基于视觉的网页重要变化检测方法

史存会[1,2]，俞晓明[1]，刘悦[1]，靳小龙[1,2]，程学旗[1,2]
（1.中国科学院网络数据科学与技术重点实验室，中国科学院计算技术研究所，北京 100190；
2. 中国科学院大学计算机科学与技术学院，北京 100049）

论文摘要

很多网页是动态变化的。通过检测网页重要变化，判断页面核心内容是否发生变化，可以有效降低数据采集中重复索引的数量，这对于很多互联网应用（如搜索引擎、变化检测与通知系统，以及互联网存档系统）的优化非常重要。本文提出了一种基于视觉特征的重要变化检测模型**VICD**，用以检测页面不同语义区域的变化，并可将页面压缩为一个低维向量进行表示。该方法从用户视觉的角度，理解页面不同区块语义重要度的差异。与现有方法相比，该方法独立于基于**HTML**类基础文档的分析方法，因此在新媒体，如移动互联网上，也有一定的适用性。通过在实验数据上的评估，验证了该方法的有效性。

关键词：Web内容；变化检测；视觉特征

There are always dynamic changes in Web pages. Therefore, we can effectively reduce duplicate Web indexes by detecting changes of essential content. This is very crucial for the optimization of many Web applications, such as search engines, change detection and notification system, and web archiving system. This paper proposes a model,VICD, based on visual features, which is used to detect changes in different semantic regions of the page and compress the page into a low dimensional vector representation. This method helps to understand importance of semantic in different regions from users' perspectives. Comparing with existing method, this method is independent of the analysis of HTML, which makes it also suitable for new media such as mobile Internet. Experiments show the effectiveness of the proposed method.

Key words: Web content; change detection; visual feature

论文简介

本文提出一种基于视觉特征的重要变化检测方法VICD(Vision based important Change Detection)，通过将同一页面在不同时刻的页面渲染后的图像作为输入，保留了页面的视觉特征，这与用户的感知输入保持一致，利用深度神经网络来模拟人的判定，从而有效的进行网页重要变化的检测。本文所提方法的优点在于：首先这是一种基于视觉特征的网页重要变化感知方法，能够利用视觉信息进行变化检测；其次该方法能够将原始图像进行压缩，形成一个低维表示，有利于页面比较，及互联网应用的优化；最后本文的方法具有很好的泛化能力，针对互联网不同网站对网页的设计差异性很大的特点，学习的模型可以用于检测从未见过的页面上的重要变化。

此外，随着移动互联网的快速发展，大量的信息产生于移动互联网。由于移动端的相对封闭性，以及与传统互联网的差异，导致对于移动互联网的检索分析还存在一定的困难。而本文提出的基于视觉特征分析的方法，对移动端的变化检测具有一定的适用性。

实验仿真

数据集

表2 VisualHtml数据集页面对应域名样例

dataset-popular	dataset-random
eu.real.com	digilander.libero.it
www.alistapart.com	www.animis.de
www.bing.com	ecoceco.com
www.ama-assn.org	www.cadixtour.com
www.break.com	www.elisanet.fi
www.fda.gov	www.greencove.fr

表3 VisualHtml数据集详情

	train	test
site	56	10
data-popular	17	5
data-random	39	5
page/site	11	11
pair_data	3080	550
triple_data	8400	_

对比方法

（1）SimHash： Manku等人[9]提出一种利用SimHash算法用于对网页不同部分进行签名形成指纹，并利用汉明距离进行变化判断，该算法用于近似重复网页的检测，并被应用到Google爬虫。

（2）VICD_2_EU：利用2分支的网络架构训练模型，并使用欧式距离进行变化判断。

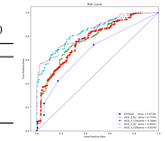
（3）VICD_2_COS：利用2分支的网络架构训练模型，并使用Cosine相似度进行变化判断。

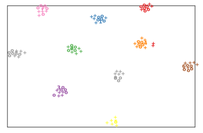
（4）VICD_3_EU：利用3分支的网络架构训练模型，并使用欧式距离进行变化判断。

（5）VICD_3_COS：利用3分支的网络架构训练模型，并使用Cosine相似度进行变化判断。

表4 不同模型在VisualHtml上的效果(AUC)

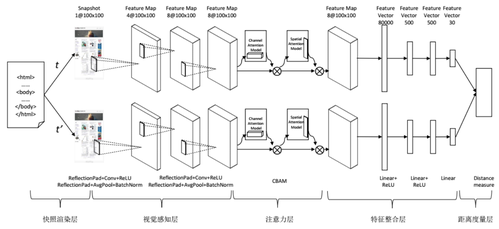
model	AUC
Simlash	0.6719
VICD_2_EU	0.7743
VICD_2_COS	0.7694
VICD_3_EU	**0.8441**
VICD_3_COS	0.8374





算法原理

基于视觉的重要变化检测模型VICD



快照渲染层

我们将页面快照大小缩放为100 × 100，将图像RGB模式转换为灰度图

视觉感知层

名称	输出维度	卷积核	步长
RefPad+Conv+ReLU	4 × 100 × 100	3 × 3	1
RefPad+AvgPool+Norm	4 × 100 × 100	3 × 3	1
RefPad+Conv+ReLU	8 × 100 × 100	3 × 3	1
RefPad+AvgPool+Norm	8 × 100 × 100	3 × 3	1
RefPad+Conv+ReLU	8 × 100 × 100	3 × 3	1
RefPad+AvgPool+Norm	8 × 100 × 100	3 × 3	1

注意力层

给定中间特征图，CBAM依次增加通道注意力和空间注意力：

$$F' = M_c(F) \otimes F,$$
$$F'' = M_s(F') \otimes F'$$
$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F)))$$
$$M_s(F) = \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)]))$$

特征整合层

在特征整合层，我们将得到的视觉特征向量v作为输入，然后输入到3层的前馈网络中，得到页面向量的低维表示v_p

$$V_p = W^2\sigma(W^1\sigma(W^0 v + b^0) + b^1) + b^2$$

距离度量层

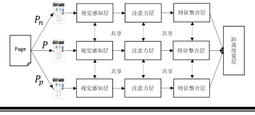
$$Sim(P_t, P_{t'}) = D\left(v_p^{P_t}, v_p^{P_{t'}}\right)$$

模型训练

对比损失函数(contrastive loss function)

$$L\left(W, Y, \vec{X_1}, \vec{X_2}\right) = (1-Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{max(0, m - D_W)\}^2$$

基于视觉的重要变化检测模型变形



模型训练

三元组损失函数(triplet loss function)

$$L(a, p, n) = \max(0, D(a, p) - D(a, n) + margin)$$

系统模型

基于视觉的网页重要变化检测模型

问题定义

定义1：一个Web页面是由不同的页面语义块组成，并且这些语义块具有一定的层级关系。
1级语义块：Header、Footer、Content、Sidebar
2级语义块：Logo, Menu, Title, Link-list, Table, Comments, Ad, Image, Video, Article, Searchbar

定义2：页面发生重要变化，当且仅当Content中的内容发生变化，无论其他语义块是否发生变化。

$$Sim(P_t, P_{t'}) = D(\emptyset(P_t), \emptyset(P_{t'}))$$

$$Change(P_t, P_{t'}) = \begin{cases} 0, & Sim(P_t, P_{t'}) < threshold \\ 1, & Sim(P_t, P_{t'}) \geq threshold \end{cases}$$

难点分析

问题的难点在于互联网具有多源异构的特性，其由各自独立设计的网站所组成。每个网站在样式设计上可能均不相同，我们不可能对每个网站，甚至每个网页设计函数$\emptyset$和D。针对互联网不同网站对网页的设计差异性很大的特点，如何利用模型对从未见过的页面进行变化检测，是本文的研究重点。

论文结论

检测网页重要内容的变化，对于很多互联网应用，如搜索引擎，变化检测与通知系统，以及互联网存档系统的优化非常重要。本文提出了一种基于视觉特征的重要变化检测模型，该模型可以用一个低维向量表示网页，并基于低维向量间的距离检测页面是否发生了重要变化。重要的是，模型可用于对从未见过的页面进行检测。此外，利用视觉特征将可以在新媒体上具有一定的适用性，如移动互联网。
从实验结果中可以看到，视觉特征可以有效提升对网页的理解。

在未来工作中，我们考虑将视觉特征应用到更多的网页分析任务上，如镜像网站识别、网页的语义分割与标注和信息抽取等。此外，本文将页面进行了相同大小处理，但实际上每个页面的实际大小可能均不同，未来我们考虑设计对于不同大小页面的处理方法。

中国科学院网络数据科学与技术重点实验室
Key Laboratory of Network Data Science & Technology ,CAS
中国科学院 CHINESE ACADEMY OF SCIENCES
西安电子科技大学 XIDIAN UNIVERSITY