一种针对大规模网络图像的自动标注改善算法

王 斌[①] 俞能海[①②]
[①](中国科学技术大学电子工程与信息科学系 合肥 230026)
[②](中国科学技术大学多媒体计算与通信教育部-微软重点实验室 合肥 230026)

摘 要：在对网络图像进行索引时，人们往往利用网页中图像周围的文字作为其近似标注信息，但是这些文字信息质量不高，不足以良好地描述图像内容。该文提出一种综合利用图像视觉特征、相关文本信息以及词汇间语义关系的方法对这些不精确的文本信息进行改善，从而提高图像的索引和搜索质量。在大规模数据集上的实验证明了所提出的方法能够有效改善图像的标注。

关键词：自动图像标注；标注改善；多模态学习
中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2009)02-0270-05

An Algorithm for the Automatic Annotation Refinement on Large-Scale Web Images

Wang Bin[①] Yu Neng-hai[①②]
[①](*Department of Electrical Engineering and Information Science University of Science and Technology of China*, *Hefei* 230026, *China*)
[②](*MOE-MS Key Laboratory of Multimedia Computing and Communication*, *University of Science and Technology of China*, *Hefei* 230026, *China*)

Abstract: When Web images are indexed, the textual information in the hosting web pages are usually used as approximate image description. However, such information is not accurate enough. In this paper, a framework is proposed to utilize the visual content, the textual context, and the semantic relations between keywords to refine the image annotation. Experiments on large-scale dataset demonstrate the effectiveness of the proposed method.

Key words: Automatic image annotation; Annotation refinement; Multimodality learning

1 引言

随着数码相机、可拍照手机等数字化影像设备的普及，图像数量快速增长。而 Internet 的普遍应用使图像的发布与获取变得更加容易。但是日益丰富的图像数量也使人们难以在浩如烟海的大量数据中找到自己需要的信息。为了满足人们对搜寻图像的需求，许多搜索引擎及专业图像网站都向用户提供了对图像的检索和浏览等服务。在检索图像时，用户可以有多种方法输入检索信息。目前，基于关键词的检索是一种最有效的方式。

基于关键词的搜索要求图像具有能够表达其内容的对应文本描述。文本描述的质量越高，对图像内容的描述越准确，搜索得到的结果就会越正确，相关图像的数量就会越多。但是对于大规模图像数据，例如上亿张图像，依靠人们手工标注文本描述是不可能的。在近几年，人们开始研究自动图像标注(automatic image annotation)问题。

2007-05-19 收到，2008-11-25 改回
国家自然科学基金(60672056)资助课题

文献[1]将自动标注类比于自然语言处理中不同语言间的翻译问题，利用机器翻译的方法建立视觉词汇(图像区域)和文本词汇之间的映射关系，将一幅图像转换为其对应的文本描述。文献[2]首先计算各幅图像之间的相似性，并根据相似性进行标注词汇的传播。文献[2]改进了文献[1]中图像相似性的计算方法，结果更加准确。此外，文献[3, 4]都尝试了利用词汇之间的语义关系以提高标注性能。还有一类普遍采用的模型是生成式模型，即假定存在未知的、不可观察的若干隐变量，并依此建立起视觉特征和文本词汇的联合概率分布模型，如文献[5]提出的隐式 Dirichlet 模型(Latent Dirichlet Model, LDM)。这些方法一般都要求有高质量标注的、干净的训练集供学习使用，这使它们难以扩展到大规模应用当中。而且，传统的自动图像标注方法主要利用待标注图像的视觉信息，网络图像本身的文本信息没有得到充分利用。网络图像的一个显著特点是可以获得一些相关的低质量文本信息，如 URL,锚文本(anchor text)，ALT 标记文本等等。这些文本提供了与图像相关但是并不完整准确的信息，现有的图像搜索引擎往往依靠这些低质量文本信息实现图像的

索引和检索。

为了提高网络图像标注信息的质量，本文提出了一种对低质量标注信息进行改善的方法。首先，利用图像的相似性在图像之间用基于流形(manifold)的方法进行标注信息传播。这样，一幅图像缺失的标注信息可以从与其相似的图像中获得。图像相似性可以来自视觉信息，也可以来自其它特性，如文本信息、所在网页信息等等。其次，进一步利用词语之间的语义相关性，从一幅图像的众多候选关键词中挑出最相关、最具代表性的词，滤除掉无关噪声词汇。文中提出综合利用基于词库的统计信息以及结构化电子词典来计算词语间的语义关系。实验证明该方法能够综合利用网络图像的各种信息，对标注信息进行有效改善。

2　基于流形的学习方法

基于流形的学习方法是一种半监督学习方法[6]。半监督学习方法是指在训练阶段，训练数据和测试数据本身是已知的，但是只有训练数据的标记(分类)信息已知，而测试数据的标记信息未知。与传统的有监督学习和无监督学习相比，半监督学习可以利用更多数据，所以往往能够取得更好的性能。基于流形的学习方法适用于总体数据量较大、训练数据量较小的情况，主要思想是充分利用数据的总体分布特征和原始标注信息，使最终标注既充分平滑，又能与训练数据充分拟合。

设有图 $G=(V,E)$，其中顶点集合 V 表示图像集合，即每个顶点对应一幅图像， E 为有权重的边，描述了顶点(图像)间的相似性，每一顶点 V_i 有相应的标注信息 Y_i 。基于流形的学习方法是在该图 G 上最小化目标代价函数：

$$Q(F)=\frac{1}{2}\left(\sum_{i,j=1}^{n}W_{i,j}\left\|\frac{F_i}{\sqrt{D_{i,i}}}-\frac{F_j}{\sqrt{D_{j,j}}}\right\|^2+\mu\sum_{i=1}^{n}\|F_i-Y_i\|^2\right) \quad (1)$$

其中 F_i 是顶点 V_i 的最终分类信息。式(1)中右边的第 1 项是标注信息的平滑性度量，第 2 项是标注结果与原始标注信息之间偏差的度量，反映了与训练数据之间的拟合程度。两种度量通过系统参数 μ 进行加权。给定 μ 之后，最优化目标可以通过迭代方式得到，其过程如下所示：

(1)计算邻接矩阵 $\boldsymbol{W}$ ；

(2)正则化 $\boldsymbol{W}$ 为 $\boldsymbol{S}=\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ [2]，其中 $\boldsymbol{D}$ 为对角阵，$D(i,i)$ 为 $\boldsymbol{W}$ 中第 i 行元素之和；

(3)迭代：$F^{(t+1)}=\alpha\boldsymbol{S}F^{(t)}+(1-\alpha)Y$ ，其中 t 为迭代次数，$\alpha\in[0,1]$， $F^{(0)}=Y$ 为初始标注信息；

(4)按照最终状态 F^* ，对数据集中的点进行分类。

可以证明[6]，由于相似性矩阵 $\boldsymbol{S}$ 的特征根满足 $-1\le e_k\le 1$，所以，迭代过程最终收敛到。

$$F^*=(1-\alpha)(I-\alpha\boldsymbol{S})^{-1}Y \quad (2)$$

3　网络图像标注改善

网络图像的一个重要特征是可以从其所在网页中提取出一定的相关文本信息，如 URL，锚文本(anchor text)，ALT 标签(ALT tag)等等。这些文本信息与图像内容相关，是一种低质量的描述。但是这种文本常出现的问题是：(1)遗漏关键词。这是因为网页文本一般只偏重于图像与该网页相关内容的描述。(2)存在噪声词汇。由于图像是为网页内容服务的，所以文本信息中经常包含一些与图像无关的词汇。这两种现象都极大地降低了图像搜索的质量。

针对这些问题，本文提出一种改善网络图像标注信息的方法，其框架如图 1 所示。首先将标注信息在相似图像之间进行传播，从而，缺失的标注词汇可以从相似图像那里得到。这里的相似性可以是指视觉相似性，也可以利用其它方式得到，例如图像文本信息等。框架中包含了综合利用多种不同相似性的方法。其次，对于一幅图像可能获得的大量标注词汇，需要挑选出最相关的若干关键词，滤除掉噪声词汇。本文提出综合利用基于统计的方法和利用结构化电子词典的方法。

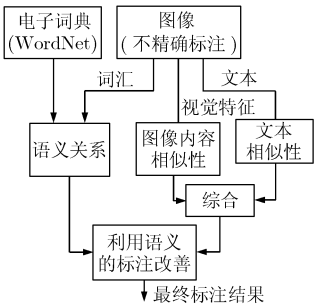


图 1 网络图像标注改善系统示意图

3.1 基于图像间相似性的方法

图像是一种视觉特征媒体，视觉上的相似往往意味着它们有一些共同的语义信息。这一节首先讨论利用视觉相似性在网络图像之间进行标注信息的传播，然后指出类似的方法也可用到基于文本信息得到的图像间关系，最后给出了综合利用多种相似性的方法。

3.1.1 基于图像内容的相似性　设两幅图像 I_i 与 I_j 间基于视觉内容的相似性可以按照某种相似性度量表示为 $\mathrm{Sim}_c(i,j)$。对于所有图像，可以得到相似性矩阵 $\mathbf{Sim}_c=\{\mathrm{Sim}_c(i,j)\}$，经过正则化后，对应传播矩阵为 $\boldsymbol{S}_c=\boldsymbol{D}_c^{-1/2}\mathbf{Sim}_c\boldsymbol{D}_c^{1/2}$，其中的 $\boldsymbol{D}_c$ 为对角阵， $D_c(i,i)$ 为 $\mathbf{Sim}_c$ 中第 i 行元素之和。

设原有标注信息矩阵 $\boldsymbol{T}^0$ (每一行对应一幅图像，每一列对应一个关键词)，那么，标注信息可以根据内容相似性 $\boldsymbol{S}_c$ 进行迭代传播：

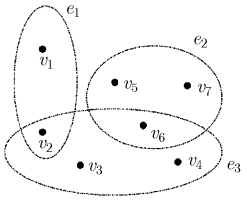
$$T_c^{(t)}=\alpha\boldsymbol{S}_cT_c^{(t-1)}+(1-\alpha)\boldsymbol{T}^0 \quad (3)$$

上标 t 表示迭代次数，$0<\alpha<1$ 是参数，用来调整传播速度。在许多传统标注方法中，这种传播过程只进行一次。由于网络图像具有数据量大的特点，采用基于流形的学习方法更加合理。根据式(2)可以写出基于流形的学习方法得到的结果为

$$T_c=(1-\alpha)(1-\alpha\boldsymbol{S}_c)^{-1}\boldsymbol{T}^0 \quad (4)$$

3.1.2 基于网络文本的相似性 对于网络图像，其周围的文本信息也具有一定的内容相关性，所以，可以利用文本信息计算图像之间的相似性。文本信息往往表示为高维矢量，每一维对应一个词汇，可以利用矢量点积或 KL 距离计算文字相关性。但是这种计算两两之间关系的方法丢掉了大量有用信息。例如，三幅图像使用了同一个词作为其标注，那么只考虑两幅图像之间文本信息的方法无法反映这种关系。所以，需要从图像的完整文本描述信息出发计算它们的文本相似度。

超图(hypergraph)是普通图模型的一种扩展[7](图 2)。假设 $G=(V,E)$ 表示一超图，V 表示顶点集合，E 表示超边集合，每一条超边连接了多个点(不少于两个点)并具有权重 $w(e)$ 。可以认为每一条超边定义了一个子图。当对所有超边有 $w(e)=2$ 时，超图退化为普通图模型。假设超图 G 中每一个顶点对应一幅图像，每一条超边表示一个关键词，这样可以将词与图像之间的关系完整的表达出来。对每一条超边 e 和顶点 v，可以定义它们之间的邻接关系 $h_{e,v}$，它反映了词与图像的联系，是词汇标注质量的一种先验描述。从而，可以得到每条超边的权重 $w(e)$，每个顶点的权重 $w(v)$ (定义为所有经过该顶点的超边权重之和)，以及顶点与超边之间的邻接关系 $\boldsymbol{H}$ 。利用这些关系，标注信息可以在超图上按照以下方式进行传播[7]：



	e_1	e_2	e_3
v_1	1	0	0
v_2	1	0	1
v_3	0	0	1
v_4	0	0	1
v_5	0	1	0
v_6	0	1	1
v_7	0	1	0

图 2 超图模型[7]

$$T_t^{(t)}=\beta \boldsymbol{S}_t T_t^{(t-1)}+(1-\beta)\boldsymbol{T}^0 \tag{5}$$

$$\boldsymbol{S}_t=\boldsymbol{D}_v^{-1/2}\boldsymbol{H}\boldsymbol{W}\boldsymbol{D}_e^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{D}_v^{-1} \tag{6}$$

其中 $\boldsymbol{D}_v$ 是顶点权重构成的对角矩阵，$\boldsymbol{D}_e$ 是超边权重构成的对角矩阵，而 $\boldsymbol{H}$ 是邻接矩阵，$0<\beta<1$ 是系统参数。类似地，可以根据式(2)直接写出收敛后的状态：

$$T_t=(1-\beta)(1-\beta \boldsymbol{S}_t)^{-1}\boldsymbol{T}^0 \tag{7}$$

3.1.3 多种相似性的综合 当有多种图像间相似性可用时，如何综合利用这些相似性成为一个重要的问题。文献[8]提出了两种在基于流形的框架下综合利用多种相似性度量的方法。一种称为线性综合，即

$$S_d=\frac{\mu \boldsymbol{S}_c+\eta \boldsymbol{S}_t}{\mu+\eta} \tag{8}$$

它等效于对两种相似性进行线性加权，然后进行标注传播。另一种是进行序列综合，即

$$S_d=\frac{\mu \boldsymbol{S}_c+\eta \boldsymbol{S}_t-\mu\eta \boldsymbol{S}_c\boldsymbol{S}_t}{\mu+\eta-\mu\eta} \tag{9}$$

它等效于先用一种相似性进行传播之后，再利用另一种相似性进行传播。

3.2 基于关键词语义关系的方法

传统的自动图像标注的一个重要问题是“语义鸿沟”(semantic gap)，即图像所能抽取出的底层视觉特征难以同图像表示的语义信息建立起明确有效的关联关系。与比较低层次的视觉特征相比，词汇有着更加明确、清晰的语义信息。利用这些词语之间的相互关系，可以改善自动图像标注的质量。

3.2.1 词义及其关系 词汇与词汇间存在着多种多样的语义关系。最常见的是层次关系，如“汽车”是“交通工具”的一种。除此之外，词汇之间还存在许多种相关性，如“汽车”与“道路”之间有着很强的联系，而且这种相关性不依赖于特定数据集，是人们在生活中大量知识的积累和反映。所以，当一幅图像已经被标上了“汽车”、“人”等词汇后，“道路”作为该图像标注词汇的概率会相应提升。为了获取这种相关信息，一种方法是利用数据集中的词汇统计信息，但是该方法受限于特定数据集。本文提出利用具有大量词汇、包含了人的知识的结构化电子词典 WordNet 来计算词汇间的关系。与统计方法相比，词典包括了更加完整的语义信息，而且不会受到数据集中噪声的影响。

3.2.2 WordNet WordNet 是自然语言处理中得到广泛应用的一种结构化电子词典[9]。在 WordNet 中，语义按照层次关系构造了多棵语义树，即子节点是父节点语义的细化。不同树上的节点之间也建有多种语义关系连接，如“is-made-of”(组成关系)、“is-an-attribute-of”(属性关系)等等。所有这些语义关系都可以用来衡量词汇之间的语义相关性。当前的 WordNet 2.1 中总共收录了超过 11 万词汇。在 WordNet 的基础上，自然语言处理领域提出了多种定量计算词汇间相关性的度量，jcn 算法被证明是最有效的方法之一[10]。它能够计算出两个概念 c_1 与 c_2 之间的关系 $\mathrm{sim}_{\mathrm{jcn}}(c_1,c_2)$ 。当一个词汇有多种语义时，它会出现在语义树上的多个概念结点。由于自动图像标注的目标是关键词，所以，定义两个词之间的语义关系为它们所有可能的语义对之间相似性的最大值，即

$$\mathrm{Sim}(w_1,w_2)=\max_{c_1\in w_1,c_2\in w_2}(\mathrm{sim}_{\mathrm{jcn}}(c_1,c_2)) \tag{10}$$

3.2.3 基于统计的方法与基于词典的方法相结合 基于词典方法的优点是语义信息及词汇关系明确、噪声少，质量高，缺点是不易扩展，对词典外词汇(如网络中出现的新词汇)不能解释。基于统计的方法能够有效地处理新词汇，缺点是受噪声影响较大。所以综合使用两种方法得到词汇相关性可以进一步提高系统性能：

$$S_w=\eta S_s+(1-\eta)S_{\mathrm{jcn}} \tag{11}$$

与基于图像的相似性类似，可以根据式(2)给出基于流形的学习方法的结果为

$$T_w^{(t)}=(\delta S_w(T_w^{(t-1)})^{\mathrm{T}}+(1-\delta)(\boldsymbol{T}^0)^{\mathrm{T}})^{\mathrm{T}},\quad 0<\delta<1 \tag{12}$$

$$T_w=(1-\delta)\boldsymbol{T}^0(1-\delta S_w)^{-1} \tag{13}$$

3.3 综合方法

上面已经提到了基于图像间相似性以及基于词汇语义关系两种标注传播方式。文献[8]讨论了两种综合利用 $\boldsymbol{S}_c$ 和 $\boldsymbol{S}_t$ 的方法。当引进了词义相关性后，它们可以整合在一个完整的算法框架之内。根据文献[6]，我们经过简单的推导可以得出系统的最终标注结果为

$$T_F=(1-\varepsilon)(1-\varepsilon S_d)^{-1}\boldsymbol{T}^0(1-\delta S_w)^{-1}(1-\delta) \qquad (14)$$

其中 S_d 是根据文献[8]对 $\boldsymbol{S}_c$ 和 $\boldsymbol{S}_t$ 进行综合的结果。从式(14)可见，两类传播过程可以交换顺序而不影响系统的最终标注。

4 实验

4.1 实验数据集

为了能够与其它文献进行比较，本文采用普遍使用的 Corel 数据集。该数据集包含 5000 幅图像，其中 4500 幅作为训练图像，500 幅作为测试图像。每幅图像有 1–5 个词作为标注，词的总数量为 371 个。每幅图像被分割为 1–10 个区域。对每一个区域，本文使用文献[1]中使用的 36 维图像特征，据此构建了两个数据集。在数据集 1 中(称为 Corel1)，每幅图像具有两个准确的标注词汇，我们将从这两个词出发改善图像标注。在数据集 1 的基础上，为了模拟网络噪声词汇，每幅图像随机添加 2 个错误的词汇,构成了数据集 2(称为 Corel2)。

第二类数据集是从网络上下载的大规模图像集合(http://office. microsoft.com)。该数据集包括 34172 幅图像和 17194 个词汇，本文利用 PorterStemming[11]算法对词汇进行处理以去除复数等变型，最终词汇量为 8985。采用与构造 Corel2 相同的方法，我们构建了一个带有大量噪声的数据集(称为 CA)。图像的内容特征为 64 维 HSV 空间的颜色直方图，而计算语义信息利用了 WordNet::Similarity 软件[12]。

最后，我们构造了一个网络图像集合，首先将 125 个词汇输入到 Google 图像搜索引擎当中，下载了每个词汇返回的前 100 幅图像及其对应网页，并利用网页分析工具从对应网页中提出每幅图像的文本描述。该数据集(称为 WebDB)与网络图像搜索引擎具有相同的处理方式和文本信息，是最贴近应用环境的数据集。

为了衡量性能，本文利用常用的两种指标，查准率(precision)是指正确返回数据与所有返回数据的比值，而查全率(recall)是返回的正确数据与所有正确数据的比值，两种指标均是对测试集中所有词进行平均得到。

查准率＝(正确数)/(正确数＋错误数)

查全率＝(正确数)/(正确数＋遗漏数)

4.2 超图性能的比较

我们首先比较超图模型与传统图模型在计算文本相似性上的性能。表 1 给出了不同的数据集上超图模型与普通图模型仅利用文本信息进行基于流形的学习方法的性能。从表 1 中可以看到超图模型在无噪数据集上可以获得更高的查准率，同时只损失很少的查全率。而在有噪数据集上，超图模型查准率受噪声影响略大，但是综合性能较传统图模型略高。

表 1 超图模型(H)与图模型(P)性能对比

	Corel1		Corel 2	
	H	P	H	P
查准率	0.844	0.784	0.462	0.464
查全率	0.712	0.732	0.540	0.534

4.3 综合性能的比较

由于以前基本没有与本文类似的工作，所以我们实现了类似于文献[13]的算法作为对比(称为 CRM-like 算法，与 CRM 的区别在于计算图像相似度时利用了与对照算法相同 HSV 颜色直方图，而不是原始 CRM 算法中的分块特征)。这是由于文献[13]是现有可公开获得的 Corel 数据集上性能最好的方法。表 2 给出了综合利用图像内容信息、文本信息以及词汇语义信息的方法与 CRM-like 的性能对比。从表 2 中可以看到，利用本文提出的方法，图像得到了更多的正确标注词汇。这是因为传统的图像标注方法没有考虑到原始数据中的噪声造成的。

表 2 本文算法与 CRM-like 性能比较

	Corel1		Corel2	
	本算法	CRM-like	本算法	CRM-like
正确标注数目	3041	2025	2279	1540
查准率	0.835	0.434	0.461	0.279
查全率	0.716	0.470	0.555	0.351

表 3 给出了本文算法应用在网络图像数据集上的结果。由于网络图像没有经过完全的标注，所以给出的评价只有查准率这一指标，而查全率无法计算。从表中可以看到，利用本文提出的算法对标注信息进行改善之后，图像搜索的准确度有了相当的提高。

表 3 本文算法在 WebDB 数据集上的性能

	改善前	利用本文算法改善后
查准率	0.669	0.714

图 3 给出了 CA 集中的几幅示例图像。可以看到，原有的图像标注中包含了两个错误的标注词汇，而经过算法处理后，错误的词汇被去除，正确的标注词汇则被添加了进来。

5 结束语

本文提出了一种综合利用多种信息对图像的文本描述进行改善的方法。该方法采用基于流形的半监督学习方法，适用于具有大量有噪数据的情况。针对网络图像的特点，本

	正确	错误
训练	Hardware, screwdrivers	Lighter, mausoleums
最终	Industrial, household. hardware, screwdrivers	lighter
实际	Industrial, household. hardwar, screwdrivers, tool	

	正确	错误
训练	Insect, wildlife	Gambia, Uranus
最终	Animal, nature, insect, wildlife	Arrive
实际	Animal, nature, insect, wildlife, creature, ladybug, beetles	

图 3 标注改善的示例

文提出利用图像内容信息、相关文本以及词汇间语义信息等多种类型的信息进行图像标注的改善工作。其中利用了超图模型来刻画文本信息之间的相关性，与传统的图模型相比，超图利用了更多的信息。计算词汇语义关系时，不仅利用了统计信息，而且利用了 WordNet 电子词典，以获取更多无偏的、数据集之外的信息。本文提出了一种统一的框架，能够使多种关系综合在一起。在大规模数据集上的实验证明了文中所提出方法的有效性。

本文没有就图像内容相似性的计算进行探讨，但是该部分具有非常重要的意义。所以，在以后的研究中如何更加准确、快速的计算图像相似度将是一个重要的课题。此外，当前的算法中只利用了 jcn 测度，我们相信，多种语义测度的综合将会进一步改善系统的性能。

参 考 文 献

[1] Duygulu P, Barnard K, and Freitas J, *et al.*. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary, European Conference on Computer Vision, Copenhagen, Denmark, 2002: 97–112.

[2] Jeon J, Lavrenko V, and Manmatha R. Automatic image annotation and retrieval using cross-media relevance models. In Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, 2003: 119–126.

[3] Wang X, Zhang L, and Jing F, *et al.*. AnnoSearch: image auto-annotation by search. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, USA, 2006, Vol. 2: 1483–1490.

[4] Deschacht K and Moens M. Text analysis for automatic image annotation. 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech, 2007: 1000–1007.

[5] Zhang R, Zhang Z, and Li M, *et al.*. A probabilistic semantic model for image annotation and multi-modal image retrieval, Springer Multimedia Systems, 2006, Vol. 12: 27–33.

[6] Zhou D, Weston J, and Gretton A, *et al.*. Ranking on data manifolds. MPI Technical Report (113), Max Planck Institute for Biological Cybernetics, Tübingen, Germany, June 2003.

[7] Zhou D, Huang J, and Schölkopf B. Beyond pairwise classification and clustering using hypergraphs. MPI Technical Report (143), Max Planck Institute for Biological Cybernetics, Tübingen, Germany, 2005.

[8] Tong H, He J, and Li M, *et al.*. Graph based multi-modality learning. In Proceedings of ACM Multimedia, Singapore, 2005: 862–871.

[9] Miller G A. WordNet: A lexical database for English. *Communication of ACM*, 1995, 38(11): 39–41.

[10] Jiang J and Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of International Conference Research on Computational Linguistics, Taipei, 1997.

[11] Porter M F. An Algorithm for suffix stripping. *Program*, 1980, 14(3): 130–137.

[12] Perdersen T, Patwardhan S, and Michelizzi J. WordNet: Similarity —— Measuring the relatedness of concepts. Proceedings of 5th NA-ACL, Boston, USA, 2004: 267–270.

[13] Lavrenko V, Manmatha R, and Jeon J. A model for learning the semantics of pictures. Proceedings of Advance in Neutral Information Processing Systems, Vancouver, Canada, 2003.

王 斌： 男，1975 年生，博士生，研究方向为网络图像搜索.

俞能海： 男，1964 年生，教授，博士生导师，研究方向为图像处理、多媒体通讯、信息检索与信息安全.