

一种基于半监督学习的多模态 Web 查询精化方法

姜 远 黎 铭 周志华

(南京大学计算机软件新技术国家重点实验室 南京 210093)

摘 要 Web 搜索系统往往通过与用户的交互来精化查询以提高搜索性能. 除文字之外, 网页中还存在着大量其它模态的信息, 如图像、音频和视频等. 以往对于查询精化的研究很少涉及对多模态信息的利用. 文中提出了一种基于半监督学习的多模态 Web 查询精化方法 M^2S^2QR , 将 Web 查询精化转化为一个机器学习问题加以解决. 首先, 基于用户判断后的网页信息, 分别为不同模态训练相应的学习器, 然后利用未经用户判断的网页信息来提高学习器性能, 最后将不同模态学习器结合起来使用. 实验验证了文中方法的有效性.

关键词 机器学习; 半监督学习; 多模态信息; Web 搜索; 查询精化

中图法分类号 TP18 **DOI 号:** 10.3724/SP.J.1016.2009.02099

Multi-Modal Web Search Query Refinement Based on Semi-Supervised Learning

JIANG Yuan LI Ming ZHOU Zhi-Hua

(National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093)

Abstract Web search systems usually improve search performance by interacting with users to refine queries. In addition to text information, usually a large amount of information of other modalities, such as image, audio and video, exist in Web pages. Few previous researches on Web query refinement, however, try to exploit the multi-modal information. This paper proposes a multi-modal Web search query refinement method M^2S^2QR based on semi-supervised learning, which transforms Web search query refinement into a machine learning problem. First, based on the information given by Web pages judged by users, classifiers are trained for different modalities, respectively. Then, Web pages that have not been judged by users are used to help improve the performance of the classifiers. Finally the classifiers of different modalities are combined to use. Experiments validate the effectiveness of the proposed method.

Keywords machine learning; semi-supervised learning; multi-modal information; Web search; query refinement

1 引 言

随着互联网的迅猛发展, Web 已经成为世界上最大的信息库, 用户往往难以高效地从中获取自己所需的信息, 因此 Web 搜索技术受到了越来越多的

关注. 在 Web 搜索中, 用户往往习惯于使用字数较少的查询关键词. 有研究指出, 用户在进行检索时使用的查询的平均长度为 2.3 个单词^[1]. 根据如此短的查询来准确地判断用户的信息需求是十分困难的. 因此, 搜索过程往往分为若干轮进行, 在每一轮中, 搜索引擎通过与用户的交互来获取信息对查询

收稿日期: 2009-07-15; 最终修改稿收到日期: 2009-09-04. 本课题得到国家自然科学基金(60635030, 60721002)、江苏省自然科学基金(BK2008018)资助. 姜 远, 女, 1976 年生, 博士, 副教授, 主要从事机器学习、信息检索、数据挖掘等方面的研究工作. E-mail: jiangyuan@nju.edu.cn. 黎 铭, 男, 1980 年生, 博士, 讲师, 主要从事机器学习、数据挖掘, 特别是利用未标记数据学习方面的研究工作. 周志华, 男, 1973 年生, 博士, 教授, 博士生导师, 主要从事人工智能、机器学习、数据挖掘、信息检索、模式识别、演化计算、神经计算等领域的研究工作.

进行精化. 目前常用的查询精化方法主要有相关反馈和关键词推荐. 相关反馈要求用户对搜索结果中的部分网页进行标记, 搜索引擎依据用户的标记信息重新产生出与用户查询更加相关的检索结果; 在关键词推荐技术中, 搜索引擎自动产生出一些推荐的关键词供用户选择, 用户从中选择某些关键词加入原有的查询中.

值得注意的是, 网页中经常包含丰富的图片、音频和视频等多种模态信息. 与文字信息相比, 这部分信息对于表示网页的内容也起到了重要作用. 然而, 以往对于 Web 搜索查询精化的研究却很少利用网页中的文字以外的其它模态的信息. 多模态信息之间的相容性与互补性使得借助它们提升交互式搜索的性能成为可能. 同时, 由于用户往往不愿意提供大量反馈信息, 这使得单个模态情况下难以获得较优的 Web 查询精化的性能.

针对上述问题, 本文提出了一种基于半监督学习的多模态 Web 查询精化方法 M^2S^2QR . 该方法通过同时利用网页中的文字和图像这两种不同模态的信息来获得更优的精化结果. 为保证每个模态上的性能, 本文采用半监督学习中的协同训练^[2]技术有效利用用户未查看的网页信息来提升查询精化的性能.

本文第 2 节简要介绍本文的研究背景, 主要包括 Web 查询精化、多模态技术和半监督学习技术; 第 3 节详细介绍本文中提出的基于半监督学习和多模态信息的 Web 查询精化方法; 第 4 节报告实验结果; 第 5 节总结本文工作, 并对以后的工作进行展望.

2 相关工作

2.1 Web 搜索查询精化

Web 搜索在与用户交互的过程中通过各种方式收集用户的反馈信息, 并利用这些信息精化搜索结果. 相关反馈和关键词推荐是交互式搜索中被广泛运用的两大技术.

相关反馈是信息检索中的一项经典技术. 在相关反馈中, 用户对搜索结果中排序靠前的若干文档进行相关性标记, 系统通过分析用户反馈对查询进行精化. 目前主要的相关反馈方法有两种: 检索词加权 (term reweighting), 即改变查询中不同检索词的权值; 检索词扩展 (term expansion), 即向查询中添加新的检索词. Rocchio^[3] 基于向量空间模型在

SMART 系统中完成了最初的相关反馈实验, 他同时使用了检索词加权和检索词扩展的方法. 在此基础上, Ide^[4] 提出了两种不同的反馈机制, 并且验证了相关反馈的有效性. Salton 和 Buckley^[5] 对多种相关反馈算法进行了实验比较, 发现这些方法之间的差距并不明显. Robertson 和 Jones^[6] 以及 Robertson^[7] 分别提出了基于概率模型的检索词加权和检索词扩展方法. Harman^[8] 通过实验验证了在检索词加权的基础上, 检索词扩展对于概率模型的重要性以及多轮相关反馈的有效性.

在关键词推荐中, 搜索引擎首先自动地从当前搜索结果中产生出一些推荐的关键词供用户选择, 然后基于用户的选择进行更进一步的搜索. 早期关于关键词推荐的研究可以分为两大类, 即全局分析和局部分析. 在全局分析中, 产生候选关键词所用的信息来源于所有文档; 而在局部分析中, 只利用搜索结果中排序靠前的文档. 全局分析中的主要技术包括词聚类方法^[9]、相似度辞典 (similarity thesauri) 方法^[10]、Phrasefinder^[11] 等. 局部分析中的主要技术包括局部聚类^[12]、局部上下文分析^[13] 等.

2.2 多模态信息利用

Web 网页中多模态信息之间的关联性和互补性已引起了研究者的关注. Woodruff 等人^[14] 提出了一种将整个网页缩略为一张图片的方法, 该图片对网页中的重要文字进行加亮处理, 使得用户可以清楚阅读. Wang 等人^[15] 通过对网页中的图像和图像周围的文字之间的关联矩阵及相似度矩阵的叠代来进行相似度的传播. Xue 等人^[16] 提出了“图像摘录” (image snippet) 的概念, 并在搜索结果提交时将文字摘录和图像摘录一起提供给用户. 网页的图像摘录是反映网页的部分或全部内容的、存在于该网页中的图片. 如果网页包含多幅图像, 则在接收到不同的查询时所产生的图像摘录可能是不同的, 在理想情况下, 与查询主题最相关的图像才会成为图像摘录. 由于人类对图像的阅读和理解速度要快于文字, 通过使用图像摘录, 可以帮助用户更加快速地判断搜索结果是否确实是自己所需要的. Zhou 和 Dai^[17] 提出了 WebSEIC 框架, 在搜索过程中利用网页中包含的图像的视觉特征来改善搜索效果, 大致来说, 如果一个网页同时在多个模态中都被判断为与查询相关, 那么这个网页将比另一个仅仅在文字上被判断为相关的网页更可能与查询主题相关. 在多模态信息综合利用中, 显式考虑不同模态的相关性通常可获得更优的性能^[18-19]. 最近, Li 等人^[20] 提

出了利用多模态信息的相互作用来帮助学习的统一框架,该框架能有效利用文字、图像等不同模态信息的相似度进行相互增强,从而学得更能反映高层语义的相似度度量。

2.3 半监督学习

在许多现实应用中往往很容易获得大量的未标记数据(unlabeled data),但是对数据进行标记却需要花费较大代价,因此,研究如何利用未标记数据的半监督学习成为机器学习研究中的一个热点领域。目前已经有很多成功的半监督学习技术,与本文工作最有关的是协同训练(co-training)。最早的协同训练算法是 Blum 和 Mitchell^[21]提出的,该算法要求问题域存在两个充分冗余视图(sufficient and redundant views),即两个不同的属性子集,每个子集都包含了足够训练出完美学习器的信息。该算法在每个视图上利用有标记数据训练一个学习器,然后两个学习器通过对未标记数据进行预测,互相提供新的有标记训练样本。在 Blum 和 Mitchell 的工作之后,此类技术受到了广泛的重视,在理论分析、算法设计、实际应用等方面都有很多重要的工作出现,目前被称为基于不一致性的半监督学习方法^[22],逐渐发展成为半监督学习的主流技术之一。值得注意的是,原始协同训练算法^[21]要求问题域存在两个充分冗余视图,这是一个非常强的假设,最近, Wang 和 Zhou^[23]对此类算法所能奏效的理论条件进行了研究,其结论是在 PAC 条件下,只要能够产生两个有充分差异的学习器,协同训练过程就可以发挥作用,这一结果为不存在两个充分冗余视图的任务利用协同训练技术提供了理论支撑。关于这方面的研究进展可以参考文献^[22, 24],本文不再赘述。

3 M²S²QR 方法

互联网上的页面通常借助图像、声音、视频等多种媒体形式来配合页面中的文字对该网页所表达的主题进行说明。由于不同模态的数据都从一个侧面对相同主题进行描述,这些多模态信息具有相容且互补的特性。对这些多模态信息进行有效利用,可更好地把握该网页所包含的语义,从而有助于对查询进行精化。在此,本文对同时包含文字和图像两个模态的信息网页进行处理。设同时包含文字和图像的网页集合为 $X = \{(x_{i1}, x_{i1}), (x_{i2}, x_{i2}), \dots\}$, 其中 x_{ik} 对应于网页中的文字信息, x_{ik} 对应于网页中图像摘

录信息($k=1, 2, \dots$)。

在查询精化过程中,用户首先需要向系统提交初始查询,获得该查询由文字摘录和图像摘录组成的搜索结果,并提供系统相关反馈以备精化查询。然而,在实际搜索过程中,绝大部分用户并不愿意专门为系统提供相关反馈。因此,为了从用户获取足够的反馈信息,通常可以参考用户在初始搜索结果排序靠前的网页的点击信息(click-through information)^[12]。本文将用户浏览的网页标记为相关,未被点击的网页标记为不相关。设标记为相关的网页集合为 P , 标记为不相关的网页为 N 。 P_{txt} 和 N_{txt} 分别表示 P 和 N 中网页文字信息构成的文档集合, P_{ima} 和 N_{ima} 分别表示 P 和 N 中网页的图像摘录所构成的集合。在获得用户反馈以后,查询精化算法需要根据用户反馈信息学习一个排序函数 $R(x|P, N)$, 使得排序函数对与查询相关网页的输出尽可能大。

和传统查询精化算法仅利用文字信息来构造排序函数不同,本文提出综合考虑文字和图像上所提供的排序信息来定义最终网页的排序函数。受文献^[17]的启发,本文定义排序函数如下:

$$R(x|P, N) = \eta R_{\text{txt}}(x_{\text{txt}} | P_{\text{txt}}, N_{\text{txt}}) / \text{Max}_{\text{txt}} + (1 - \eta) R_{\text{ima}}(x_{\text{ima}} | P_{\text{ima}}, N_{\text{ima}}) / \text{Max}_{\text{ima}} \quad (1)$$

其中 $R_{\text{txt}}(\cdot)$ 为基于文字信息构造的排序函数, $R_{\text{ima}}(\cdot)$ 为基于图像信息构造的排序函数, Max_{txt} 和 Max_{ima} 分别为 $R_{\text{txt}}(\cdot)$ 和 $R_{\text{ima}}(\cdot)$ 在 $P \cup N$ 上的最大输出值,用于进行规范化。 η 用于调整两个学习器在最后合并结果中的权重。然而,由于本文所用的用户反馈中可能包含一定数量的噪声,但是考虑到用户在反馈过程中引入的噪声,学习器的性能有可能在反馈过程中产生波动。因此简单地将 η 设为常数有可能使得某个性能良好的学习器受到另一个性能较差的学习器的影响。为了应对上述问题,本文利用每个学习器对已标记数据的检索结果来计算 η 。具体来说,设文本学习器和图像学习器对于所有网页的排名分别为 Rank_{txt} 和 Rank_{ima} , 已经被用户标记为相关的网页为 W 。分别计算出 W 在两个排名中的名次总和:

$$S_{\text{txt}} = \sum_{x \in W} \text{rank}_{\text{txt}}(x) \quad (2)$$

$$S_{\text{ima}} = \sum_{x \in W} \text{rank}_{\text{ima}}(x) \quad (3)$$

设 m (本文中令 $m=0.5$) 为之前对于两个学习器权重的估计,则本文中用 $1/S_{\text{txt}}$ 和 $1/S_{\text{ima}}$ 对原权重进行调整:

$$\eta = \left[\frac{S_{\text{ima}}}{S_{\text{txt}} + S_{\text{ima}}} + m \right] = \frac{(m+1)S_{\text{ima}} + mS_{\text{txt}}}{2(S_{\text{txt}} + S_{\text{ima}})} \quad (4)$$

确定如何结合不同模态上的排序函数后,下一步就需要确定每个模态下应该如何构造排序函数,使得排序函数在与查询相关文档或图像上的输出值较大.直观上说,如果一个网页对象和多数相关网页比较相似并且和所有不相关网页相似度较小,那么该网页极有可能也是一个相关网页.基于此,本文分别定义图像和文字两个模态下的排序函数为

$$R_{\text{txt}}(x_{\text{txt}} | P_{\text{txt}}, N_{\text{txt}}) = \frac{\sum_{y \in P_{\text{txt}}} \text{Sim}_{\text{txt}}(x_{\text{txt}}, y)}{|P_{\text{txt}}|} - \frac{\sum_{y \in N_{\text{txt}}} \text{Sim}_{\text{txt}}(x_{\text{txt}}, y)}{|N_{\text{txt}}|} \quad (5)$$

$$R_{\text{ima}}(x_{\text{ima}} | P_{\text{ima}}, N_{\text{ima}}) = \frac{\sum_{y \in P_{\text{ima}}} \text{Sim}_{\text{ima}}(x_{\text{ima}}, y)}{|P_{\text{ima}}|} - \frac{\sum_{y \in N_{\text{ima}}} \text{Sim}_{\text{ima}}(x_{\text{ima}}, y)}{|N_{\text{ima}}|} \quad (6)$$

式(5)和式(6)中 Sim_{txt} 和 Sim_{ima} 分别为文字相似度和图像相似度, $|A|$ 表示集合 A 中元素的个数.在利用式(6)计算基于图像信息的相似度时,若网页中含有多个图像,则该网页的相似度为网页中所有图像相似度的最大值.对于 Sim_{txt} 的计算这里采用余弦夹角,如式(7)所示.其中, $(w_{x_1}, w_{x_2}, \dots, w_{x_m})$ 和 $(w_{y_1}, w_{y_2}, \dots, w_{y_m})$ 分别为文档 x 和 y 的向量表示.

$$\text{Sim}_{\text{txt}}(x, y) = \frac{\sum_{k=1}^m w_{xk} w_{yk}}{\sqrt{\sum_{k=1}^m w_{xk}^2} \sqrt{\sum_{k=1}^m w_{yk}^2}} \quad (7)$$

对于 Sim_{ima} 的计算,为了充分利用网页中文字和图像之间的关联信息,这里使用了 Wang 等人^[15]的方法.首先将网页分割成多个块,每个块由若干图像以及这些图像周围的文字组成,在块内的信息被认为具有相近的语义.在此,本文采用利用 VIPS(Vision-based Page Segmentation)算法^[25]进行网页分割.然后,计算出所有图像之间的相似度矩阵,如式(8)所示.式(8)中 \mathbf{X}_i 为编号为 i 的图像基于视觉特征的向量表示, Eud 为欧氏距离.

$$K_{ij} = 1 - \frac{Eud(\mathbf{X}_i, \mathbf{X}_j)}{\max_{i,j} Eud(\mathbf{X}_i, \mathbf{X}_j)} = 1 - \frac{\sqrt{(\mathbf{X}_i - \mathbf{X}_j)(\mathbf{X}_i - \mathbf{X}_j)^T}}{\max_{i,j} \sqrt{(\mathbf{X}_i - \mathbf{X}_j)(\mathbf{X}_i - \mathbf{X}_j)^T}} \quad (8)$$

接着计算出所有块中文字之间的相似度矩阵,如式(9)所示.式(9)中 x_i 为编号为 i 的块中的文字信息.

$$G_{ij} = \text{Sim}_{\text{txt}}(x_i, x_j) \quad (9)$$

在得到 \mathbf{K}, \mathbf{G} 之后,图像和块之间的传递矩阵的定义如下.

$$Z_{ij} = \begin{cases} 1/\theta_i, & \text{图片 } i \text{ 属于块 } j \\ 0, & \text{图片 } i \text{ 不属于块 } j \end{cases} \quad (10)$$

当上述 3 个矩阵计算完毕时,通过下式进行叠代直至收敛,最后得到的 $\hat{\mathbf{K}}$ 是初始 \mathbf{K}, \mathbf{G} 的非线性组合.

$$\begin{cases} \hat{\mathbf{K}} = \alpha \mathbf{K} + (1-\alpha)\lambda \mathbf{Z} \hat{\mathbf{G}} \mathbf{Z}' \\ \hat{\mathbf{G}} = \beta \mathbf{G} + (1-\beta)\lambda \mathbf{Z}' \hat{\mathbf{K}} \mathbf{Z} \end{cases} \quad (11)$$

式(11)中 $0 < \alpha, \beta, \lambda < 1$, α 和 β 用以控制两个矩阵在迭代过程中所占的权重, λ 保证迭代最终收敛.最后 Sim_{ima} 的计算方法为

$$\text{Sim}_{\text{ima}}(x_i, x_j) = \hat{K}_{ij} \quad (12)$$

对于上述矩阵的计算和迭代均可预先计算好,在检索时直接使用.由于形式简单,可满足网页搜索高实时性的要求.类似的排序函数定义方式已成功在文献[26]用于图像检索.

然而,在实际搜索问题中,特别是在初始搜索过程中,用户往往不愿意尝试点击更多的网页,从而导致相关网页的数量极少.同时,在该情况下盲目认为用户没有点击的网页就为不相关网页势必会引入很多噪声,从而使得从每个模态上所得的排序函数性能极差,从而影响整体的查询精化结果.事实上,除少数在初始排序中靠前的网页在用户反馈中使用以外,还有大量排序靠后的网页未被使用,造成了资源的浪费.基于此,本文引入半监督学习技术,有效利用大量未被用作反馈的网页来帮助提升排序函数的性能.在此,本文采用协同训练算法来进行半监督学习.具体来说,将每个模态的排序函数计算视为一个简单的学习器,让其相互标注一些未使用的网页,并以此更新排序函数.值得注意的是,学习器未必能为所有查询都从未标记数据中找出相关网页(因为未标记网页也许都是与查询主题无关的).受 Zhou 等人^[26-27]启发,本文对标准协同训练进行了改进,在每一轮反馈中每个学习器都将自己认为与当前查询最不相关的 n 个未标记网页视为不相关网页,用以扩充另一个学习器的训练集.为了抑制学习器在利用未标记数据时产生的错误,本文中只让这 $2n$ 个被标记的网页在当前轮中发挥作用,不将其带入下一

轮反馈.

此外,考虑到与某个不相关网页相似的网页与该不相关网页属于同一语义集的概率很大,并且对于一个特定的查询,往往只有少量网页与其相关,其余大多数网页均无关.因此,与 Zhou 等人^[26]的方法相似,本文对不相关网页进行泛化,使其能够对其对应的语义集有更好的代表性.具体来说,对于文字和图像学习器的每个不相关网页,分别通过 Sim_{txt} 和 Sim_{img} 找出与其最相似的 k 个样本.对于基于文字信息的学习器,将这 k 个样本经过规范化后的向量空间表示进行求和然后平均得到新的样本,并用该样本直接代替原来的反例;对于基于图像信息的学习器则直接用这 k 个样本代替原来的不相关网页,因为 Sim_{img} 的计算可通过查询矩阵 \mathbf{K} 直接获得,所以用一个不相关网页的集合代替一个不相关网页并不会显著增加计算开销.

基于所有上述考虑,本文提出了 $M^2 S^2 QR$ 方法.具体来说,首先训练出文字模态学习器和图像模态学习器,然后利用协同训练技术^[2]从未获得用户反馈的网页中选取若干信息更新上一步得到的学习器.最后,结合不同学习器的输出得到最终的结果.值得注意的是,互联网页面的文字模态信息和图像模态信息实际上很难满足原始协同训练算法^[21]的充分冗余视图假设,但如前所述,Wang 和 Zhou^[23]最近的研究结果表明,即使不存在两个视图,在 PAC 条件下,只要能够获得有较大差异的学习器,就可以利用协同训练技术提高性能.在本文的问题中,利用文字模态信息和图像模态信息可以产生差异较大的学习器,因此,上述结果也为本文方法提供了理论支撑.

4 实验测试

4.1 实验设置

为了模拟真实的搜索环境,本实验的网页库由 3 类查询所得到的网页构成:模糊查询、非模糊查询、热点查询.每类查询中有 6 个具体的查询,具体情况见表 1.表中的模糊查询和非模糊查询之间一一对应,每个模糊查询和与之对应的非模糊查询表达了同样的用户需求.比如“jordan basketball”和“jordan”这两个查询均试图查找关于篮球运动员乔丹的网页.

表 1 查询列表

模糊查询	apple, dove, eagle, jaguar, jordan, tiger
非模糊查询	apple fruit, dove chocolate, eagle brid, jordan basketball, tiger beer
热门查询	brad pitt, digital camera, ipod nano, janet jackson, myspace, xbox

对于每个具体查询,通过 Google 对其进行检索,将搜索结果中的前 200 个网页保存,然后利用文献^[16]中的方法获得图像摘录,舍弃其中不含有图像摘录的网页.最后将这些查询的网页汇总构成实验中用到的网页库,其中包含 3600 个网页.对每个查询,通过人工浏览网页内容将其标记为与该查询相关或不相关.

本文实验中对于文本向量的表示采用 TFIDF,图片视觉特征的提取使用了 Zhang 等人^[28]的方法.式(11)中 α, β, λ 均按照 Wang 等人^[15]文章中的说明选取, $\alpha = 0.3, \beta = 0.8, \lambda = 0.8$.矩阵的迭代次数为 10 次,反例泛化时的参数 k 设为 10,即用与每个反例最相似的 10 个样本来进行泛化.

在每轮协同训练中,两个学习器分别从未标记网页中选取 4 个网页用来扩充对方的训练集.每轮搜索结果的前 10 个网页被展示给用户并请用户进行反馈.由于用户是通过搜索结果中的文字摘录和图像摘录来进行反馈的,实验中聘请了多位志愿者根据文字摘录和图像摘录人工判断网页的相关性以作为用户的反馈信息.由于涉及昂贵的人力代价,实验中使用了 5 个具有明显视觉特征的查询,分别是 apple fruit, eagle bird, jaguar car, jordan basketball 和 tiger beer.

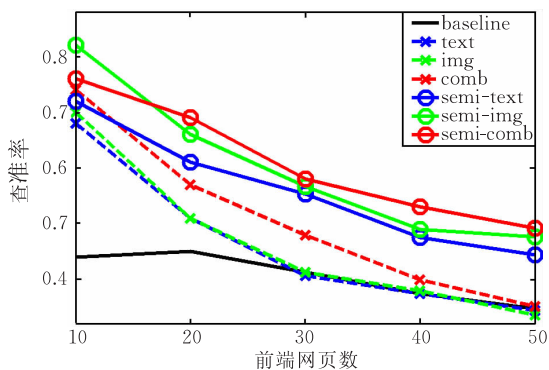
4.2 实验结果

对每个查询,未经反馈的搜索结果作为 baseline.实验中对 6 种不同的反馈方法进行了比较:(1)只使用文本信息的反馈(text), (2)只使用图像信息的反馈(img), (3)同时使用文本信息和图片信息的反馈(comb), (4)使用未标记数据后只使用文本信息的反馈(semi-text), (5)使用未标记数据后只使用图片信息的反馈(semi-img), (6)使用未标记数据后同时使用文本信息和图像信息的反馈(semi-comb).其中(1)与(4), (2)与(5), (3)与(6)用来验证使用未标记数据后搜索性能的提高.(1)与(3), (4)与(6)用以验证使用网页中图像信息对于反馈性能的帮助.对于每种反馈方法,将检索结果前端相关网页数所占的百分比作为查准率,该反馈方法在 5 个查询中的前 10、20、30、40、50 个网页中的平均

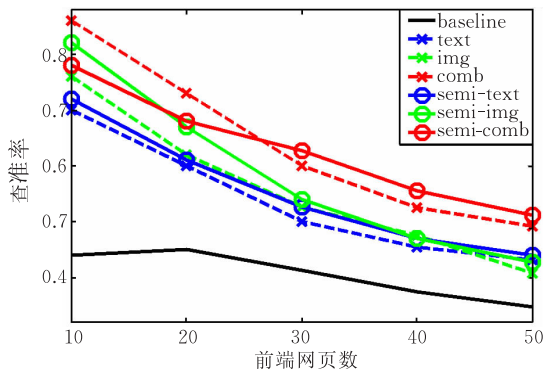
查准率被用以衡量该方法的性能.

图 1 显示了以上 6 种反馈方法在前 5 轮反馈中的查准率. 从实验结果中可以看出在前 3 轮反馈中, 利用未标记数据后明显提高了查准率, 随着反馈次数的增加, 由于有标记网页增多, 未标记数据的作用

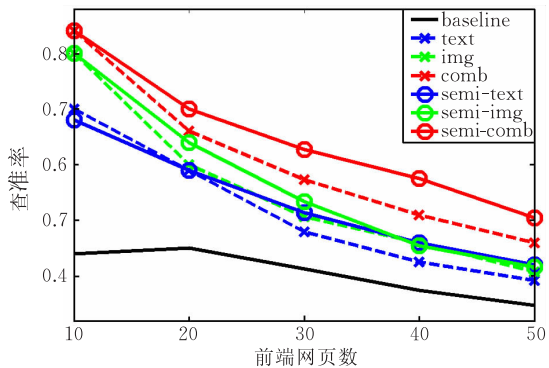
逐渐下降. 在所有前 5 轮反馈中, 同时使用图像和文字信息的检索精度均明显高于只使用文字的检索精度. 综合以上两点, 该实验验证了本文中对于未标记数据和多模态信息的使用能够有效提高搜索性能.



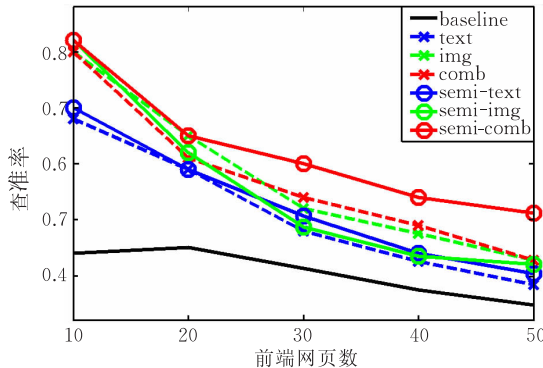
(a) 第1轮反馈



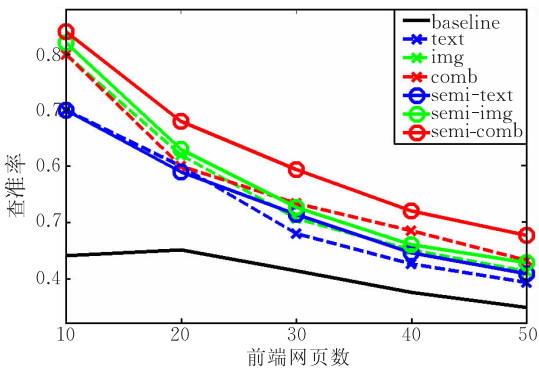
(b) 第2轮反馈



(c) 第3轮反馈



(d) 第4轮反馈



(e) 第5轮反馈

图 1 前 5 轮反馈的结果

5 结束语

本文提出了一种新型 Web 查询精化方法 M^2S^2QR . 该方法通过同时利用网页中的文字和图像这两种不同模态的信息来获得更优的精化结果. 为保证每个模态上的性能, 本文采用半监督学习技

术有效利用用户未查看的网页信息来提升查询精化的性能, 实验结果验证了该方法的有效性.

将本文方法与 Google 提示词等相结合来进一步提高查询精化的效果, 是一个值得尝试的方向. 本文直接使用网页的图像摘录作为供用户反馈的图像信息. 由于从网页中抽取的图像摘录本身可能有一定的噪音且网页中可能包含其他与网页主题相关的图

片,因此如何选择更为合适的图像提供给用户反馈是一个值得进一步研究的问题.另一方面,由于目前广泛使用的图像的视觉特征不能很好地表达高层语义,本文方法对于某些没有明显视觉特征的抽象查询有可能失效.如何针对这一类查询设计更为有效地查询精化的方法也是一个未来值得深入的重要的问题.

参 考 文 献

- [1] Silverstein C, Henzinger M, Marais H, Moricz M. Analysis of a very large AltaVista query log. Digital Systems Research Center, Technical. Report 1998-014, 1998
- [2] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training//Proceedings of the 11th Annual Conference on Computational Learning Theory. Madison, WI, 1998: 92-100
- [3] Rocchio J J. Relevance feedback in information retrieval//Salton G ed. The SMART Retrieval System. Englewood Cliffs, NJ; Prentice-Hall, 1971
- [4] Ide E. New experiments in relevance feedback//Salton G ed. The SMART Retrieval System. Englewood Cliffs, NJ; Prentice-Hall, 1971
- [5] Salton G, Buckley C. Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, 1990, 37(4): 288-297
- [6] Robertson S, Jones K S. Relevance weighting of search terms. Journal of the American Society for Information Science, 1976, 27(3): 129-146
- [7] Robertson S. On term selection for query expansion. Journal of Documentation, 1990, 46(4): 359-364
- [8] Harman D. Relevance feedback revisited//Proceedings of the 15th ACM International Conference on Research and Development in Information Retrieval. Copenhagen, Denmark, 1992; 1-10
- [9] Spark Jones K. Automatic Keyword Classification for Information Retrieval. London, UK; Butterworths, 1971
- [10] Qiu Y, Frei H-P. Concept based query expansion//Proceedings of the 16th ACM International Conference on Research and Development in Information Retrieval. Pittsburgh, PA, 1993; 160-169
- [11] Jing Y, Croft W. An association thesaurus for information retrieval//Proceedings of the Intelligent Multimedia Information Retrieval Systems. New York, NY, 1994; 146-160
- [12] Attar R, Fraenkel A. Local feedback in full-text retrieval systems. Journal of the ACM, 1977, 24(3): 397-417
- [13] Xu J, Croft W. Improving the effectiveness of information retrieval with local context analysis. ACM Transactions on Information Systems, 2000, 18(1): 79-112
- [14] Woodruff A, Faulring A, Rosenholtz R, Morrison J, Pirolli P. Using thumbnails to search the Web//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Seattle, WA, 2001; 198-205
- [15] Wang X-J, Ma W-Y, Xue G-R, Li X. Multi-model similarity propagation and its application for Web image retrieval//Proceedings of the 12th ACM International Conference on Multimedia. New York, NY, 2004; 944-951
- [16] Xue X-B, Zhou Z-H, Zhang Z. Improve Web search using image snippets//Proceedings of the 21st National Conference on Artificial Intelligence. Boston, WA, 2006; 1431-1436
- [17] Zhou Z-H, Dai H-B. Exploiting image contents in Web search//Proceedings of the 20th International Joint Conferences on Artificial Intelligence. Hyderabad, India, 2007, 2928-2933
- [18] Rege M, Dong M, Hua J. Clustering web images with multi-modal features//Proceedings of the 15th ACM International Conference on Multimedia. Augsburg, Germany, 2007; 317-320
- [19] Jeon J, Lavrenko V, Manmatha R. Automatic image annotation and retrieval using cross-media relevance models//Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Toronto, Canada, 2003; 119-126
- [20] Li M, Xue X-B, Zhou Z-H. Exploiting multi-modal interactions: A unified framework//Proceedings of the 21st International Joint Conference on Artificial Intelligence. Pasadena, CA, 2009; 1120-1125
- [21] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training//Proceedings of the 11th Annual Conference on Computational Learning Theory. Madison, WI, 1998; 92-100
- [22] Zhou Z-H, Li M. Semi-supervised learning by disagreement. Knowledge and Information Systems, to appear
- [23] Wang W, Zhou Z-H. Analyzing co-training style algorithms//Proceedings of the 18th European Conference on Machine Learning. Warsaw, Poland, 2007; 454-465
- [24] Zhou Zhi-Hua. The co-training paradigm in semi-supervised learning//Zhou Zhou-Hua, Wang Jue eds. Machine Learning and Application. Beijing: Tsinghua University Press, 2007; 259-275(in Chinese)
(周志华. 半监督学习中的协同训练风范//周志华, 王珏主编. 机器学习及其应用. 北京: 清华大学出版社, 2007; 259-275)
- [25] Cai D, Yu S, Wen J-R, Ma W-Y. Block based Web search//Proceedings of the 27th ACM International Conference on Research and Development in Information Retrieval. Sheffield, UK, 2004; 456-463
- [26] Zhou Z-H, Chen K-J, Dai H-B. Enhancing relevance feedback in image retrieval using unlabeled data. ACM Transactions on Information Systems, 2006, 24(2): 219-244
- [27] Wang W, Zhou Z-H. On multi-view active learning and the combination with semi-supervised learning//Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland, 2008; 1152-1159
- [28] Zhang R, Zhang Z, Li M, Ma W-Y, Zhang H-J. A probabilistic semantic model for image annotation and multi-modal image retrieval//Proceedings of the 10th IEEE International Conference on Computer Vision. Beijing, China, 2005; 846-851



LI Ming, born in 1980, Ph. D., assistant professor.

JIANG Yuan, born in 1976, Ph. D., associate professor. Her research interests include machine learning, information retrieval and data mining.

His research interests mainly include machine learning and data mining, especially in learning with labeled and unlabeled examples.

ZHOU Zhi-Hua, born in 1973, Ph. D., professor, Ph.D. supervisor. His research interests mainly include artificial intelligence, machine learning, data mining, information retrieval, pattern recognition, evolutionary computation and neural computation.

Background

Web search systems usually improve search performance by interacting with users to refine queries. In addition to text information, usually a large amount of information of other modalities, such as image, audio and video, exist in Web pages. How to use such multi-modal information to help improving the search performance has drawn much attention of many researchers. Previous studies include combining features of different modality for final use, fusing decision from multiple modalities, exploiting the correlation between multi-

modal information for better performance, etc. These methods have been applied to many real-world applications such as video retrieval, events detection, speaker identification, etc. However, few previous studies attempt to exploit the multi-modal information to help refining the Web search queries. Thus, this paper proposes a multi-modal Web search query refinement method M^2S^2QR based on semi-supervised learning, which transforms Web search query refinement into a machine learning problem.