

本科毕业论文



|  |  |
| --- | --- |
| 论文题目： | 基于隐模型的双视图财经数据聚类和分析 |
| 院 系： | 计算机科学技术学院 |
| 专 业： | 计算机科学与技术 |
| 姓 名： | 朱恬骅 |
| 学 号： | 09300240004 |
| 指导教师： | 黄萱菁 池明旻 |

2013年 6月21日

**摘要**

摘要……

**关键词：**多视角，视觉特征，文本特征，三分图，等周问题

**Abstract**

Abstract…

**Keywords:** ….

目录

[第一章 引言 1](#_Toc225579808)

[第二章 图模型简介 3](#_Toc225579811)

[2.1 齐次图 3](#_Toc225579812)

[2.2 二分图 4](#_Toc225579813)

[2.3 三分图 6](#_Toc225579813)

第三章 底层特征提取………………………………………………………………..8

3.1 文本特征提取……………………………………………………………….8

3.2 视觉特征提取……………………………………………………………….9

3.2.1 图像全局特征…………………………………………………………9

3.2.2 SIFT特征……………………………………………………………..10

3.2.3 SURF特征……………………………………………………………11

3.2.4 Bag-of-visual-word算法……………………………………………...11

第四章 多视角聚类算法……………………………………………………………13

4.1 二分图的等周算法………………………………………………………...13

4.1.1 算法推导……………………………………………………………..13

4.1.2 系数矩阵转化………………………………………………………..16

4.1.3 阈值选择……………………………………………………………..16

4.2 三分图的等周算法………………………………………………………...17

4.2.1 原始模型……………………………………………………………..17

4.2.2 带权重模型…………………………………………………………..18

4.2.3 最终模型……………………………………………………………..18

4.3 算法总结…………………………………………………………………...20

4.4 算法复杂度分析…………………………………………………………...20

第五章 测试和实验结果……………………………………………………………22

5.1 数据准备…………………………………………………………………...22

5.2 聚类结果评判方法………………………………………………………...25

5.3 图像聚类实验结果………………………………………………………...25

5.3.1 多视角算法必要性…………………………………………………..25

5.3.2 SIFT特征……………………………………………………………..29

5.3.3 SURF特征……………………………………………………………31

5.4 图像聚类算法比较………………………………………………………...32

5.5 计算速度…………………………………………………………………...33

第六章 总结及展望…………………………………………………………………34

[参考文献 35](#_Toc225579814)

[致谢 37](#_Toc225579815)

# 第一章 引言

在过去的十多年里，互联网发生了翻天覆地的变化。

# 第二章 图模型简介

## 鉴于本文所使用的方法都是基于图模型理论的，在本章中我们将对后面所使用到的图模型概念作一些简单的定义和介绍。

## 2.1 齐次图

一个。

**定义1：**图的邻接矩阵A可以定义为：

= （1）

。

## 2.2二分图

一集。

## 2.3三分图

通过。

# 第三章 底层特征提取

## 第二章．

## 3.1文本特征提取

网络图片与普通图片的很大不同就在于，前者的周围通常都会存在一些描述性的文字，而这些描述性的文字，可以帮助我们解决图像特征在语义上所不能解决的一些问题。对于这些文本信息的描述，我们采用了TF-IDF的词频统计方法，下面就来介绍一下TF-IDF算法。

TF-IDF（term frequency–inverse document frequency）是一种用于资讯检索与资讯探勘的常用加权技术。它一般用来评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

TFIDF实际上是：TF \* IDF，TF是词频(Term Frequency)，IDF是反文档频率(Inverse Document Frequency)。TF-IDF的主要思想是：如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。

在一份给定的文件里，词频指的是某一个给定的词语在该文件中出现的次数，。这个数字通常会被归一化，以防止它偏向长的文件。

 （12）

其中，是该词在文件中出现的次数，而分母则是在文件中所有字词的出现次数总和。

逆向文件频率是一个词语普遍重要性的度量。某一个词语的IDF，可以由总文件数目除以包含该词语的文件数目，再将得到的商去对数。

 （13）

其中，|D|表示语料库中的文件中数，而表示语料库中包含字词的文件数。

最后将TF和IDF相乘得到最终的词频统计结果：

 （14）

一个字词拥有较高的TF-IDF权重表示它在文件中出现的次数多，并且出现该字词的文件数在整个语料库中比较少，所以该权重能够有效地过滤掉那些常用的字词。

## 3.2视觉特征提取

聚类。

**3.2.1 图像全局特征**

聚类。

**3.2.2 SIFT特征**

特点如下：

**3.2.3 SURF特征**

下降。

**3.2.4 Bag-of-visual-words算法**

效果。

# 第四章 多视角聚类算法

与第二章相同。

## 4.1二分图的等周算法

方法。

**4.1.1 算法推导**

如果是这样，

**4.1.2 系数矩阵转化**

对于前面得到。

**4.1.3 阈值选择**

对方程。

## 4.2三分图的等周算法

**4.2.1 原始模型**

在介绍。

**4.2.2 带权重模型**

为了。

**4.2.3 最终模型**

为了

。

## 4.3算法总结

本文

## 4.4 算法复杂度分析

本文。

# 第五章 测试和实验结果

## 5.1数据准备

为了。

## 5.2 聚类结果评判方法

为了。

## 5.3 图像聚类实验结果

**5.3.1多视角算法必要性**

首先。

**5.3.2 SIFT特征**

这次。

## 5.4 图像聚类算法比较

最后。

## 5.5 运算速度估计

现在。

# 第六章 总结及展望

此外，虽然我们通过同时利用视觉和文本特征可以在一定程度上提高聚类的准确率，但是，我们离完全正确地进行图片聚类还有一定距离。对于那些图像视觉特征难以分辨、文本特征又非常稀缺的网络图片，我们可以考虑使用网络图片数据集中关于链接的信息，这些信息可以有效地帮我们解决语义上的问题，填补缺少文字信息所造成的不足，得到更好的聚类结果。

**参考文献**

[1] S. Gordon, H. Greenspan, and J. Goldberger. Applying the information bottleneck principle to unsupervised clustering of discrete and continuous image representations. In proc.of IEEE ICCV, 2003.

[2] G. Qiu. Image and feature co-clustering. In proc. of IEEE ICPR, 2004.

[3] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. ACM Comput. Surv., 31(3):264–323, 1999.

[4] M. Cascia, S. Sethi, and S. Sclaroff. Combining textual and visual cues for content-based image retrieval on the world wide web. In proc. of IEEE CBAIVL, 1998.

[5] S. Deerwester, S. T. Dumais, G.W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6):391–407, 1990.

[6] R. Zhao and W. I. Grosky. Narrowing the semantic gap-improved text-based web document retrieval using visual features. IEEE Trans. on Multimedia, 4(2):189–200, 2002.

[7] D. Cai, X. He, Z. Li,W.-Y. Ma, and J.-R.Wen. Hierarchical clustering of www image search results using visual, textual and link information. In proc. of ACM Multimedia, 2004

[8] Z. Li, G. Xu, M. Li, W.-Y. Ma, and H.-J. Zhang. Grouping www image search results by novel inhomogeneous clustering method. In proc. of MMM, 2005.

[9] B. Gao, T.-Y. Liu, T. Qin, X. Zheng, Q.-S. Cheng, and W.-Y. Ma. Web image clustering by consistent utilization of visual features and surrounding texts. In proc. of ACM Multimedia, 2005.

[10] M. Rege, M. Dong, J. Hua. Graph theoretical framework for simultaneously integrating visual and textual features for efficient web image clustering. In proc. of WWW, 2008

[11] S. Lazebnik, C. Schmid, J. Ponce. Beyond bag of features: spatial pyramid matching for recognizing natural scene categories. In proc. of CVPR, 2006

[12] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In Proc. CVPR, 2005.

[13] David G. Lowe, "Object recognition from local scale-invariant features," *International Conference on Computer Vision,* Corfu, Greece (September 1999), pp. 1150-1157.

[14] J. Dodziuk, “Difference equations, isoperimetric inequality and the transience of certain random walks,” Trans. of the American Mathematical Society, vol. 284, 1984.

[15] J. Dodziuk and W. S. Kendall, “Combinatorial laplacians and isoperimetric inequality,” From Local Times to Global Geometry, Control and Physics of Pitman Research Notes in Mathematics Series, Longman Scientific and Techical, vol. 150, pp. 68–74, 1986.

[16] B. Mohar, “Isoperimetric numbers of graphs,” Journal of Combinatorial Theory, Series B, vol. 47, pp. 274–291, 1989.

[17] L. Grady and E. L. Schwartz, “Isoperimetric graph partitioning for image segmentation,” IEEE Trans. on PAMI, vol. 28, no. 3, pp. 469– 475, 2006.

[18] M. Rege, M. Dong and F. Fotouhi. Co-clustering documents and words using Bipartite Isoperimetric Graph Partitioning. In proc. of ICDM, 2006

[19] L. Hagen and A. B. Kahng, “New spectral methods for ratio cut partitioning and clustering,” IEEE Trans. on CAD of Integrated Circuits and Systems, vol. 11, no. 9, 1992

[20] C. L. Lawson and R. J. Hanson. Solving Least Squares Problems. Soc for Industrial and Applied Math, 1995

[21] [H. Bay](http://en.wikipedia.org/wiki/Herbert_Bay), A. Ess, T. Tuytelaars, L. V. Gool. ["SURF: Speeded Up Robust Features"](http://www.vision.ee.ethz.ch/~surf/papers.html), Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346--359, 2008

[22] Michael Grubinger, Clement H. C. Leung: A Benchmark for Performance Calibration in Visual Information Search. Proceedings of The 2003 International Conference on Visual Information Systems (VIS 2003), pages 414 - 419, Miami, FL, USA, September 2003. Knowledge Systems Institute.

# 致谢

在本文写作的各阶段，我都得到了许多人帮助。在本文的最后，谨敬上我对他们的微薄意与诚挚感谢。

经过半年多的紧张工作，这篇凝结了多方汗水的论文终于得以完成。在此之际，我要特别感谢指导老师池明旻老师，为这篇论文的开篇立意，框架指导，写作过程，最终定稿付出了辛勤劳动。她对于论文的每一稿都会认真细致地修改，直到满意为止，闪烁着严谨的治学精神。感谢鲍江峰、刘隽、刘文昭、陈峰科学长，邹杨修、赵林佳等同学的帮助。感谢陈一星学姐长久来的鼓励，理解的目光都推动着我去向未知的远方。

最后，感谢父母的养育之恩，感谢他们无怨无悔的付出和对我将来规划的支持。谨以此文献给他们。