



基于 Python 的图像搜索系统的设计与实现

齐亚莉, 张 磊

(北京印刷学院 信息与机电工程学院, 北京 102600)

摘 要: 信息技术和网络技术的飞速发展, 图像信息量激增, 人们对检索图像信息的需求也日益增加。为了从 Flickr 网站信息库中获取大量城市资料信息和文件数据, 采用 Web 挖掘技术流程, 利用 Python 语言, 通过 Flickr 网站提供的 API 服务, 设计了基于 Python 的图像搜索系统方案。运行结果表明, 该方法速度快, 图片信息的准确性高, 在没有带宽限制的情况下尽可能多地向服务器发送请求, 可更加快速地获取数据信息, 以备后续的应用需要。

关键词: Web 数据挖掘; 模式分析; 图像信息检索

中图分类号: G2, TP311.13

文献标识码: A

文章编号: 1004-8626(2010)02-0048-04

The System of Image Retrieval Based on Python

QI Ya-li, ZHANG Lei

(School of Information & Mechatronic Engineering, Beijing Institute of Graphic Communication, Beijing 102600, China)

Abstract: Based on the development of information technology and network, the amount of image information increase sharply. The image retrieval requirement augments increasingly. This paper proposes a system use the process of web data mining technology. The system gain a large scale of city image information based on Python image data mining system and the API service that Flickr website supplied. The experiment result shows it has good effectiveness on speed and accuracy. In the case of without bandwidth limitations it can send request as much as possible to sever and get data information faster for application of the follow-up.

Key words: web data mining; association analysis; image retrieval

随着 Internet 的飞速发展, 网络已成为全球最大的信息资料来源。搜索引擎 (如 Google Image Search)、图像目录 (如 WebSEEK) 和专类收藏 (如 Art Museum Image Consortium) 是可用于图像信息检索的 3 种途径^{[1]5}。其中, 搜索引擎要求用户提

出具体的、直接的、简练的检索需求; 图像目录则适用于用户不是很清楚自己所要的图像信息的情形, 按照事先编制好的目录或超链接等, 浏览自己需要查找的图像; 专类收藏借助搜索引擎和图像目录, 达到搜寻某些专题图像的目的。本研究项目设计的系统采用 Python 语言^[2-4]和 Flickr 网站提供的开放式应用程序设计界面 (API), 设计 MYSQL 数据库, 按照要求稳定地下载搜索到的大量图像。

1 Web 搜索流程设计

Web 上的信息是非结构化或半结构化的、动态的、并且是容易造成混淆的, 所以很难直接以 Web 网页上的数据进行数据挖掘, 必须经过必要的数据处理。查找资源是从目标 Web 文档中得到数据, 值得注意的是, 有时信息资源不仅限于在线 Web 文档, 还包括电子邮件、电子文档、新闻组、网站的日志数据, 甚至通过 Web 形成的交易数据库中的数据。信息选择和预处理是从取得的 Web 资源中剔除无用信息和将信息进行必要的整理^{[1]10}。例如, 从 Web 文档中自动去除广告连接、去除多余格式标记、自动识别段落或者字段并将数据组织成规整的逻辑形式或关系表; 在同一个站点内部或在多个站点之间进行模式分析, 验证、解释上一步骤产生的模式。

1.1 信息选择和预处理

信息选择和预处理是从取得的 Web 资源中剔除无用信息和对信息进行必要的整理。例如, 从 Web 文档中自动去除广告连接、去除多余格式标记、自动识别段落或者字段并将数据组织成规整的逻辑形式或关系表。

1) 相片来源 URL。知道了图片的 ID、服务器 ID 等后, 便可创建连接至相片的来源 URL。URL 采用以下格式:

`http://farm{farm-id}.static.flickr.com/{server-id}/{id}_{secret}.jpg or`

```
http://farm{farm-id}.static.flickr.com/{server-id}/{id}_{secret}_{mstb}.jpg or
```

```
http://farm{farm-id}.static.flickr.com/{server-id}/{id}_{o-secret}_{o}.(jpg|gif|png)
```

2) 后缀分类。字母后缀如下:s 小正方形 75dpi ×75dpi;t 缩略图,最长边为 100;m 小,最长边为 240;-中等,最长边为 500;b 大,最长边为 1024;o 原始图片,根据来源,格式可以是.jpg、.gif 或.png。

呼叫 flickr.people.getInfo 来寻找其自订 URL,只需简单地使用大多数相片 API 呼叫传回的使用者 ID,然后就可以轻松建立至个人档案、所有相片、个人相片或者影集的 URL:

```
http://www.flickr.com/people/{user-id}/-profile
```

```
http://www.flickr.com/photos/{user-id}/-photostream
```

```
http://www.flickr.com/photos/{user-id}/{photo-id}-individual photo
```

```
http://www.flickr.com/photos/{user-id}/sets-all photosets
```

```
http://www.flickr.com/photos/{user-id}/sets/{photoset-id}-single photoset
```

图1所示的文件有固定的文件格式,以便于Python程序可以成功解析其中的内容信息。文件第一行是此列表每项的名称,featureid表示这个特征的id号,这个id号是唯一的,不能与其他项重复;featuretext是表示此城市某一特征的关键词;orig_featuretext是表示这个城市特征原始名称的关键词;urlcity表示一个城市名称在数据挖掘系统里所能识别的字符串,一般是城市名称首字母小写状态;city表示城市名称的规范写法,首字母大写,后面字母小写的标准格式;citycode表示这个城市的代码,每个城市有唯一的citycode,citycode与featureid是一对多的关系,一个城市有很多特征标识,而一个特征标识只代表唯一的一个城市。

1.2 模式发现和模式分析

若要使用 Flickr API 自动进行模式发现,需要

```
featureid,featuretext,orig_featuretext,urlcity,city,citycode
132708-St. Bride Library-London-L001-
239285-Kensington Palace-London-L001-
205157-Green Line-London-L001-
132700-Krug Room-London-L001-
232687-Screen on the wall-London-L001-
13732-Backpacker-London-L001-
13730-Angel-London-L001-
336639-North End Road Market-London-L001-
3104026-Hepburn's-London-L001-
3130295-Perfume-London-L001-
3130294-Perfume-London-L001-
3130297-Bride's-London-L001-
```

图1 feature_list.txt 图

拥有一个应用程序 key。

首先注册成为 Flickr 用户,登录 Flickr 访问 <http://www.flickr.com/services/apikeys> 页面即可申请到免费的个人使用的 FlickrAPI Key 了。申请到 API key 后即可使用 Flickr API 提供的服务,进行模式自动发现。模式建立起来后,进行模式分析,验证、解释上一步骤产生的模式,可以是机器自动完成,也可以是与分析人员交互完成。挑选出符合要求的图片,例如,图片名称为这张图片的 photo id 加文件后缀“.jpg”,因为数据量非常大,所以分为 100 个文件夹存放,dir00 至 dir99。文件最后两位数的数值代表存在相对应编号的文件夹内,如 198965501.jpg 表示文件存放在 dir01 文件夹中,2241816488.jpg 表示文件存放在 dir88 文件夹中。

1.3 数据库设计

本系统使用 MySQL 数据库存储所有数据信息,需要 3 个数据表:flickrphotoset、flickrphototags、word_values。其中,flickrphotoset 与 flickrphototags 是两个模板表,表内容均为空。下载一个城市信息时,程序会调用 sql 语句“create table xxx like flickrphotoset”与“create table xxx like flickrphototags”,复制生成与城市名对应的 flickrphotoase 和 flickrphototags 表。命名规则为 flickrphotoset_e6_cityname、flickrphototags_e6_cityname。

flickrphotoset 表是图片信息表,包含 18 项字段,表中主要字段为:

1) Photoid,图片编号,此图片编号从 flickr 获取,该编号在 flickr 的数据库中是唯一的,即在本系统数据库中也是唯一值。此字段为数据库 flickrphotoset 表的主键对整条数据进行唯一标识。

2) Title,图片标题,是从 flickr 获取的主要信息,图片标题由发布者添加。在 flickr 网站上,title 属性是图片最概要的解释。

3) Descripton,图片描述信息,对图片内容做进一步描述说明,内容可以为空。

4) url.是通过数据挖掘后,存放在本地服务器发布的 URL 地址。在该系统工作流程中,此字段为空。

5) owner,图片拥有者,此项基本没有工作流程,均为空值。

6) postdate,图片发布到 flickr 网站时的日期记录信息。

7) tstamp,数据挖掘下载图片到本地,添加到数据库的日期。

8) photoldwn、photomdwn、photosdwn, 下载图片尺寸标识分别为 large、middle、small, 系统没有涉及到这三项字段, 都为空值。

9) photos_url, 服务器端此图片的真实 URL 地址, 即搜索并下载保存在本地的图片的地址。

10) license, 此字段与 userid 是联合主键。编号代表版权类型

11) infofetched, 系统没有涉及到该字段, 均为空值。

12) username, 用户名, 该用户在 flickr 注册的用户名。

13) takendate, 图片认证日期。

14) userid, 用户编号, 与用户名一对一关系。一个用户只有一个用户编号。可以通过 <http://www.flickr.com/photos/用户编号> 来访问该用户。

15) citycode, 城市代码, 此编码由 city 提供, 城市代码唯一。

16) withoutcity, 识别是否有城市标识, 有些图片内容与城市相符, 但在标签上没有标识出来, 称其为 nocity photos。

flickrphototags 表是图片标签表, 包含 3 项字段, 表中主要字段为:

1) tagid, 每张图片的标签, 对其进行编号, 此编号是自增主键。

2) photoid, 此图片编号与 flickrphotoset 表中的 photoid 一致, 通过 photoid, 可以找到对应的 tag 标签。

3) tag, 图片标签, 通过标签可以分析出此图片的真实内容。tag 与 photoid 是多对多关系, 一个 photoid 可以有多个标签, tag 也不是唯一的, 多个 photoid 可拥有同一个 tag。

word_values 表是收集的关键词列表, 包含以下 2 个字段:

1) word, 关键词或词组, 此字段为主键。

2) value, 给这个关键词打分, 如果关键词比较少见, 关键词的分数就比较高, 此字段功能是为后面评审打分、评定图片质量所用。

2 系统的实现

2.1 主流程的实现

Web 挖掘流程需要 4 个阶段: 查找资源、信息选择和预处理、模式发现、模式分析。本系统的代码实现其中的信息选择和预处理、模式发现这两个环节的程序编写。本系统的主程序文件为 dig-

万方数据

ger3.py, 主要流程是: 读取配置函数变量, 读取文件的数据信息; 分析、处理读取的数据信息; 连接数据库; 将数据信息放入线程中运行直至结束。需要说明的是, 该搜索系统包含 8 种搜索方式, 在实际操作中, 8 种搜索方式都需要逐一完成, 这 8 种方式为:

1) tags full feature name+city, 标签搜索特征单词+城市名。

2) text full feature name+city, 文本搜索特征单词+城市名。

3) tags full feature name, 标签搜索特征单词。

4) text full feature name, 文本搜索特征单词。

5) tags rarest word of feature+city, 标签搜索稀少单词+城市。

6) text rarest word of feature+city, 文本搜索稀少单词+城市。

7) any tags full+city, 标签搜索城市名。

8) any text full+city, 文本搜索城市名。

2.2 系统的实现

本系统界面在命令行界面下执行, 在 Windows 开始菜单中选择“运行”, 输入 cmd 后, 进入命令提示界面, 执行系统主程序 digger3.py 后面加入的参数是城市名称 (见图 2~图 6)。

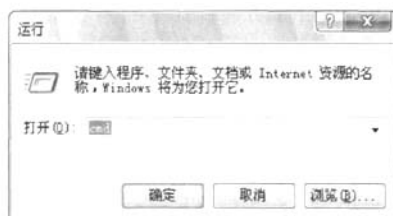


图 2 Windows 运行

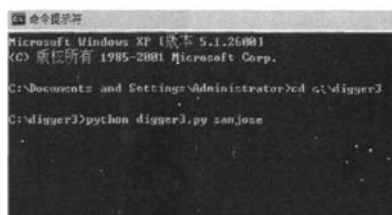


图 3 cmd 命令行

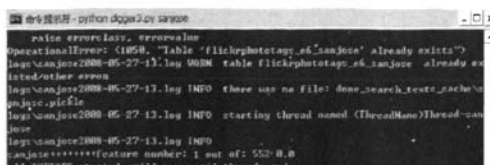


图 4 主程序运行界面



图5 下载后的文件夹分类

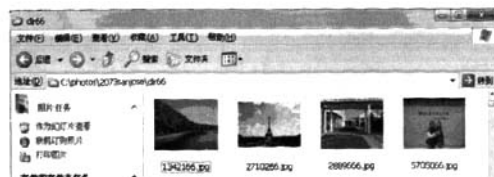


图6 下载的缩略图

3 结 语

采用Python语言和Flickr网站提供的开放式应用程序设计界面(API),设计MySQL数据库,系统能准确、快速、大批量地收集到图像信息。系统基于Python进行大数据量的数据结构存储,可以

通过flickrAPI稳定地向服务器发送大量请求下载数据,系统目标是从Internet进行图像数据挖掘。应用结果表明,该系统速度快,图片信息的准确性高,在没有带宽限制的情况下,系统可尽可能多地向服务器发送请求,更加快速地获取丰富的数据信息,以备后续的应用需要。

参考文献:

- [1] 储荷婷,敖毅. 怎样在Internet上查找图像[M]. 辽宁科学技术出版社,2007:5-10.
- [2] Allen B. Downey. Think Python: How to Think Like a Computer Scientist [EB/OL]. (2009-03-16) [2009-05-25]. <http://www.greenteapress.com/thinkpythonhtmlindex.html>.
- [3] Python 从新手到专家[EB/OL]. (2007-06-10) [2008-05-25]. <http://www.woodpecker.org.cn/diveintopython/>.
- [4] Guido van Rossum, Fred L. Drake. Python 手册[EB/OL]. (2003-07-29) [2009-05-25]. <http://doc.chinahtml.com/Manual/Python/tut/>.

(责任编辑:邱林华)

(上接第41页)

$$E(x, y) = (x, y)M(x, y)^T \quad (16)$$

这里, M 是一个 2×2 的对称阵, 即: $M = \begin{bmatrix} A & C \\ C & B \end{bmatrix}$,

可以看出, E 是和局部自相关函数相关的, 设 α, β 是 M 的特征值, 则 α, β 将和自相关函数的主曲率成比例。通过考察自相关函数的曲率来决定角点的存在。如果两个曲率都高, 这时, 对应的局部自相关函数的峰值达到最高点, 在任何方向上的移动都会使 E 增加, 这就说明这是角点。

4 结 语

角点作为图像上的特征点, 包含有重要的信息, 在图像融合和目标跟踪及三维重建中有重要的应用价值。基于实际应用需求, 从角点检测的快速性、准确性、鲁棒性等方面看, 上述各种角点检测算法各有利弊。近年来, 随着数学理论的完善和发展, 利用小波分析和数学形态学的角点检测方法已经成为人们研究的热点。本文较为详细地分析了

角点检测的几个有代表性的方法, 回顾了各种处理方法的发展现状, 希望能对相关领域的科研人员和工程技术人员有所帮助。

参考文献:

- [1] ROSENFELD, A. JOHNSTON, E. ANGLE. Detection on digital curves[J]. IEEE Trans. Comput, 1973(22):875-878.
- [2] FREEMAN H., DAVIS L. S. A corner-finding algorithm for chain-coded curves[J]. IEEE Trans. Comput, 1977(26):297-303.
- [3] CHI-HAO YEH. Wavelet-based corner detection using eigenvectors of covariance matrices[J]. Patter Recognition Letters. 2003(24):2797-2806.
- [4] 任雁. 角点检测方法研究[J]. 机械工程与自动化, 2009, 1(2):198-200.
- [5] CHEN Bai-fan, CAI Zi-xing. Harris corner detector based multi-scale space theory[J]. Journal of Central South University (Science), 2005, 36(5):751-754.
- [6] 闫龙, 赵正旭, 周以齐. 图像质量对Harris角点检测的影响研究[J]. 山东大学学报, 2006, 36(5):21-24.

(责任编辑:邱林华)

作者: [齐亚莉, 张磊](#)
作者单位: [北京印刷学院, 信息与机电工程学院, 北京, 102600](#)
刊名: [北京印刷学院学报](#)
英文刊名: [JOURNAL OF BEIJING INSTITUTE OF GRAPHIC COMMUNICATION](#)
年, 卷(期): 2010, 18(2)

参考文献(4条)

1. [Guido van Rossum; Fred L. Drake Python 手册](#) 2009
2. [Python 从新手到专家](#) 2008
3. [Allen B. Downey Think Python: How to Think Like a Computer Scientist](#) 2009
4. [储荷婷; 敖毅 怎样在互联网上查找图像](#) 2007

本文链接: http://d.g.wanfangdata.com.cn/Periodical_bjysxyxb201002013.aspx