基于网页上下文分析的图片检索

刘金松　于浩　西野文人
富士通研究开发中心，北京 100016
E-mail: jinsong@frdc.fujitsu.com

摘　要：基于网页上下文分析的图片检索是指利用 HTML 文档源代码，通过分析文档结构自动获取图片的说明，并以此创建图片索引的一种图片检索方法．在本篇论文中我们提出了一种能更加有效的创建图片索引的新方法．本方法在传统的计算图片与文本的距离的方法的基础上提出了利用识别出的主要文本块和重复图片块来提高说明文字提取精度，将图片说明分为个别图片说明和公共图片说明，并识别图片与 HTML 文档标题之间的联系的新设想．经过试验验证该方法能够显著提高系统性能，精度和召回率由原来的 57% 和 90%，提高到 86%和 95%．

关键词：图片检索，主要文本块，重复图片块，个别图片说明，公共图片说明，HTML 标题

Web-based Image Retrieval by Image Context Analyzing

Liu Jinsong, Yu Hao, Nishino Fumihito
Fujitsu R&D Center, Beijing 100016
E-mail: jinsong@frdc.fujitsu.com

Abstract: Web-based Image Retrieval is to retrieve image from WWW by text index automatically collected by Image Context analyzing. In this paper we propose a new method of extracting the explanation of Content Image on web page. Including the traditional method which recognizes the explanation of image mainly by distance, we try to get the explanation by recognition of Main Text Block and Repeating Image Block. With this method we can not only extract image' s explanation with higher precision, but also can extract the Common Explanation of image and find out the relationship between HTML Title and Content Image. By this method we improve the precision from 57% to 86% and recall from 90% to 95%.

Keyword: Web-based Image Retrieval, Content Image, Main Text Block, Repeating Image Block, Individual Explanation, Common Explanation Extraction, HTML Title.

1 简 介

随着网络技术的迅速发展，网上的图片信息越来越多。有大量的公司在网上发布各种版权图片如著名卡通形象，著名影星的照片等。但由于复制和发布非常容易，网上的盗版现象也越来越猖獗。这极大地损害了原创机构的利益，为了打击此类违法行为，维护公司的正当权益，强烈需要能够迅速追踪发现此类违法行为的软件工具。基于这样的应用背景我们开发完成了“非法图片检索系统”用于搜索网上的非法图片。它的核心技术是图片检索。自从上世纪 70 年代以来，图片检索一直是广大学者研究的热点。该领域技术发展经历了两个阶段，早期以 Anna Bjarnestan 为代表，通过人工为图片添加标注，然后以这些标注为索引来进行图片检索[1]。但是，对图片进行人工标注需要消耗大量的人力物力，成本太高，而且不同的人对同一张图片会有不同的理解，标注的结果很难统一，导致检索精度较低。同时随着网络的迅猛发展，图片数量的急剧上升，人工标注方法已逐渐被淘汰。取而代之的是上世纪 90 年代兴起的基于图片内容的检索。该方法根据图片自身的属性包括颜色、纹理、形状等特征来检索图片[2]。该方法主要优点在于可以自动创建图片索引，节省了人工标注的时间和成本，取得了很大的成功。但它也同样存在着检索精度低的致命弱点。

综合以上对图片检索技术的描述可以发现它们的共同缺点是没有充分利用网络技术迅速发展所带来的便利。事实上，网页中存在很多与图片相关的有用信息。网页图片随机调查[3]发现网页中 93%的图片有一个以上的说明，仅有 7%的图片没有对应的说明性文字。因此，近来越来越多的学者开始关注基于网页的图片检索。他们利用网页中的各种信息如 HTML 文档标题、图片的文件名、URL 地址、别名、链接[4、5、6]，并综合图片颜色、纹理、形状信息[7、8、9]来进行网页图片检索。这些尝试取得了很好的成果，很多商业图片检索系统也已经开发成功，比较著名的有 Google、WebSeer、AltaVista 等。

尽管如此，这些系统的性能还有待进一步提高。首先，现有的网络图片检索系统（以下简称现有系统）在提取图片说明信息时主要利用图片与文字的距离信息，具体来说就是统计图片与文字之间有多少个 HTML 标记，不同的标记有不同的赋值，将这些标记的赋值累加就得到图片与文字之间的距离。当距离小于某一阈值时就认为该文字是图片的说明，否则认为不是。这种算法过于简单，容易遗漏很多有用信息，导致检索性能不高。我们注意到网页中常常有一些大段的文字，这些文字描述了该网页的主要内容（本文称之为主要文本块）。或者网页中的多张图片与其对应的说明性文字以特定的格式循环往复地出现（本文称之为重复图片块）。如果在提取说明之前先行识别出主要文本块和主要图片块，那么对于提高说明的提取精度将非常有帮助。其次，显而易见 HTM 文档标题与其中的图片之间有某种联系。但标题一般只是与其中的部分图片而不是全部图片相关。但现有系统由于没有对网页结构进行详细分析无法判断图片与标题是否相关。在制作所索引时只好要么认为所有图片都与标题相关，要么认为都无关。这显然没有充分利用好标题信息。本系统通过识别页面的主要部分（主要文本块与重复图片块的合称），将位于主要部分中的图片设定为与标题相关，其他图片与标题无关。由此改善了系统的性能。第三，在包含有多个图片的网页中，各张图片除了有各自的说明之外，往往还有一些字段描述这些图片的共同主题。现有系统要么将这些字段设为某张图片的说明，要么遗漏这些说明性文字，导致图片的召回率大大下降。本系统通过对网

页主要部分的识别，可以将图片的说明划分为个别图片说明和公共图片说明两个类别。在保证检索精度的前提下大大提高了图片的召回率。经过以上改进，我们设计制作的非法图片检索系统的综合性能指标（F 值）由原来的 70%提高到 90%。

F 值 = （召回率*精度）*2/（召回率+精度）

2 基于网页的图片检索

非法图片检索系统的总体框架如图 1 所示。首先，由智能机器人自动从网上抓取网页存放到本地数据库。然后，由主要部分识别模块通过分析网页的 HTML 文档结构，识别描述页面主题的主要部分。网页的主要部分分为主要文本块和重复图片块两个类别。

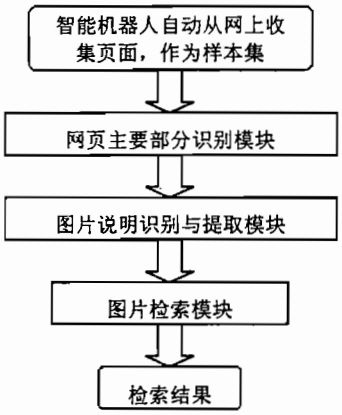

图 1：总体框架图

为了使所要表达的主题内容更加形象生动，增加读者的兴趣，一般在网页的主要部分中镶嵌有说明图片，用于辅助表达所描述主题内容。网页主要部分中的文字通常包含了对图片内容的说明，这些说明性文字是自动制作图片索引的关键。同时这些镶嵌在主要部分中的图片往往与页面的标题有一定的联系。这样页面的标题也可以作为这些图片的索引的一部分。下面用两张图为例对以上概念作一详细说明。这两张图片反映了两种比较典型的页面风格。图 2 是包含有主要文本块的页面样例。

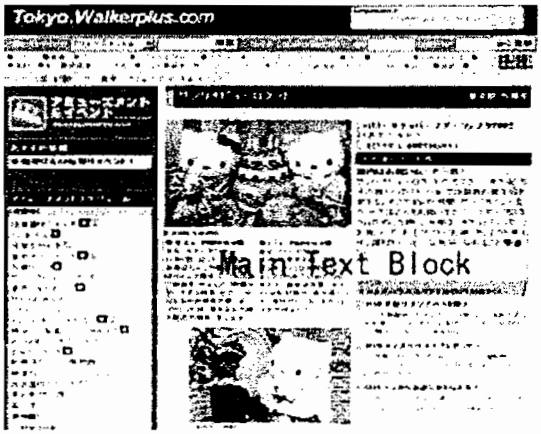

图 2：主要文本块

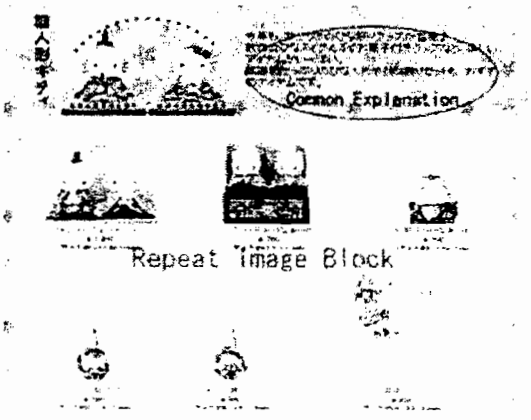

图 3：重复图片块

图中以黑色圆圈框出的部分就是该网页的主要文本块。在主要文本块中有两张日本著名卡通形象“凯蒂猫”的图片。主要文本块中的文字就是这两张图片的相关文字说明。图 3 是包含有重复图片块的页面样例。图中以黄色椭圆框出的部分就是一个重复图片块，它包含有 6 张图片及对应的图片说明（位于每张图片的正下方）。这里所指的重复并非是图片本身的重复，椭圆框中的图片显然各不相同，而是指 HTML 源代码中代码模式的循环往复。从外在表现上我们可以看到，这里第一行的图片对应于第二行的文字说明，第三行的图片对应于第四行的文字说明，第一行和第三行图片的排列方式是一样的。从原始的网页中我们可以看到，这两张网页中，页面的标题与这些图片是相关的。第三，在主要部分识别的基础上，系统调用图片说明识别与提取模块，识别并提取与图片对应的说明。说明字段包括图片别名、文件名、图片的 URL 地址、网页标题、个别图片说明、公共图片说明、图片周围文字 7 个部分。图片别名和文件名很容易提取，统计表明它们与图片内容的相关性非常高，但可惜的是有很多图片的命名不规范，仅有少量的图片有别名，因此单独使用文件名和别名作为索引的话，精度虽然可以很高，接近于 100%，但是召回率非常低大约不到 20%。图片的 URL 地址与图片内容有一定的相关性，但直接利用的话精度很低。网页标题的提取也非常容易，但与现有系统的不同之处在于我们所提取出的网页标题仅作为网页主要部分中包含的图片的说明的一部分，而不作为所有图片的一种说明。这样做可以在大幅提高召回率的同时保持精度基本不变。个别图片说明是指图片周围对单个图片内容进行描述的说明性文字。公共图片说明是指对页面中的多张图片的内容进行统一描述的说明性文字。个别图片说明和公共图片说明是网页正文中存在的内容图片的主要说明性文字，其正确提取与否直接影响到系统的性能指标。这两种说明的提取方法将在本文后续部分详细介绍。图片周围文字是指除了以上两种说明以外的与图片距离较近的文字。当个别图片说明和公共图片说明没有或者无法识别时，系统将使用传统方法根据图片与文本之间的距离提取说明文字。由于这样提取的说明文字与图片内容间的联系不是很密切，有时与图片内容非常相关，有时又没有任何关联，有时还会发生错误，为了区分起见，我们称之为图片周围文字。它对增加检索的召回率有一定帮助，当系统对召回率要求很高，而对精度并不苛求时可以考虑使用这种说明。第四，图片索引模块将根据图片说明的提取结果为图片创建索引。索引的格式如图 4 所示。<WebPage>和</WebPage>标记之间的是一个网页中所有图片的索引。

```
<WebPage>
    <Head> <URL>Web page's URL</URL>
            <LocalPath>Local Path of the page</LocalPath>
            <Title>Page's Title</Title> </Head>
    <Body> <Image>
            <FileName>Image's File Name</FileName>
            <ImageURL>Image's URL</ImageURL>
            <Size>Image's Size</Size>
            <Alt>Image's alternative field</Alt>
            <IndividualExplanation>Individual Explanation of the Image</IndividualExplanation>
            <CommonExplanation>Common Explanation of Images</CommonExplanation>
            <Surrounding>Other text near the Image</Surrounding>
            <IsInMainTextBlock>true/false<IsInMainTextBlock>
            <IsInRepeatingImageBlock>true/false</IsInRepeatingImageBlock></Image>
            <Image>Other Image</Image> </Body>
</WebPage>
```

图 4: XML Format Output

在<Head>部分记录该页面的总体信息，包括该网页的 URL 地址、本地存储为止、网页标

题。在<Body>部分分别记录各个图片的索引信息，包括图片的文件名、原始 URL 地址、图片大小、图片别名、个别图片说明、公共图片说明、图片周围文字、图片是否位于主要文本块中、图片是否位于重复图片块中。最后，根据图片的索引和输入的关键字进行计算，判断图片是否为目标图片并对检索到的结果进行排序，将结果输出给用户。

3 实验及评价

本项目的背景是从网上检索被非法使用的商业图片，如著名卡通形象或者著名影视明星的照片等。因此实验的样本集是与 Sanrio 公司的著名卡通形象凯蒂猫相关的网页。

3.1 样本集的制作

i) 网页收集；实验使用的样本集的收集方法如下。在 Google 的网页检索输入框中依次输入 kitty, 吉蒂猫，凯蒂猫三个关键词（这三个关键词对应于同一个卡通形象）。从检索结果的 URL 中依次下载排在前面的数百个页面作为样本集。ii)人工标注；为了便于进行人工标注提高效率，我们设计制作了标注工具。它能自动提取网页中的图片，简化标注过程。为了减少标注工作量，同时保持正确性的前提下，网页中存在的重复图片和细长条形图片被预先过滤。iii)标注结果统计；标注结果如表 1 所示。

表 1: 样本统计

网页总数	320 页
正解图片总数	594 张
非正解图片总数	1463 张
图片数合计	2057 张

3.2 试验结果及评价

经过实际测试实验结果如表 2 所示。试验结果表明，主要文本块和重复图片块的识别和提取对于基于网页的图片检索来说非常关键，它显著提高了个别图片说明的提取精度，为公共图片说明的提取创造了条件，并使得图片与网页标题之间的关系的识别成为可能。由此最终大大提高系统的性能指标，其结果甚至超出了最初的设计预期。

表 2: 试验结果

自动/人工	正解图片	非正解图片
正解图片	566	88
非正解图片	28	1375
合计	594	1463
召回律	精 度	F 值
95.3 %	86.5%	90.7%

4. 总 结

本文提出了在网页中分析和提取图片说明的一种方法。该方法通过识别网页中普遍存在的主要文本块和重复图片块显著提高了个别图片说明的提取精度，为提取公共图片说明创造了条件，并使得图片与网页标题的关系的判别成为可能。通过使用这些方法系统的综合性能指标由原来的 70％提高到 90％，其中精度从 57%提高到 86%，召回率从 90%提高到 95%.

但系统还有以下几个方面需要改进。i)包含关键词的非目标图片的识别；举个简单的例子，有一张宠物狗的图片旁边的说明是“总统的爱犬-斑点狗”，那么在检索“总统”时会把该小狗的图片错误地当成是“总统”被检索出来。这就需要进一步对说明的语义进行分析。ii)主要文本块的分割；有时主要文本块中会有多个子文本块分别描述对应的图片，此时就需要对主要文本块作进一步分割。这些是将来需要进一步探讨的课题。

参 考 文 献

[1] Anna Bjarnestam: “Text-based Hierarchical Image Classification and Retrieval of Stock Photography”, The Challenge of Image Retrieval Conference, February 25-26, 1999, Newcastle upon Tyne, UK

[2] Eakins, J P and Graham, M E: “Content-based image retrieval”, Report to JISC Technology Applications Programme, January 1999.

[3] Neil C. Rowe. 1999. Precise and Efficient Retrieval of Captioned Images, The MARIE Project.

[4] V. Harmandas, M. Sanderson, and M. D. Dunlop. 1997. Image retrieval by hypertext links. In the Proceedings of the 20th ACM SIGIR conference, Pages 296-303, 1997

[5] Neil C. Rowe and Brian Frew. 1998. Automatic Caption Localization for Photographs on World Wide Web Pages. The MARIE Project.

[6] Guojun Lu and Ben Williams. 1999. An Integrated WWW Image Retrieval System. AusWeb99- Proceedings

[7] La Cascia, M., S. Sethi, and S. Sclaroff. 1998. Combining Textual and Visual Cues for Content-Based Image Retrieval on the WorldWide Web. Proceedings. IEEE Workshop on Content-Based Access of Image and Video Libraries (Cat. No. 98EX173). IEEEComput. Soc Los Alamitos CA USA, 1998. viii+115.

[8] Thijs Westerveld. 2000, Image Retrieval: Content versus Context. Content-Based Multimedia Information Access, RIAO 2000 Conference Proceedings, (pp276-284) C. I. D. -C. A. S. I. S., Paris, France, 2000

[9] Rong Zhao and William I. Grosky. 2002, Narrowing the Semantic Gap Improved Text-Based Web Document Retrieval Using Visual Features , IEEE Transactions on Multimedia , 4(2), pp. 189-200, 2002.