一种时变对象加权概率辨识模型

吴诗贤
(重庆工商大学计算机科学与信息工程学院 重庆 400067)

摘 要 以对象为基本检索单位的Web对象搜索技术正成为下一代智能搜索引擎的主要发展方向之一,而由于一些对象的部分属性具有时变性,高精度的时变对象辨识技术就成为实现高精度Web对象搜索的重要前提之一。从Web对象的不同属性具有对对象本质不同的表征能力、概率分布型属性值的演变大多服从某种分布以及确定型属性通常能比概率分布型属性更大程度地反映对象的区分度这些基本思想出发,提出了一种基于相似度计算的时变对象加权概率辨识模型。

关键词 时变对象,辨识,相似度,确定型属性,概率分布型属性
中图法分类号 TP391 **文献标识码** A

Weighted Probabilistic Model for Identifying Time-varying Objects

WU Shi-xian
(Computer Science and Information Engineering College, Chongqing Technology and Business University, Chongqing 400067, China)

Abstract Web object retrieval is becoming one of the main trends in the development of Intelligent Search Engines. High-precision time-varying objects recognition technology is one of the important prerequisite for high-precision Web object retrieval. Usually, different attributes have different capabilities to token the essence of the object, stochastic attribute obeys some type of distribution, and determinate attribute has better separating capacity than stochastic attribute. In this paper, starting from the above-mentioned thought a weighted probabilistic model for identifying time varying objects was proposed to improve the accuracy of identify objects.

Keywords Time-varying objects, Identification, Similarity, Determinate attribute, Stochastic attribute

1 引言

近年来,随着自然语言处理以及智能搜索引擎技术的发展,各种新的检索方法和技术不断涌现,其中,为了更加符合人类的信息处理和认知规律,一些研究学者已开始了以Web对象(而非传统页面)为目标的对象搜索技术研究[1]。在对象搜索引擎中,往往需要根据从Web中检索、抽取得到的信息来构建或识别出属于某类的对象,因此,对象识别技术是实现对象检索的基础技术之一,一般的手法是根据待识别对象的各种特征的相似度计算、归类来实现。有的Web对象是在其生命周期中各种属性不随时间而变化的时不变对象,对这样的时不变对象,只需简单计算其属性值的相似度即可完成对象识别;而有的Web对象的某些属性值却是随时间而变化的时变对象,如,"学术论文"这样一类对象,虽然其中的"作者"、"出版机构"、"出版时间"等属性值在出版时就已确定,并不随时间而变化,但其"被引用次数"等属性值却是随时间而不断变化的,对这样的时变对象,如果不把握其属性值的变化情况,很容易造成误认或漏认,大大降低对象检索的精度,随着Web对象检索需求的快速扩大,这个问题必将变得越来越重要。目前,虽然相关的学术研究报道还不多见,但也有研究学者注意到了这个问题,并提出了基于两次观测到的对象的各个属性值来自于同一对象的概率和来自于不同对象的概率之比值的时变对象辨识模型[2]。

本文把时变对象的时变属性分为3类:第一类是具有确定的随时间变化映射关系的属性(包括恒值属性),称之为确定型属性;第二类是服从某种概率分布的属性,称之为概率分布型属性;第三类是除上述两类之外的属性。本文以不含上述第三类时变属性的时变对象的辨识为研究目标,从对象的不同属性具有对对象本质的不同表征能力以及确定型属性通常能比概率分布型属性更大程度地反映对象的固有特性这一观点出发,提出了一种新的时变对象识别模型。

2 相似度模型

2.1 对象模型

设待识别对象具有N个属性,其中$1\sim M$属性为确定值属性,$M+1\sim N$属性为概率分布型属性,在前一时刻t_i、后一时刻t_j观测到的对象分别用下列向量表示:

$$o_i((x_i^{(1)},w_1),(x_i^{(2)},w_2),(x_i^{(3)},w_3),\cdots,(x_i^{(M)},w_M),(x_i^{(M+1)},w_{M+1},\cdots,(x_i^{(N)},w_N))$$

$$o_j((x_j^{(1)},w_1),(x_j^{(2)},w_2),(x_j^{(3)},w_3),\cdots,(x_j^{(M)},w_M),(x_j^{(M+1)},w_{M+1},\cdots,(x_j^{(N)},w_N))$$

其中,$x_i^{(n)}$,$x_j^{(n)}$分别表示在t_i,t_j时刻观测到的对象的第n个

到稿日期:2008-12-22 返修日期:2009-02-10 本文得到"日元贷款人才培养计划"(C01-P158)赴日访问学者项目资助。
吴诗贤 博士研究生,主要研究方向为智能情报处理,E-mail:wsx_19730813@163.com。

属性值，w_n 表示对象的第 n 个属性的权重。

2.2 相似度计算模型

设根据第 n 个属性（$M<n\leqslant N$，即该属性为概率分布型属性）在 t_i，t_j 时刻的观测值 $x_i^{(n)}$，$x_j^{(n)}$ 确定的两次观测的对象为同一对象的概率为 $p(o_i=o_j|(x_i^{(n)},x_j^{(n)}))$，两次观测的对象为不同对象的概率为 $p(o_i\neq o_j|(x_i^{(n)},x_j^{(n)}))$，利用参考文献[2]的思想，用它们的比值 $\rho(x_i^{(n)},x_j^{(n)})$ 反映两观测值为同一对象的随机属性值的可能性的大小，即

$$\rho(x_i^{(n)},x_j^{(n)})=\frac{p(o_i=o_j|(x_i^{(n)},x_j^{(n)}))}{p(o_i\neq o_j|(x_i^{(n)},x_j^{(n)}))} \tag{1}$$

根据贝叶斯定理，有

$$p(o_i=o_j|(x_i^{(n)},x_j^{(n)}))=\frac{p((x_i^{(n)},x_j^{(n)})|o_i=o_j)\times p(o_i=o_j)}{p(x_i^{(n)},x_j^{(n)})} \tag{2}$$

$$p(o_i\neq o_j|(x_i^{(n)},x_j^{(n)}))=\frac{p((x_i^{(n)},x_j^{(n)})|o_i\neq o_j)\times p(o_i\neq o_j)}{p(x_i^{(n)},x_j^{(n)})} \tag{3}$$

讨论 $p((x_i^{(n)},x_j^{(n)})|o_i=o_j)$，$o_i=o_j$ 意为前（t_i 时刻）后（t_j 时刻）两次观测到的对象为同一对象，有

$$p((x_i^{(n)},x_j^{(n)})|o_i=o_j)=p(x_j^{(n)}|x_i^{(n)})\times p(x_i^{(n)}) \tag{4}$$

讨论 $p((x_i^{(n)},x_j^{(n)})|o_i\neq o_j)$，因为 $o_i\neq o_j$，即前（t_i 时刻）后（t_j 时刻）两次观测到的对象为不同对象，$x_i^{(n)}$，$x_j^{(n)}$ 为不同对象的属性值，所以可以认为相互独立，则有

$$p((x_i^{(n)},x_j^{(n)})|o_i\neq o_j)=p(x_i^{(n)})\times p(x_j^{(n)}) \tag{5}$$

将式(4)代入式(2)、式(5)代入式(3)，再将式(2)、式(3)代入式(1)，得到

$$\rho(x_i^{(n)},x_j^{(n)})=\frac{p(x_j^{(n)}|x_i^{(n)})\times p(o_i=o_j)}{p(x_j^{(n)})\times p(o_i\neq o_j)} \tag{6}$$

由于 $p(o_i=o_j)$，$p(o_i\neq o_j)$ 均为先验概率分布，并不依赖于两个对象的观测值，在基于具有相同分布的两个对象观测值进行对象辨识时，其比值为常数 C，因此，有

$$\rho(x_i^{(n)},x_j^{(n)})=C\times\frac{p(x_j^{(n)}|x_i^{(n)})}{p(x_j^{(n)})} \tag{7}$$

这里，如果直接使用 $\rho(x_i^{(n)},x_j^{(n)})$ 作为式(12)中的 $h(x_i^{(n)},x_j^{(n)})$ 项，则当 $\rho(x_i^{(n)},x_j^{(n)})$ 非常大时，此项会在相似度计算中占统治地位，会“淹没”其它项属性在相似度计算中的作用，而这在一般情况下是应避免的，因此，取它的对数 Sigmoid 变换，使其取值在 0 到 1 之间。则有

$$h(x_i^{(n)},x_j^{(n)})=\frac{1}{1+e^{-\ln\rho(x_i^{(n)},x_j^{(n)})}} \tag{8}$$

定义两个待识别对象 o_i，o_j 的第 n 个属性值相似度为 $\mathrm{sim}(o_i^{(n)},o_j^{(n)})$，则有

$$\mathrm{sim}(o_i^{(n)},o_j^{(n)})=\begin{cases}h(x_i^{(n)},x_j^{(n)}), & M<n\leqslant N\\ \text{（即对概率分布型属性）}\\ f(x_i^{(n)},x_j^{(n)}), & 1<n\leqslant M\\ \text{（即对确定型属性）}\end{cases} \tag{9}$$

$$f(x_i^{(n)},x_j^{(n)})=(\delta(x_i^{(n)},x_j^{(n)}))^{\alpha} \tag{10}$$

其中，$\delta(x_i^{(n)},x_j^{(n)})$ 为 $x_i^{(n)}$，$x_j^{(n)}$ 的近似度，其值小于等于 1，对于字符型向量，可采用字符串比较等方法获得。对于数字型向量值，可通过计算误差等获得，如，可取

$$\delta(x_i^{(n)},x_j^{(n)})=(1-\left|\frac{x_i^{(n)}-x_j^{(n)}}{|x_i^{(n)}|+|x_j^{(n)}|}\right|) \tag{11}$$

α 是为了反映 $x_i^{(n)}$，$x_j^{(n)}$ 之间近似的非线性特点而设置的惩罚系数，比如，当根据上式计算的 $\delta(x_i^{(n)},x_j^{(n)})$ 从 0.99 变为 0.9 时，许多情况下，两比较数据的相似性就从非常大变为非常小，这时它们之间近似度的比值显然不应该为 0.99 除以 0.9。α 可根据各个属性的情况取不同的数值，但大多数情况下应大于 1。

最后，定义 o_i，o_j 的相似度为

$$\mathrm{Sim}(o_i,o_j)=\prod_{n=1}^{M}f(x_i^{(n)},x_j^{(n)})\times(\sigma+\sum_{n=M+1}^{N}(w_n\times h(x_i^{(n)},x_j^{(n)}))) \tag{12}$$

式(12)中，各个确定型属性的相似度连乘，而各个概率分布型属性的相似度加权求和，是为了反映：在一般情况下，相较于概率分布型属性，确定型属性能更多地反映对象的固有特性，从而对两个待识别对象的同一性更具有决定权（特别是否决权）；对各个概率分布型属性加权，是为了反映不同概率分布型属性对对象本质的不同表征能力。式中，σ 是为了防止对象的各个概率分布型属性的 $h(x_i^{(n)},x_j^{(n)})$ 取 0 或趋近于 0 时 $\mathrm{Sim}(o_i,o_j)$ 得 0 的异常现象，一般取很小的正数值。

下面给出两点说明。

（1）模型利用条件

利用本文提出的相似度模型进行某类时变对象辨识时，需要知道该类对象各个属性的分布，然后才能利用式(7)－式(12)计算两个时变对象间的相似度。

（2）属性值不全时相似度的确定

在有些情况下，对象的属性值没有全部观测到，这时就需要确定没有观测到的属性的 $f(x_i^{(n)},x_j^{(n)})$ 或 $h(x_i^{(n)},x_j^{(n)})$。对于 $h(x_i^{(n)},x_j^{(n)})$，可直接赋值为 0，并将它的权重值按比例添加到其它概率分布型属性上，实际上就是该属性不参与两次观测对象相似度计算；但 $f(x_i^{(n)},x_j^{(n)})$ 则一般不宜直接赋值为 0，因为这会导致 $\mathrm{Sim}(o_i,o_j)=0$，从而容易造成对同一对象的漏认，但也不宜直接赋值为 1，使 $\mathrm{Sim}(o_i,o_j)$ 偏大，这又容易造成误认，这种情况下可采取先对 $f(x_i^{(n)},x_j^{(n)})$ 赋值为 1，并增大相应的 α 来提高对已知确定型属性值之间相似度的要求。

3 模型应用领域举例

本文提出的相似度计算模型可以应用到许多需要对象辨识的场合，如：

（1）对象搜索

在以对象为基本检索单位的搜索引擎中，可以利用模型对多个疑似时变对象与目标对象进行相似度计算，以提高搜索命中精度。

（2）对象数据集成、更新、剔除

数据质量是影响数据挖掘效果的关键因素之一。为提高被挖掘数据源的数据质量，数据清理变得很重要。在许多数据来源广泛的数据仓库中数据集成的一个重要问题是一些不同记录可能代表现实世界中的同一实体，即数据仓库中存在着大量的同一对象的不同时刻的变体，导致数据的重复、冗余、不一致等现象时有发生，因此相似重复数据的检测成为数据清理中的一个关键环节[3]。此时，对象的高精度辨识就可成为解决这些问题的重要手段之一，同时，随着同一目标对象自身的演变，其属性数据的高效准确的集成、更新也离不开高精度的对象辨识。比如，有时需要从 Web 上搜索集成某人的学

（下转第 273 页）

(1)首先读入一幅图像,并在无噪声图像中先加入了高斯噪声(均值为 0,方差为 0.01),并再加入椒盐噪声(强度为 0.01)进行实验。选取的滤波器大小为 5 * 5,小波包去噪域值为 35。原始图像及含噪图像如图 1 所示。

图 1 原始图像及含噪图像

(2)对图像进行小波包分解、计算最佳树,并对分解系数进行阈值量化,最后进行重构,再进行一次中值滤波得到去噪图像,如图 2 所示。

图 2 小波包第一次去噪后图像

图 3 中值滤波后的图像

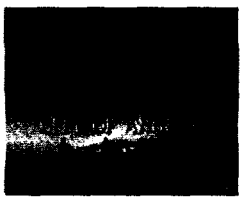

图 4 小波包第二次去噪后图像

图 5 最终去噪后图像

(3)图 2 由小波包去噪处理后,再进行一次中值滤波,能从整体上保留较多的细节,并且,椒盐噪声明显得以削弱。在对图 3 进行第二次小波包去噪后,得到图 4,我们看到小波包处理的图像在椒盐性噪声处有明显的过度性质的平滑削减,高斯噪声大幅减少,视觉效果明显。但经两次小波包去噪的图像清晰度明显减弱。所得图 4 再与图 3 进行小波融合,得到最终去噪图 5。图 5 比起融合前的二次小波包去噪图 4,背景图像明显更清晰。

结束语 综合上述分析可得如下结论:从含噪图像可以看出噪声含量非常强,而从去噪的结果可以看出,含混合噪声的图像,经小波包去噪,再对所得的图像进行小波融合,能得到去噪效果良好又能保持细节而且背景清晰的图像。因此,本文所提出的方法在图像含多种混合噪声的情况下是相当有效的。

参 考 文 献

[1] Ding Xing-hao, Deng Shan-xi, Yang Yong-yue. Image denoising based on wavelet packet[J]. Journal of Electronic Measurement and Instrument, 2003, 17(2): 35-39

[2] 小波分析理论与 MATLAB7 实现[M]. 北京:电子工业出版社, 2005

[3] 成礼智,王红霞,罗永. 小波的理论与应用[M]. 北京:科学出版社, 2004

[4] He H, Cheng S. Home network power—line communication signal processing based on wavelet packet analysis [J]. IEEE Transaction on Power Delivery, 2005, 20(3): 1879-1885

[5] Zhang N, Wu X. Lossless Compression of ColOr Mosaic Images [J]. IEEE Transaction on Image Processing, 2006, 15(16): 1379-1388

[6] Grossman A, Morlet J, Paul T. Transforms Assoeiated to Square Integeable Representation, II, Examples, Am. Inst. Henri Poineare, 1986, VDl. 45(3): 293-309

[7] Chen Q, Cheng Z, Wang C. Existence and construction of compactly supported biorthogonal multiple vector-valued wavelets [J]. Journal of Applied Mathematics and Computing, 2006, 22(3): 101-115

[8] 陈清江,程正兴,冯晓霞. 高维多重双正交小波包[J]. 应用数学, 2005, 18(3): 358-364

[9] Coffman R D. Wavelet and their application [J]. Wavelet and signal processing, 1992

(上接第 247 页)

术论文,若仅以姓名等不变性属性进行搜索,由于可能存在大量属性值相同(如,同名同姓)的人,会极大影响检索精度,而表征这个人的另外一些属性(如,体重、爱好等)又可能是随时间变化的,因此,要实现高精度检索该人,就需要进行含时变属性的多属性值相似度计算来提供辨识依据。

(3)对象分组

有时候,需要对观测到的大量不同时间序列的对象进行分组,此时也可以利用本文提出的相似度模型。比如说,已知各个属性有相同分布的初始对象 A,B,后来又观测到来自于这两个初始对象的若干时间序列变体后代,在需要把它们辨识为到底是 A 的变体还是 B 的变体的时候,可以利用式(7)一式(12)分别计算这些对象与 A,B 的相似度,然后对每一个待辨识对象,若其与原始对象 A 的相似度大于与原始对象 B 的相似度,则将其作为 A 的时变对象,反之,则将其作为 B 的时变对象。

结束语 为了实现更有效的检索、更有效的数据更新、重复数据剔除以及数据演变分析等,以对象为基本检索单位取代以页面为基本检索单位的 Web 对象搜索技术正逐渐得到有关学者的重视。由于大量的 Web 对象的部分属性具有时变性,要实现对象的高精度搜索,高精度的时变对象识别技术是重要的基础技术之一。在这个背景下,本文提出了一种时变对象辨识模型,该模型基于以下观点:①相当部分对象属性的演变服从某种分布;②对象的不同属性具有对对象本质不同的表征能力;③确定型属性通常能比概率分布型属性更多地反映对象的固有特性。

参 考 文 献

[1] Nie Zaiqing, Ma Yunxiao, Shi Shuming, et al. Web object retrieval[R]. MSR-TR-2006-70

[2] 小山聡,白砂健一,田中克己. 属性値が時間変化するオブジェクトを識別する確率モデル[C]// The 22nd Annual Conference of the Japanese Society for Artificial Intelligence, 2008

[3] 李星毅,包从剑,施化吉. 数据仓库中的相似重复记录检测方法[J]. 电子科技大学学报, 36(6): 1273-1277

[4] 张鹏,王国胤,陶春梅,等. 基于本体粗糙集的程序代码相似度度量方法[J]. 重庆邮电大学学报:自然科学版, 2008, 20(6): 737-741