其中,“?”代表“0”或“1”,“+”代表“ ≥ 1 ”,*代表“ ≥ 0 ”。

为了使该模式的实用性更强,对“(形容词)?(名词)”中名词短语的形式进行了限制,对名词以外的其他词语进行设定。将模式 1 应用到查询的前 1 000 个相关文档中,由 BM25 模型^[4]提取。因为该模式包含真实的短语和文档中的完整查询及名词短语,所以发现的候选子主题都与之相关。比如,句子“我们提供糙米粥的饮食食谱”,如果查询是“饮食”,使用模式 1 提取出来的候选子主题是“糙米粥的饮食食谱”。

如果查询包含 ≥ 2 个词语,模式 1 并不足以将各种样式的候选子主题提取完整,因为候选子主题和查询之间的完整匹配减少。为了改善这一缺陷,在原始查询中使用可部分替换的查询 q_{left} 和 q_{right} 进行部分查询。首先连续删除查询的右边词语产生所有可能的左边部分短语,针对每一个短语,提取前 N 位的相关文档并与原始查询比较。如果这些文档与原始查询有超过 50% 的相同度,那么该左边部分短语就是可部分替换查询 q_{left} 的候选。在包含文档最多的候选中,选出最短的 q_{left} 来提高匹配的可能性。如果没有包含超过 50% 相关文档的候选短语,选取最长的短语 q_{left} ,因为这样可以包含最多的查询信息。同理,右边部分查询 q_{right} 选取方法相似。所以每个原始查询只有一个 q_{left} 和 q_{right} 。用 q_{left} 和 q_{right} 代替查询,创建新的简单模式如下:

模式 2:((形容词)?(名词)*(非名词))*?(q_{left})(词语)*(q_{right})((非名词)*(形容词)?(名词))*?

模式 3:(q_{right})(非名词)*(形容词)?(名词)*

模式 4:(形容词)?(名词)+(非名词)*(q_{left})

值得注意的是, q_{right} 前面和 q_{left} 附近的名词短语并没有反映在模式中,因为可应用的名词短语的范围未知,使用新的模式从检索文档中提取不同短语,比如,句子“你必须注意饮食的副作用”。如果查询的是“粥饮食”,用模式 3 找出

的子主题为“粥饮食的副作用”。即使这些子主题并不是文档中真实的短语,仍可以降低数据稀疏度,并提高多样性。

为了降低冗余度,过滤相似的子主题。设 s_{wp} 是每一个子主题开始或末端的名词短语“(形容词)?(名词)+”中的一组词汇。如果有不少于两个候选子主题有完全相同的 s_{wp} ,则认为它们相似,因为 s_{wp} 包含重要的关键词,正是这些关键词决定了每一个子主题的意义。所以将相似子主题的频率信息合并,并选择其中频率最高的候选子主题,该频率反映了用户的喜好程度。比如,假设给定的查询“粥饮食”有 3 个频率较高的候选子主题:“<“米粥的饮食食谱,”9>、<“米粥所有的饮食食谱,”9>和<“加米的粥饮食食谱,”7>。因为这些候选子主题有相同的 s_{wp} {食谱,米},将频率信息合并,如 $9+9+7=25$,然后选择“米粥的饮食食谱”。最后,如果候选子主题的频率小于阈值(实验设置为 3),该候选子主题将被排除在外,因为在实际文档中基本不出现。

2.2 使用子主题的分层结构排序

对于给定的查询,若使用排序的方法,其候选子主题的数量十分有限,因为用户并不想看到太多的子主题。然而,若仅依靠子主题的流行性来排序,大多数排序靠前的子主题极可能只包含一小部分的搜索意图。相反,若仅依靠子主题的多样性来排序,普通用户可能对排序靠前的子主题并不感兴趣,因为其流行性可能较低。因此,本文提出子主题的三级分层结构,获取数量相对较小的子主题,但是这些子主题涵盖了原始查询各式各样的搜索意图。

将查询的子主题组织成层级结构。如果查询模糊不清,有不少于两个子主题,会变成查询(根部)的语句子辈(层级 1),更具体的子主题是孙辈(层级 2)。因为可能的查询搜索意图不断缩小,层次结构的深度加深,由此可知高层级的子主题反映查询更广泛的搜索意图,层级较低对应的范围较窄。

相比低层级而言,少数层级较高的子主题包含了查询全部的意图。如果任一高层级子主题的搜索意图不够具体,其子节点(子主题)对涵盖更具体的搜索意图用途很大。然而,这种结构有时候会太具体,为了简单起见,本文提出子主题的三层分层结构,如图 1 所示。给定查询为根,子节点为“主要子主题”,每一个叶节点为“次级子主题”。主要子主题对搜索意图进行去歧化和具体化,也可能会选取一组主要子主题来满足搜索意图较高的多样性。次级子主题更加具体,缩小了主要子主题搜索意图的范围,满足了主要子主题搜索意图的多样性。

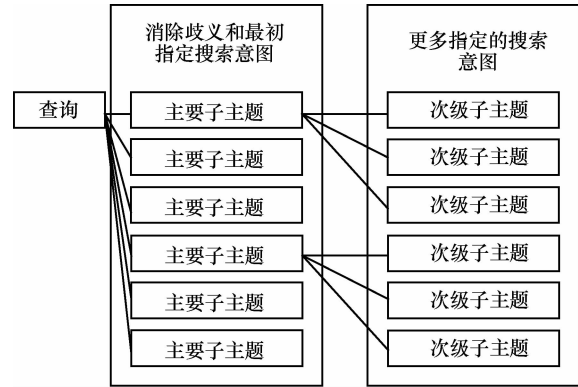


图1 子主题的三级层次结构

为了创建所提出的分层结构,假设与给定查询相关度较高的文档代表了用户所有可能的搜索意图,文档中候选子主题代表了一些搜索意图。一个主要候选子主题必须满足两个条件。首先,能包含许多高相关文档,从而反映较广泛的搜索意图。一般情况下每个相关文档只有一个主要子主题。因此,包含主要子主题的一组相关文档(如图2中集合A)很少与包含其他子主题的其他集合重叠。而且,该文档集可能包含一些子集(如图2中集A-1、A-2、A-1-1和A-1-2),从相应文档集重叠的角度来看,主要子主题和次级子主题或其他子主题截然不同,主要子主题的文档集比其他文档集的清晰度更高。

本文提出了评分衡量(selection score,SS)法,这一参数核实候选子主题包含高相关文档的数量以及相应的文档集与其他的不同之处为:

$$SS(st,US) = Coverage(st,US) \times Distinctness(D(st))$$
$$= \left| D(st) \cap US^c \right| / \left| \bigcup_{st' \in ST} D(st') \right| \times Distinctness(D(st)) \quad (1)$$

其中, st 是候选子主题, ST 是从查询排名前 N 的相关文档中提取出来所有候选子主题的集合; $D(st)$ 是包含查询排名前 N 的相关文档中的 st 文档集合; US 是联合文档集,包含之前选出的主要子主题; US^c 是 US 集合的补集。

在上述SS测量中,将覆盖率和清晰度因素相加,从而避免选择偏见,为了包含新的高相关文档(隐藏的搜索意图),这些文档并不包含在之前的主要子主题中。接着,再测量文档集的清晰度,本文选择了清晰度熵(distinctness entropy,DE)的方法^[15]。

$$DE(st) = - \sum_{sc \in ST, sc \neq st} \frac{|D(st) \cap D(sc)|}{|D(st)|} \lg \frac{|D(st) \cap D(sc)|}{|D(st)|} \quad (2)$$

这里将传统的文档聚类问题转换为短语排序问题。因为文档聚类中存在差异巨大的文档聚类,所以为了保证文档聚类应用与所提出方法的各个步骤对应,在处理过程中

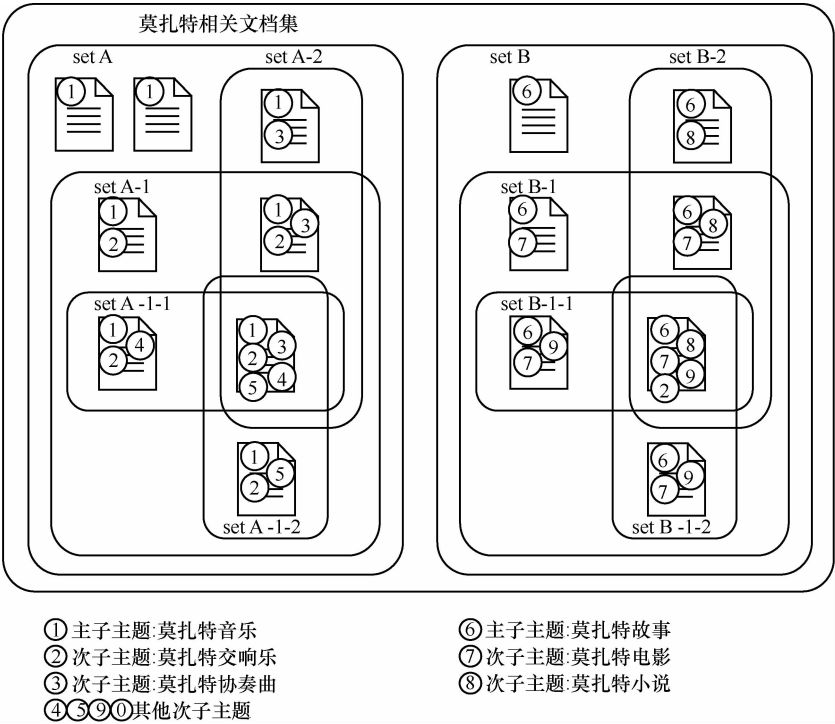


图2 查询“莫扎特”的相关文档集



添加了一些额外操作,即提取候选子主题,用子主题为聚类命名,并按照聚类间差异进行排序。

主要子主题不断使用进行选择操作。如果 $|US| = \left| \bigcup_{st' \in ST} D(st') \right|$, 则停止选择过程。因为此时已选的主要子主题能够覆盖所有高相关文档。此外,可通过增加每个主要子主题所覆盖的相关文档数量,来减少主要子主题的个数。如果不同主要子主题的相关文档相似,则相似文档集可以合并成新的单一集合,原来具有最高流行度的主要子主题名称作为合并后新文档集的名称。除了主要子主题直接关联相关文档外,可通过递归方法计算每个主要子主题的次级子主题。

如前文所述,主要子主题和次级子主题分组都是相对较小的分组,用以满足父节点的多样性搜索意图,在主要子主题和次级子主题之间存在着继承的流行度。基于这些特点,本文给出了一种依赖于三级层次结构的子主题排序序列,用以平衡子主题的流行度和多样性(如图3所示)。首先对主要子主题进行排序,通过少量子主题实现高多样性。然后选择首次排序的主要子主题的次级子主题,须考虑该节点与父节点间的流行度以及此节点搜索意图的多样性。继而,在下次循环中对主要子主题排序后,按流行度对其次级子主题进行排序。此外,经排序的主要子主题和次级子主题数量小于目标数量时,其他子主题按照流行度排序,直接添加到经排序的子主题队尾。算法1和算法2描述了子主题筛选和排序的过程。

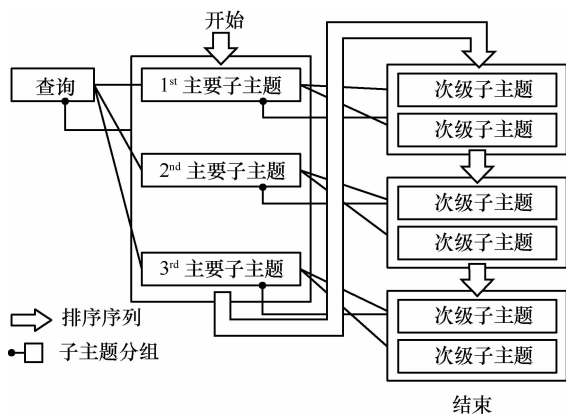


图3 子主题分组满足父节点多样性搜索意图

算法1 子主题选择

SubtopicSelection(R, ST)

输入: 文档集合 R , 和子主题候选集合 ST

输出: 子主题顺序列表 L

```

US ← {} //空集
L ← [] //空列表
while |US| < |∪_{st' ∈ ST} D(st')| do
    按照 SS(st, US) 最大值选择 st
    增加 st 到 L, 增加 D(st) 到 US
end
for i ← 1 to L do
    AD ← D(ei)
    AS ← {ei}
    for j ← (i+1) to L do
        if CosineSimilarity(D(ei), D(ej)) > 0.5 then
            AD ← AD ∪ D(ej)
            增加 ej 到 AS, 从 L 移除 ej
            j ← j - 1
        end
    end
    D(ei) ← AD
end
通过流行度来排序 L 中的所有子主题
return L
    
```

算法2 子主题排序

SubtopicRanking(HR_{query}, ST_{all}, K)

输入: 给定的前 N 个相关文档集 HR_{query} , 从 HR_{query} 提取的所有候选子主题集 ST_{all} 和子主题数量 K

输出: 最终的子主题排列列表 L_{final}

```

Lfinal ← SubtopicSelection( $HR_{query}, ST_{all}$ )
PS ← Lfinal
for i ← 1 to PS do
    R ← D(psi) // psi 是 PS 中第 i 个子主题
    添加 SubtopicSelection(R, ST) 到 Lfinal 的末尾
end
if |Lfinal| < K then
    通过剩下子主题的流行度, 添加已排序主题到 Lfinal
end
return Lfinal
    
```

3 子主题流行度估算

在网络搜索环境中,流行度与重要程度相关。因为如果消息是重要的,则表明有大量用户对此感兴趣。因此,本文通过关注子主题的重要性来估计它们的流行度。子主题

重要性衡量基于的假设是:如果在给定查询时,某一子主题是许多相关文档的重要短语,那么它就具有高重要性。本文采用词频—逆向文档频率(TF-IDF)表示目标词在文档中的重要性。子主题(st)和查询目标相关文档集(R_{query})用于 TF-IDF 的完整语料库。计算 TF-IDF 累加和与 R_{query} 中的所有相关文档作为 st 的重要性,计算方式如下:

$$\begin{aligned} STFIDF(st) &= \sum_{doc \in R_{query}} SubtopicImportance(st, doc) \\ &= \sum_{doc \in R_{query}} SubtopicImportance(st, doc) \\ &= freq(st, doc) \lg \frac{|R_{query}|}{|D(st, R_{query})|} \end{aligned} \quad (3)$$

其中, $freq(st, doc)$ 是文档 st 中 doc 的频率。 $D(st, R_{query})$ 表示 R_{query} 中包含 st 的文档集。

由于文档比子主题包含更多的信息,如果在估计子主题流行度时,额外考虑相关文档的重要性,将会得到区分度更高的子主题流行度。与查询高相关的文档包含更多关联信息,即这些文档对用户查询很重要。因此,假设流行度高的子主题能够检索到更重要的文档,就可用文档覆盖率函数(DC)来检测高相关子主题文档覆盖查询队列相关文档的数量。

$$\begin{aligned} DC(st) &= \sum_{doc \in (HR_{st} \cap HR_{query})} Weight(st, doc) \times \\ &\quad DocumentImportance(doc) \\ &= \sum_{doc \in (HR_{st} \cap HR_{query})} \frac{DS_{st}(doc)}{\text{maximum of } DS_{st}s} \times \\ &\quad DS_{query}(doc) \end{aligned} \quad (4)$$

HR_{st} 和 HR_{query} 分别是 st 与查询的前 N 个最高相关文档集。 $DS_{st}(doc)$ 和 $DS_{query}(doc)$ 分别是两者文档经 BM 25 模型计算得到的排序评分。

权重值是子主题与相关文档的关联度,用来标记文档的重要度和重要度评分。如果子主题与某文档高度相关,则给该文档的评分分配高权值。每个子主题的最终评分(Score)由 STFIDF 和 DC 的线性组合表示:

$$Score(st) = (1-\lambda) \frac{STFIDF(st)}{\text{average of STFIDF}} + \frac{DC(st)}{\text{average of DC}} \quad (5)$$

4 实验及分析

4.1 实验数据集及参数设置

为了进行实验评估,本文挖掘了来自 NTCIR-9 库的 100 个日文查询和来自 TREC 2009 库的 100 个英文查询子

主题和网络跟踪多样性任务,在 TREC 中,每个查询都有若干个子主题描述,用以描述其代表的搜索意图。网络跟踪多样性任务没有提供关于搜索意图的流行度信息。因此,评估人员手动将每个挖掘子主题归类为同一意图。如果两个评估人员都认为某一子主题不能归入同一类,则丢弃它。

给定的日文和英文查询意图平均评分为 10.91 和 4.60。实验仅使用日文文档集 ClueWeb09-JA 和英文文档集 TREC 的 B 类。实验使用日文 MeCab POS 标签和英文 Stanford POS 标签,执行名词短语划分和识别,用 PROP- x - y 标志,其中, x 是语言参数,如“J(日文)”或“E(英文)”; y 是方案参数,如“PT”、“HR”、“DC”或“DCA”。基本方案 PT 使用了最简单模式和 STFIDF (式 (3)), 主要方案 HR、DC 和 DAC 同时使用了简单模式和子主题层次结构。为了评估流行度,HR 仅使用了 STFIDF, 但是 DC 和 DAC 同时使用了 STFIDF 和 DC(式(4))。评分中的 λ 设置为 0.5(式(5)), 为了避免过度处理,限制最高相关文档数量 N 为 200。

4.2 结果和讨论

表 1~表 3 中的粗体表示所有方法中的最佳结果,灰色背景区域表示本文方法的结果;在 0.01/0.05 级的统计意义采用标识 Q/q、A/a、B/b 区分, 分别表示基于基准 BASE-J/E-QS、BASE-J/E-AC、BASE-J/E-BP 的改进结果。正如表中所示,在前 10、20、30 子主题搜索中,本文方法优于基准方法。本文的主要方案间的执行方式不一样,并且基准只具有统计意义。D#-nDCG 均值表示的日文意义概率(t-test, 2-tailed)为 0.05、0.004 和 0.009 ($p < 0.01$); 英文为 0.019、0.003 和 0.004 ($p < 0.05$)。在日文和英文的前 10 个子主题中, D#-nDCG 均值结果表明,HR、DC 和 DAC 能够使用少量(前 10 个)子主题覆盖查询的各种搜索意图,并能够适当地维护流行性和多样性平衡。对两种语言而言,与基准 BASE-J-BP、BASE-E-QS 比较, PROP-J-DCA 和 PROP-E-DC 是最好的。当 D#-nDCG@10 时,本文最好的方法提高了 9.96% 和 4.63%。

英文方面,通过使用 DC,设置评分中的 λ 为 0.5,得到最好的整体结果。均值 D#-nDCG 结果表明,当 λ 值介于 0.1 和 0.7 之间时,结果会更好(如图 4(b)所示),即子主题可从相关文档的重要性中获得合适的可区分流行度。然而,在日文子主题中,尽管 DCA 方案获得了最好的结果,但与 HR 结果没有显著的不同,当 λ 值较大时, D#-nDCG 均值减小(如图 4(a)所示)。



表 1 基准结果和本文方法对日英文排名前 10 个子主题的搜索结果

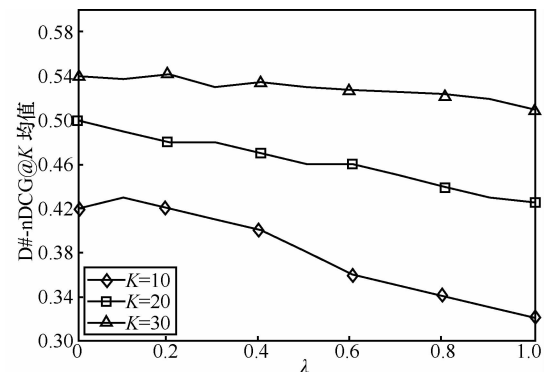
语言	方法名	均值 I-rec@10	均值 D-nDCG@10	均值 D#-nDCG@10
日文	BASE-J-QS	0.395 9	0.388 0	0.391 9
	BASE-J-AC	0.346 8	0.385 7	0.352 8
	BASE-J-BP	0.398 5	0.390 4	0.394 5
	PROP-J-PT	0.417 8	0.409 7	0.413 8 ^{qA}
	PROP-J-HR	0.437 8	0.429 3	0.433 6 ^{QAB}
	PROP-J-DC	0.428 2	0.408 7	0.418 4 ^{qAb}
	PROP-J-DCA	0.436 9	0.430 6	0.433 7 ^{QAB}
	EXT-AC	0.435 5	0.405 6	0.420 5
英文	BASE-J-QS	0.511 9	0.565 0	0.538 5
	BASE-J-AC	0.434 9	0.449 5	0.442 2
	BASE-J-BP	0.528 8	0.515 1	0.521 9
	PROP-J-PT	0.547 4	0.555 6	0.551 5 ^{Ab}
	PROP-J-HR	0.561 4	0.562 5	0.561 9 ^{qAB}
	PROP-J-DC	0.557 1	0.569 6	0.563 4 ^{qAB}
	PROP-J-DCA	0.555 2	0.562 3	0.558 7 ^{qAB}

表 2 基准结果和本文方法对日英文排名前 20 个子主题的搜索结果

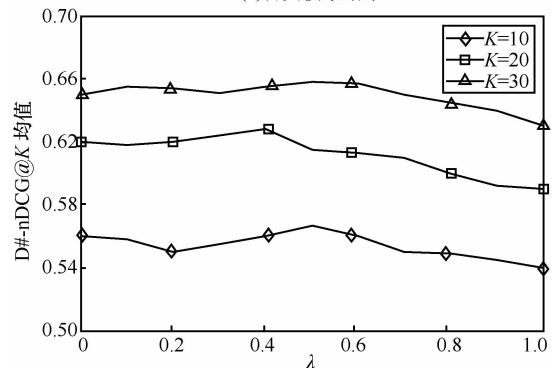
语言	方法名	均值 I-rec@20	均值 D-nDCG@20	均值 D#-nDCG@20
日文	BASE-J-QS	0.511 7	0.393 8	0.452 8
	BASE-J-AC	0.453 3	0.349 6	0.401 5
	BASE-J-BP	0.514 4	0.391 4	0.452 9
	PROP-J-PT	0.555 3	0.411 5	0.483 4 ^{qA}
	PROP-J-HR	0.581 8	0.423 9	0.502 9 ^{QAB}
	PROP-J-DC	0.559 0	0.407 2	0.483 1 ^{qAb}
	PROP-J-DCA	0.581 7	0.424 8	0.503 2 ^{QAB}
	EXT-AC	0.630 7	0.420 1	0.525 4
英文	BASE-J-QS	0.607 0	0.563 0	0.585 0
	BASE-J-AC	0.552 0	0.433 2	0.492 6
	BASE-J-BP	0.618 1	0.512 7	0.565 4
	PROP-J-PT	0.635 7	0.570 0	0.602 9 ^{Ab}
	PROP-J-HR	0.664 0	0.570 8	0.617 4 ^{QAB}
	PROP-J-DC	0.670 9	0.578 0	0.624 5 ^{QAB}
	PROP-J-DCA	0.670 4	0.571 6	0.621 0 ^{QAB}

表 3 基准结果和本文方法对日英文排名前 30 个子主题的搜索结果

语言	方法名	均值 I-rec@30	均值 D-nDCG@30	均值 D#-nDCG@30
日文	BASE-J-QS	0.567 4	0.401 6	0.484 5
	BASE-J-AC	0.515 3	0.350 8	0.433 1
	BASE-J-BP	0.566 2	0.391 5	0.478 9
	PROP-J-PT	0.624 5	0.424 2	0.524 4 ^{qA}
	PROP-J-HR	0.648 9	0.425 4	0.537 1 ^{QAB}
	PROP-J-DC	0.638 1	0.417 4	0.527 8 ^{qAb}
	PROP-J-DCA	0.649 7	0.424 7	0.537 2 ^{QAB}
	EXT-AC	0.653 3	0.386 7	0.520 0
英文	BASE-J-QS	0.660 1	0.570 6	0.615 4
	BASE-J-AC	0.592 8	0.420 6	0.506 7
	BASE-J-BP	0.663 0	0.503 2	0.583 1
	PROP-J-PT	0.711 8	0.579 8	0.645 8 ^{Ab}
	PROP-J-HR	0.725 7	0.577 6	0.651 7 ^{QAB}
	PROP-J-DC	0.727 9	0.580 5	0.654 2 ^{QAB}
	PROP-J-DCA	0.727 4	0.578 4	0.652 9 ^{QAB}



(a) 日文搜索结果



(b) 英文搜索结果

图 4 日文和英文结果的前 K 个子主题的 D#-nDCG 均值与评分 λ 值比较

此外,日文查询中的最好方法甚至优于 EXT-AC,该方法使用了来自微软内部网络搜索平台的外部网络文档。即通过使用附加的网络文档,本文方法可以更有效地利用有限的资源,并有可能获得更高的性能。同时,因为 DC 和 DCA 更关注子主题的流行度,所以前 10 个子主题并没有包含对多样性影响很大的子主题。

在每种语言的前 10、20、30 个子主题中,D#-nDCG 均值持续改进,这意味着在考虑流行度和多样性平衡时,主

要方案可以提取更多相关和多样性的子主题。当与最好的基准 BASE-J-BP/QS 和 BASE-E-QS 比较时, PROP-J-DCA 和 PROP-E-DC 的 D $\#$ -nDCG@20/30 均值分别改进了 11.10/10.89% 和 6.75/6.31%。

英文方面, 通过使用 DC, 设置评分中的 λ 为 0.5, 得到最好的整体结果。在前 20、30 个子主题中, DCA 方法取得了比 HR 更好的效果。此外, 基于各种 λ 值 (从 0.0 到 1.0, 步长为 0.1), 均值 D $\#$ -nDCG 结果表明, 当 λ 值介于 0.1 和 0.7 之间时, 结果会更好 (如图 5(b) 所示), 即子主题可从相关文档的重要性中获得合适的可区分流行度。然而, 在日文子主题中, 尽管 DCA 方案获得了最好结果, 但与 HR 结果没有显著的不同; 在所有案例中, DC 没有比 HR 执行得更好。此外, 当 λ 值较大时, D $\#$ -nDCG 均值减小 (如图 5(a) 所示)。由于存在大量无用的术语, 这导致了较低的查询性能。

此外, 在日文和英文查询中, 通过设置查询的对应评分 λ 值, DCA 方法分别取得了最好的和相对较好的搜索性能。然而, 该方案仅考虑了子主题的表面信息, 不足以决策更合适的 DC 和 STFIDF 的权重。深入分析评估结果可知: 当相关文档具有多个意图时, DC 方法更有优势, 而 STFIDF 适合于只有一个搜索意图的文档。为每个查询检查 PROP-J-DC 和 PROPE-DC 的 D $\#$ -nDCG, 发现许多查询比 PROP-J-HR 性能高, 却比 PROP-E-HR 性能低。使用 DC 的查询都获得了高性能, 在相关文档中, 会有两个或多个具有不同搜索意图的子主题以高频率同时出现, 但是其他查询的子主题却没有遵循此模式。因此, 通过给定更合适的 DC 权重和基于 STFIDF 来判断相关文档是否具有多个意图, 来获得更好的子主题挖掘性能。

在前 10、20、30 个子主题中, 本文方法在 I-rec 和 D-nDCG 均值中持续获得良好性能。即本文方法按照子主题的相关度、流行性和多样性来确保子主题的挖掘性能。与最好的基准比较, 检索日文时, I-rec@10、I-rec@20、I-rec@30、D-nDCG@10、D-nDCG@20 和 D-nDCG@30 均值获得改善程度为 9.63%、13.07%、14.52%、10.30%、8.52% 和 5.76%; 英文则为 8.84%、10.53%、10.26%、0.81%、2.67% 和 1.73%。综上所述, 在提取子主题阶段, 简单模式是有用的, 在查找各种相关子主题时也是有效的。在子主题排序阶段, 本文的层次结构在平衡流行度和多样性方面有良好的效果, 因为与其他方法相比, HR、DC 和 DCA 的结果最佳。

5 结束语

本文提出了一种基于简单模式的子主题挖掘方法和一种层次结构候选子主题, 在日文和英文搜索中采用了网络文档。在提取子主题阶段, 各种相关候选子主题使用名词短语和替代部分查询的简单模式。在子主题排序阶段, 构建候选子主题的三级层次结构, 并用该结构和估计的流行度进行排序。在实验中, 在前 10 个子主题中, 文中所提出方法性能均优于基准, 甚至优于使用外部网络文档的方法。即本文方法只需少量维持流行度和多样平衡的子主题便可覆盖查询的各种搜索意图。在前 20、30 个子主题中, 搜索结果稳步提升, 这是因为提取的相关的和多样性的子主题越来越多。

在今后研究中, 仍有进一步改进的空间, 例如精炼候选子主题, 为每种语言设计合适的排序方法, 将本文方法与资源开放型方法结合。此外, 子主题的复杂评估也是一个重要问题, 尤其是在平衡多样性和流行度方面。

参考文献:

- [1] 唐晓波, 肖璐. 基于单句粒度的微博主题挖掘研究 [J]. 情报学报, 2014, 33(6): 214-219.
TANG X B, XIAO L. Research of micro-blog topics mining based on sentence granularity [J]. Journal of the China Society for Scientific and Technical Information, 2014, 33 (6): 214-219.
- [2] 田宇辰. 专业搜索引擎的无日志查询推荐机制研究及实现 [D]. 广州: 华南理工大学, 2014.
TIAN Y C. Research and implementation of non log query recommendation mechanism for professional search engine [D]. Guangzhou: South China University of Technology, 2014.
- [3] 李胜浩. 基于 MapReduce 的 Web 文本挖掘系统的研究与实现 [D]. 北京: 北京邮电大学, 2013.
LI S H. Research and implementation of Web text mining system based on MapReduce [D]. Beijing: Beijing University of Posts and Telecommunications, 2013.
- [4] BHATIA S, MAJUMDAR D, MITRA P. Query suggestions in the absence of query logs [C]//International ACM SIGIR Conference on Research & Development in Information Retrieval, July 24-28, 2011, Beijing, China. New York: ACM Press, 2011: 795-804.
- [5] HE J, HOLLINK V, DE VRIES A. Combining implicit and explicit topic representations for result diversification [C]//The 35th international ACM SIGIR conference on Research and development in information retrieval, August 12-16, 2012, Portland, OR, USA. New York: ACM Press, 2012: 851-860.

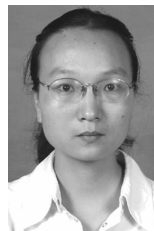


- [6] 肖璐,唐晓波. 基于句子成分的微博热点主题挖掘模型研究[J]. 情报科学, 2015, 35(11): 137-141.
XIAO L, TANG X B. Research on micro-blog hot topic mining model based on sentence composition [J]. Journal of the China Society for Scientific and Technical Information, 2015, 35(11): 137-141.
- [7] ZHU X, GUO J, CHENG X, et al. A unified framework for recommending diverse and relevant queries [C]// World Wide Web Conference Series, March 28-April 1, 2011, Hyderabad, India. New York: ACM Press, 2011: 37-46.
- [8] KIM S J, SHIN K Y, LEE J H. Hierarchical subtopic mining for topic annotation [C]// The 6th international workshop on exploiting semantic annotations in information retrieval, October 28, 2013, San Francisco, CA, USA. New York: ACM Press, 2013: 49-52.
- [9] 刘少鹏, 印鉴, 欧阳佳, 等. 基于 MB-HDP 模型的微博主题挖掘[J]. 计算机学报, 2015, 42(7): 1408-1419.
LIU S P, YIN J, OU-YANG J, et al. Topic mining from microblogs based on MB-HDP model [J]. Chinese Journal of Computers, 2015, 42(7): 1408-1419.
- [10] 岑荣伟, 刘奕群, 张敏, 等. 基于日志挖掘的搜索引擎用户行为分析[J]. 中文信息学报, 2010, 24(3): 49-54.
CEN R W, LIU Y Q, ZHANG M, et al. User behavior analysis of search engine based on log mining [J]. Journal of Chinese Information Processing, 2010, 24(3): 49-54.
- [11] 谭彩丽. 基于主题相关博客的属性挖掘模型设计 [D]. 北京: 北京邮电大学, 2011.
TAN C L. Design of attribute mining model based on topic related blog [D]. Beijing: Beijing University of Posts and Telecommunications, 2011.
- [12] DANG V, CROFT B W. Term level search result diversification[C]// International ACM SIGIR Conference on Research & Development in Information Retrieval, July 28-August 1, 2013, Dublin, Ireland. New York: ACM Press, 2013: 603-612.
- [13] 曾依灵, 许洪波, 白硕. 网络文本主题词的提取与组织研究[J]. 中文信息学报, 2008, 22(3): 64-70.
ZENG Y L, XU H B, BAI S. Research on the extraction and organization of Web text topic words [J]. Journal of Chinese Information Processing, 2008, 22(3): 64-70.
- [14] 刘德喜, 万常选, 刘喜平, 等. 基于结点权重模型的 XML 片段检索策略[J]. 计算机学报, 2013, 36(8): 1729-1744.
LIU D X, WAN C X, LIU X P, et al. XML fragment retrieval strategy based on node weight model [J]. Chinese Journal of Computers, 2013, 36(8): 1729-1744.
- [15] 刘志勇, 耿新青. 基于模糊聚类的文本挖掘算法 [J]. 计算机工程, 2009, 35(5): 44-45.
LIU Z Y, GENG X Q. Text mining algorithm based on fuzzy clustering[J]. Computer Engineering, 2009, 35(5): 44-45.

[作者简介]



史玉珍(1975-),女,平顶山学院副教授,主要从事 Web 数据挖掘、社团发现方面的研究工作。



单冬红(1976-),女,平顶山学院副教授,主要从事数据挖掘方面的研究工作。