

一种新的随机游走的异构网络节点排序算法研究

贾丽娟

(西北民族大学数学与计算机科学学院 兰州 730030)

摘 要 异构网络中往往包含多种不同类型的节点。为了满足用户对不同类型节点的检索需求,需要对不同类型的节点进行排序。同时鉴于用户对不同类型节点的检索需求不同,提出了一种基于随机游走的异构网络节点排序算法,以提高检索结果的性能。首先,对包含用户、图片和文本内容的异构网络进行了形式化描述。其次,对异构网络中节点之间的相似性进行了定义,并提出了一种包含不同类型邻居节点的相似性计算方法。接下来,基于随机游走思想提出了节点重要性的计算算法,并分析了随机游走模型中偏好向量的选取方法。最后采用大量真实数据进行了实验验证,结果证明了提出的方法在处理信息检索请求时的有效性。

关键词 异构网络, 图片检索, 随机游走, 排序算法

中图分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.1.036

Random Walk Based Node Ranking Algorithm in Heterogeneous Networks

JIA Li-juan

(School of Mathematics and Computer Science, Northwest University for Nationalities, Lanzhou 730030, China)

Abstract There are usually nodes of different types in heterogeneous networks. In order to satisfy requests of information retrieval for different types from users, different types of nodes need ranking. In order to solve the above problem and improve the effectiveness at the same time, this paper proposed a random walk based node ranking algorithm in heterogeneous networks. Firstly, we formally described a heterogeneous network including nodes of users, images and text. Secondly, we defined the similarity between nodes, and proposed a modified similarity measure that includes neighbors of different types. Thirdly, we proposed an algorithm for computing importance of node based on random walk, and analyzed the selection of bias vector in the random walk model. Finally, according to massive experiments on real dataset, we validated the effectiveness of the proposed approach while handling users' request in information retrieval.

Keywords Heterogeneous networks, Image retrieval, Random walk, Ranking algorithm

在 Web 2.0 时代,图片检索是社会网络应用中的重要研究内容之一^[1]。随着网络中图片数量的急剧增长,Facebook 和 Flickr 等社会网络聚集了大量用户上传的图片,基于用户关键字查询的图片检索方法可以有效应用图片库中的图片为用户提供不同的服务^[2]。虽然学术界和工业界对文本内容的排名和检索有广泛而深入的研究,但是基于关键字的图片搜索还处于起步状态。在文本内容检索中,系统根据文档间的链接关系对文档的重要性进行排名,最著名的文档排名方法是学者秦拯^[3]提出的 PageRank 排序算法,该方法根据已访问的链接对块进行相关性反馈,最后得出了较好的搜索精确度,但是对于数据量较大的文本网页,其效率较低;学者邱砾^[4]提出了采用倒排表对文档集合中的所有单词建立倒排索引,在响应用户的关键字查询时,将包含关键字的文档按照已经确定的重要性进行排序并返回给用户,该方法能够很好地对文档进行排序,但是对于数据量较大的文档集合,该方法排序效率同样较低。

区别于文本文档,社会网络中包含了丰富的不同层次的

异构实体信息,如用户、社团、图片、标签、评论、描述文本等,以及实体间的相互联系^[5]。在基于关键字的图片检索过程中,可以充分利用网络中的异构实体及连接信息,进一步加强图片检索的用户体验。本文将社会网络中的异构信息抽象为用户、图片和文本^[6],经过抽象后,社会网络变成了一个包含用户、图片和文本 3 种实体以及实体间相互联系的异构网络,并可以将该异构网络用图论的方式进行描述。在该异构网络的各种类型的连接中,用户与图片之间的连接表明了用户对图片的偏好,文本与图片之间的连接表明了图片的文本描述,用户与用户之间的连接表明了社会网络关系^[7]。由于该异构网络中包含着多种类型的节点和连接,因此基于关键字的图片检索变得更加复杂。

本文为了处理用户对不同类型节点的查询请求,在研究了异构网络中的节点排序基础上,提出了一种基于随机游走的异构网络节点排序算法;对包含用户、图片和文本内容的异构网络进行了形式化描述;对异构网络中节点之间的相似性进行了定义,并提出了一种包含不同类型邻居节点的相似性

到稿日期:2015-01-27 返修日期:2015-04-30 本文受国家自然科学基金项目(61163066,60902074),中央高校基本科研业务费专项资金项目(31920150078)资助。

贾丽娟(1979—),女,硕士,副教授,主要研究方向为中文多文种信息处理。

计算方法。最后采用大量真实数据进行了实验验证,结果表明了本文提出的方法在处理信息检索请求时的有效性。

1 异构网络模型

由于社会网络中含有多种不同类型的节点,如用户、图片和文本等信息,本文定义如下的异构网络模型。异构网络中含有的节点类型集合为 D , 并且 $|D| = N \geq 2$ 。异构网络为 $G = (V, E, C)$, 其中节点 $v_i^n \in V$ 的类型为 $n \in N$, 边集中的元素 $(v_i^n, v_j^{n'}) \in E$, 内容集合 C 中的元素 c_i 与节点 v_i^n 之间是一一对应的。用节点的内容计算节点之间的相似性, 从而描述节点之间的隐式连接。当 $N=1$ 时, 网络中仅仅包含一种类型的节点, 因而是同构的网络。

给定异构网络 G , 可以将其转化为 N 个同构的子网络以及子网络间的相互连接。如图 1 所示, 每个子网络 G^{T_n} 仅仅包含类型为 $T_n \in D$ 的所有节点, 以及该类型节点之间的相互连接, 对于异构的连接 $(v_i^{T_n}, v_j^{T_{n'}}) \in E^H (T_n \neq T_{n'})$, 在子网络 G^{T_n} 和 $G^{T_{n'}}$ 之间建立一条边。异构网络 G 与其转化为的子网络之间的关系表示为:

$$\begin{aligned} V(G) &= V(G^{T_1}) \cup V(G^{T_2}) \cup \dots \cup V(G^{T_N}) \\ E(G) &= V(G^{T_1}) \cup V(G^{T_2}) \cup \dots \cup V(G^{T_N}) \cup E^H \\ C(G) &= C(G^{T_1}) \cup C(G^{T_2}) \cup \dots \cup C(G^{T_N}) \end{aligned} \quad (1)$$

其中, E^H 为 G 中端节点类型不同的异构连接, 元素 $(v_i^{T_n}, v_j^{T_{n'}}) \in E^H (T_n \neq T_{n'})$ 在转化后的网络中连接两个不同的子网络 G^{T_n} 和 $G^{T_{n'}}$ 。

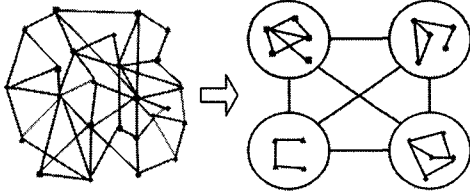


图 1 异构网络的子网络抽象

以上所述为通用的异构网络模型。为了对该模型进行实例化从而使其适用于本文研究的社会网络中的图片排名, 需要将节点的类型实例化为用户、图片和文本 3 种类型。实例化过程如下:

- 用户是社会网络中最基本的节点类型。在社会网络中, 用户可以标记图片, 进行评论, 或者加入具有相似兴趣的用户团体。用户与用户之间的连接为社会网络连接; 用户与图片之间的连接表示用户发布了该图片; 用户与文本之间的连接表示用户发布了相关的评论或者进行了相应的标记行为。

- 图片是图片共享社会网络特有的节点类型, 也是本文研究的排序算法的实体。图片与文本之间的连接表示相应的文本内容为图片的评论或者标记等。此外, 为了考虑图片之间的相似性, 在一定的相似性阈值下引入了图片和图片之间的隐式连接。

- 文本主要用于描述图片的内容。文本与用户以及图片之间连接的描述同上。

在 Web 2.0 应用中, 大多网络中都包含用户、图片和文本 3 种类型的信息, 因而都可以用上述的异构网络模型进行描述。

2 基于随机游走的图片排序

在异构网络中, 为了响应用户对图片的检索, 本文采用随机游走的方法对图片进行排名。令异构网络中的节点类型集合为 $D = \{T, I, A\}$, v^T 表示文本节点, v^I 表示图片节点, v^A 表示用户节点。为了有效地对节点进行排序, 用随机游走在异构网络上的稳态概率 r_v 表示节点 v 的重要性。

异构网络中除了包含显式的连接外, 还需要根据节点间的相似性计算出相应的隐式连接。本节首先定义了不同类型节点的相似性计算方法, 然后提出了一种异构网络上基于随机游走的节点排序算法, 最后分析了随机游走过程中偏好向量的选取方法。

2.1 相似性定义

对于类型为 $h \in D$ 的两个节点 v_i^h 和 v_j^h , 它们之间的相似性定义为 $S_{ij}^h = \text{sim}(c_i^h, c_j^h)$, 其中 c_i^h 和 c_j^h 分别为节点 v_i^h 和 v_j^h 的内容。由于随机游走过程直接对相似性值进行编码, 因而节点间的相似性影响着随机游走的方向。

相似性函数根据两个节点的内容将取值在 $[0, 1]$ 上, 即 $\text{sim}_h: (C, C) \rightarrow [0, 1]$ 。对于不同类型的节点, 应当采用不同的相似性计算方法。 S_{ij}^h 的值越大, 表明两个节点之间的内容越相似, 反之表明两个节点越不相似。对于图片、文本和用户 3 种类型的节点, 相应的相似性计算方法分别如下。

- 图片: 为了计算图片 v_i^I 和 v_j^I 之间的相似性, 分别将这两个图片用列向量 w_i 和 w_j 表示, 然后采用如下 3 种方法计算相似性。

1. 单词的共现数 (number of co-occurrence term, COT):

$$S_{ij}^I = \frac{\langle \delta(w_i), \delta(w_j) \rangle}{\|\delta(w_i)\| \cdot \|\delta(w_j)\|} \quad (2)$$

其中 $\delta(\cdot)$ 为项敏感阶跃函数, $\langle \cdot \rangle$ 为内积。

2. 词频 (term frequency, TF):

$$S_{ij}^I = \frac{\langle w_i, w_j \rangle}{\|w_i\| \cdot \|w_j\|} \quad (3)$$

3. TF-IDF (term frequency-inverse document frequency):

$$S_{ij}^I = \frac{\langle w_i, w_j \rangle_K}{\|\sqrt{K} \cdot w_i\| \cdot \|\sqrt{K} \cdot w_j\|} \quad (4)$$

其中, K 是对角矩阵, 矩阵中的元素 K_{ii} 为单词 i 的反向文档频率的对数的平方, 即 $K_{ii} = (\log \frac{\text{文档数}}{\text{包含单词 } i \text{ 的文档数}})^2$ 。

$\langle \cdot, \cdot \rangle_K$ 为经过 Hermitian 矩阵 K 特征化后的内积。

- 文本: 应用规范化后的单词共现比例作为文本间的相似性。给定文本节点 v_i^T 和 v_j^T , 分别用它们的词频来表示内容 c_i 和 c_j , 那么

$$S_{ij}^T = \frac{\langle \delta(c_i), \delta(c_j) \rangle}{\|\delta(c_i)\| \cdot \|\delta(c_j)\|} \quad (5)$$

其中, $\delta(\cdot)$ 和 $\langle \cdot \rangle$ 的定义同式 (2), 分子表示两个文本中匹配的单词的个数。

- 用户: 本文将用户分为若干组, 同一组内用户的相似性为 1, 不同组之间的用户的相似性为 0。如果采用用户组来表示矩阵, 那么有 $S^A = I$ 。

采用上述的相似性定义进行随机游走时, 仅仅考虑了相同类型节点之间的相似性, 而忽视了其它类型节点对该类型节点的影响。为此, 本文将相同类型节点之间的相似性定义进行扩展, 使其既包含相同类型节点的影响, 也包含不同类型

节点的影响。本文采用类似于 SimRank^[8] 的思想,即两个节点的相似性递归地依赖于与节点相连的邻居节点。区别于 SimRank,本文中的邻居节点既包含同类型的节点,也包含不同类型的节点。以图片节点为例,给定节点 v_i^I 和 v_j^I ,修改后的相似性 $\tilde{S}_{ij}^I$ 为:

$$\begin{aligned}\tilde{S}_{ij}^I &= f(S_{ij}^I, S^A, r^A, S^T, r^T) \\ &= S_{ij}^I + \sum_{h \in \{T, A\}} \beta^h \left(\sum_{a \in V_i^h, b \in V_j^h} S_{ab}^h r_a^h r_b^h \right)\end{aligned}\quad (6)$$

其中,节点集合 V_a^h 和 V_b^h 中的节点分别为 v_i^I 和 v_j^I 的邻居节点,并且满足 $V_i^h = \{v | v \in V^h, (i, v) \in E^h\}$, β^h 为类型为 h 的权重。当 $h=A$ 或者 T 时, $\tilde{S}_{ij}^I$ 和 $\tilde{S}_{ij}^T$ 的定义同上。如果分别用 d 和 h 表示不同类型的节点,那么修改后的节点相似性可用如下矩阵形式表示:

$$\tilde{S}^d = S^d + \sum_{h \in D - \{d\}} \beta^h L^{dh} R^h S^h R^h L^{dh} \quad (7)$$

其中, L^{dh} 为 G^d 与 G^h 之间的连接,当节点 i 和 j 之间存在连接时, $L_{ij}^{dh} = 1$, 否则 $L_{ij}^{dh} = 0$; R^h 为对角矩阵,并且 $R_{ii}^h = r_i^h$ 。

2.2 随机游走算法

对节点之间的相似性定义进行修改后(见式(7)),在每个子网络内分别进行随机游走过程,从而确定节点的相对重要性。用 α 表示随机游走过程的重启概率, p 表示重启后节点的偏好向量(该向量依赖于用户的查询),那么节点在第 t 次迭代后的重要性值向量的计算公式如下:

$$r^{(t)} = \alpha \tilde{S}^h D^{-1} r^{(t-1)} + (1-\alpha) p \quad (8)$$

其中, D 为对角矩阵,并且满足 $D_{ii} = \sum_{j=1,2,\dots} \tilde{S}_{ij}^h$ 。根据式(8)所示的随机游走模型,本文提出了 HeteroRank 算法用于计算节点的重要性,算法的伪代码如下所示。

算法 HeteroRank

输入: 每种类型节点 $d \in D$ 的相似性矩阵 S^d , 网络中的异构连接 E^h , 以及参数 α 和 β ;

输出: 每种类型节点 $d \in D$ 的重要性向量 r^d ;

1. 根据式(9)、式(10)初始化偏好向量 p ;
2. 令 $r^{(0)} = [1/n]_{N \times 1}$;
3. Repeat
4. For 每个子网络类型 $h \in D$ do
5. 根据式(7)计算修改后的相似性矩阵 $\tilde{S}^h$;
6. 应用式(8)进行随机游走计算 $r^{h,(1)}$;
7. End for
8. $t = t + 1$
9. Until r^d 收敛或者满足最大迭代次数;
10. 输出 $\forall d \in D, r^d$

在算法 HeteroRank 中,首先要确定偏好向量 p ,具体方法见 2.3 小节。在对 $r^{(0)}$ 进行初始化后,算法进入循环阶段。在应用式(8)计算 $r^{h,(t)}$ 时,采用矩阵计算时计算量会非常大,本文采用随机游走采样的方法进行。随机游走的采样方法可参见文献[9]。

2.3 查询偏好向量选取

在式(8)中,偏好向量 p 是查询敏感的,通过调整 p 的取值可以适应不同的用户查询,从而满足查询的个性化需求。给定查询 Q ,节点类型 T 的节点为 $v_j^T (1 \leq j \leq n)$ 及相应的内容 c_j^T ,其偏好向量可通过如下公式计算。

$$p_{v_i^T}^T = g\left(\sum_{q_j \in Q} \text{sim}(q_j, c_i^T)\right) \quad (9)$$

其中, g 为泛化函数,用来保证偏好向量 p 中元素的和为 1;

$\text{sim}(q_j, c_i^T)$ 为 Q 中第 j 个查询关键字与文本节点 v_i^T 之间的语义相似性。

通过指定偏好向量 p ,上述随机游走算法不仅可以用于查询关键字,还可以用来进行图片的检索。为了确保可以通过图片进行检索,需要在图片子网络内进行随机游走,并根据查询图片和图片库中图片间的相似性来初始化偏好向量 p 。在图片查询时,偏好向量 p 的计算方法如下。

$$p_{v_i^I}^I = g(\text{sim}(q, c_i^I)) \quad (10)$$

3 实验结果与分析

3.1 数据集与评价指标

实验采用公开的 Flickr 数据集^[10]。该数据集中包括 140 个用户组,每个组中的用户都关于某个领域有着相同的兴趣。此外,该数据集包含 118000 张图片,这些图片中共包含 150000 个不同的文本标签。在对文本信息去除停止词并提取词干后,选取出现频率最高的 500 个单词作为图片的标记集合。本文采用信息检索的查准率(Precision)、召回率(Recall)、AP(Average Precision)和 MAP(Mean Average Precision)^[11] 作为算法性能评估的指标。

3.2 实验结果

为了评估算法的性能,将本文提出的 HeteroRank 算法与 SVMRank^[12] 和 VisualRank^[13] 两种算法进行对比。

首先,实验对比了 3 种算法的查准率和召回率,实验结果如图 2 所示。从该图中可以看出, SVMRank 和 VisualRank 两种算法的 Precision-Recall 分布相近,而本文提出的 HeteroRank 算法的分布位于其它两种算法右上方。这表明,当查询结果具有相同的 Precision(或者 Recall)时, HeteroRank 算法的 Recall(或者 Precision)优于其它两种算法,因而 HeteroRank 算法具有更好的检索性能。

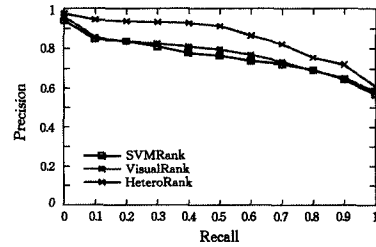


图 2 算法的检索性能(Precision-Recall)对比

其次,实验对比了 3 种算法的 MAP 随着检索结果中图片数量的变化。在该组实验中,查询项为图片,检索结果为图片的排序结果。从图 3 中可以看出,随着检索结果中图片数量的增加,3 种算法的 MAP 基本保持稳定。在 3 种结果的对比中, SVMRank 算法的 MAP 是最低的, VisualRank 居中,而本文提出的 HeteroRank 算法的 MAP 是最高的。

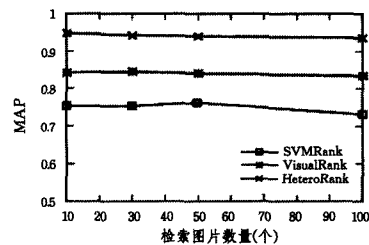


图 3 算法的 MAP 对比

接下来,通过实验分析了 HeteroRank 算法的检索性能 AP 随着参数 α 的变化情况。从图 4 可以看出,随着 α 的增大,HeteroRank 算法的 AP 先增大,在 $\alpha=0.6$ 处取最大值,然后下降。在传统的类 PageRank 算法中,如 VisualRank^[13],算法的 α 通常取值为 0.85。然而在本文提出的包含用户、图片和文本类型节点的异构模型中,在计算节点之间的相似性时,由于对相似性定义进行了修改,因此 α 的最优值发生了变化。

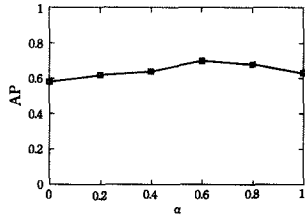


图 4 算法的 AP 对比

最后,令 HeteroRank 算法的 $\alpha=0.6$,对比了 3 种算法在检索结果为 100 个的情况下的 AP,即 AP@100。实验结果如表 1 所列,其中黑体部分为该项查询的最优结果。在该组实验中,对节点间的相似性标准进行了细分,分别包含了 COT、TF 和 TF-IDF,同时以 10 个不同的英文单词作为查询关键字进行检索。从该表中可以看出,HeteroRank 算法在处理单个关键字的查询时,其检索结果明显优于其它两种算法。

表 1 算法的 AP@100 对比

查询	SVMRank			VisualRank			HeteroRank		
	COT	TF	TF-IDF	COT	TF	TF-IDF	COT	TF	TF-IDF
balloon	0.489	0.767	0.867	0.869	0.86	0.869	0.959	0.961	0.96
bird	0.462	0.69	0.903	0.903	0.874	0.89	0.984	0.992	0.99
box	0.731	0.89	0.931	0.927	0.907	0.906	0.994	0.994	0.995
car	0.467	0.611	0.838	0.867	0.86	0.867	0.987	0.92	0.915
cat	0.439	0.516	0.622	0.623	0.651	0.643	0.985	0.98	0.981
child	0.703	0.672	0.876	0.873	0.884	0.875	0.94	0.94	0.938
flower	0.612	0.699	0.763	0.804	0.73	0.754	0.963	0.911	0.99
guitar	0.605	0.738	0.802	0.811	0.804	0.819	0.832	0.84	0.833
pump	0.727	0.741	0.77	0.802	0.781	0.781	0.774	0.774	0.774
water	0.8	0.832	0.84	0.938	0.938	0.94	0.972	0.97	0.971
均值	0.604	0.716	0.821	0.842	0.829	0.834	0.94	0.918	0.935

结束语 基于 Web2.0 的社会网络中通常包含多种不同类型的节点,为了满足用户对不同类型节点的检索需求,需要对不同的节点进行排序。本文基于随机游走的思想,研究了异构网络中不同类型节点的排序。首先,对包含用户、图片和文本内容的异构网络进行了形式化描述。其次,对异构网络中节点之间的相似性进行了定义,并提出了一种包含不同类型邻居节点的相似性计算方法。接下来,基于随机游走思想提出了节点重要性的计算算法,并分析了随机游走模型中偏好向量的选取方法。最后,通过大量真实数据的实验验证了本文提出的方法在处理信息检索请求时的有效性。

参考文献

[1] Dou Bing-lin, Li Shu-song, Zhang Shi-yong. Social Network Analysis Based on Structure[J]. Chinese Journal of Computers, 2012, 35(4): 741-753(in Chinese)
窦炳琳,李澍淞,张世永. 基于结构的社会网络分析[J]. 计算机学报, 2012, 35(4): 741-753

[2] Zhao Rong-ying, Wang Jing. Research of International Social Network Analysis in Frontier Domains in Visualized Information [J]. Document, Informaiton & Knowledge, 2011, 3(1): 88-94(in Chinese)
赵蓉英,王静. 社会网络分析(SNA)研究热点与前沿的可视化分析[J]. 图书情报知识, 2011, 3(1): 88-94
[3] Qin Zheng, Zhang Ling, Li Na. Application of an Improved PageRank in Web Crawler[J]. Journal of Computer Research And Development, 2006, 43(6): 1044-1049(in Chinese)
秦拯,张玲,李娜. 改进的 PageRank 在 Web 信息搜集中的应用 [J]. 计算机研究与发展, 2011, 43(6): 1044-1049
[4] Kuang Li, Deng Shui-guang, Li Ying, et al. Using Inverted Indexing to Facilitate Composition-oriented Semantic Service Discovery[J]. Journal of Software, 2007, 18(8): 1911-1921(in Chinese)
邝砾,邓水光,李莹,等. 使用倒排索引优化面向组合的语义服务发现[J]. 软件学报, 2012, 18(8): 1911-1921
[5] Liu Wei. Study on serveral issues of comteut-based image retrieval[D]. Hangzhou: Zhejiang University, 2007(in Chinese)
刘伟. 图像检索中若干问题的研究[D]. 杭州: 浙江大学, 2007
[6] Li Xiao-yan, Chen Gang, Shou Li-dan, et al. An image clustering algorithm in collaborative tagging system[J]. Journal of Image and Graphics, 2010, 15(11): 1635-1643(in Chinese)
李晓燕,陈刚,寿黎但,等. 一种面向协作标签系统的图片检索聚类方法[J]. 中国图象图形学报, 2010, 15 (11): 1635-1643
[7] 谢东升. 基于文本的图片搜索引擎的研究[D]. 上海: 同济大学, 2008
[8] Jeh G, Widom J. SimRank: a measure of structural-context similarity[C]// Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2012: 538-543
[9] Tabei S M A, Burov S, Kim H Y, et al. Intracellular transport of insulin granules is a subordinated random walk[J]. Proceedings of the National Academy of Sciences, 2013, 110(13): 4911-4916
[10] Hare J S, Lewis P H. Automatically annotating the mir flickr dataset: Experimental protocols, openly available data and semantic spaces[C]// Proceedings of the International Conference on Multimedia Information Retrieval. ACM, 2011: 547-556
[11] Yue Y, Finley T, Radlinski F, et al. A support vector method for optimizing average precision[C]// Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2012: 271-278
[12] Joachims T. Training linear SVMs in linear time [C]// Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2011: 217-226
[13] Jing Y, Baluja S. Visualrank: Applying pagerank to large-scale image search[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 30(11): 1877-1890
[14] Wang Hua-qi, Wang Bin. Application of Optimized Proximal Support Vector Machine in Image Retrieval [J]. Journal of Chongqing University of Technology(Natural Science), 2014, 28 (9): 66-71(in Chinese)
王华秋,王斌. 优化的邻近支持向量机在图像检索中的应用[J]. 重庆理工大学学报(自然科学版), 2014, 28(9): 66-71