北京大学
PEKING UNIVERSITY

中文电子文档的数学公式定位研究

林晓燕 高良才 汤帜

{linxiaoyan | glc | tangzhi}@pku.edu.cn

北京大学计算机科学技术研究所

目录

- 背景
- 相关工作
- 数学公式定位
- 实验结果
- 结论

目录

- *背景*
- 相关工作
- 数学公式定位
- 实验结果
- 结论

背景

- 教育、科技：电子文档
- 数学公式特点
  - 特殊编码、二维结构
- 需求
  - 显示：移动阅读，流式重排
  - 重用：教育，教材资源重用
  - 检索：数字图书馆，学术搜索
- 公式定位
  - 自动检测文档中（独立/内嵌）公式所在区域

目录

- 背景
- *相关工作*
- 数学公式定位
- 实验结果
- 结论

相关工作 – 基于规则方法

- 内容规则
  - Kacem[4]: 数学符号的作用域对周围的字符进行扩展。
  - Suzuki[5]: 传统OCR对数学符号“拒识”。

- 布局规则
  - Chang[8]: 文本行投影特征
  - Garain[9]: 行高、行间距、像素分布等，独立公式；
    语义N元模型，包含内嵌公式的行；
    分词，词的排版风格、间距，内嵌公式。

相关工作 – 基于规则方法

- 问题
  - 数学符号、中文字符混排
    - ➔ 基于运算符作用域扩展的规则失效
  - 不同的文档类型、排版风格、布局
    - ➔ 阈值设定困难

相关工作 – 基于学习的方法

- 思路
  - 文本行/词为分类单元
  - 提取特征，训练，分类
- 主要方法
  - Drake[11]: Voronoi图，连通分支邻接图；
    节点/边的几何特征。
  - Liu[12]: SVM & CRF，稀疏文本行分类 + 规则。

相关工作 – 基于学习的方法

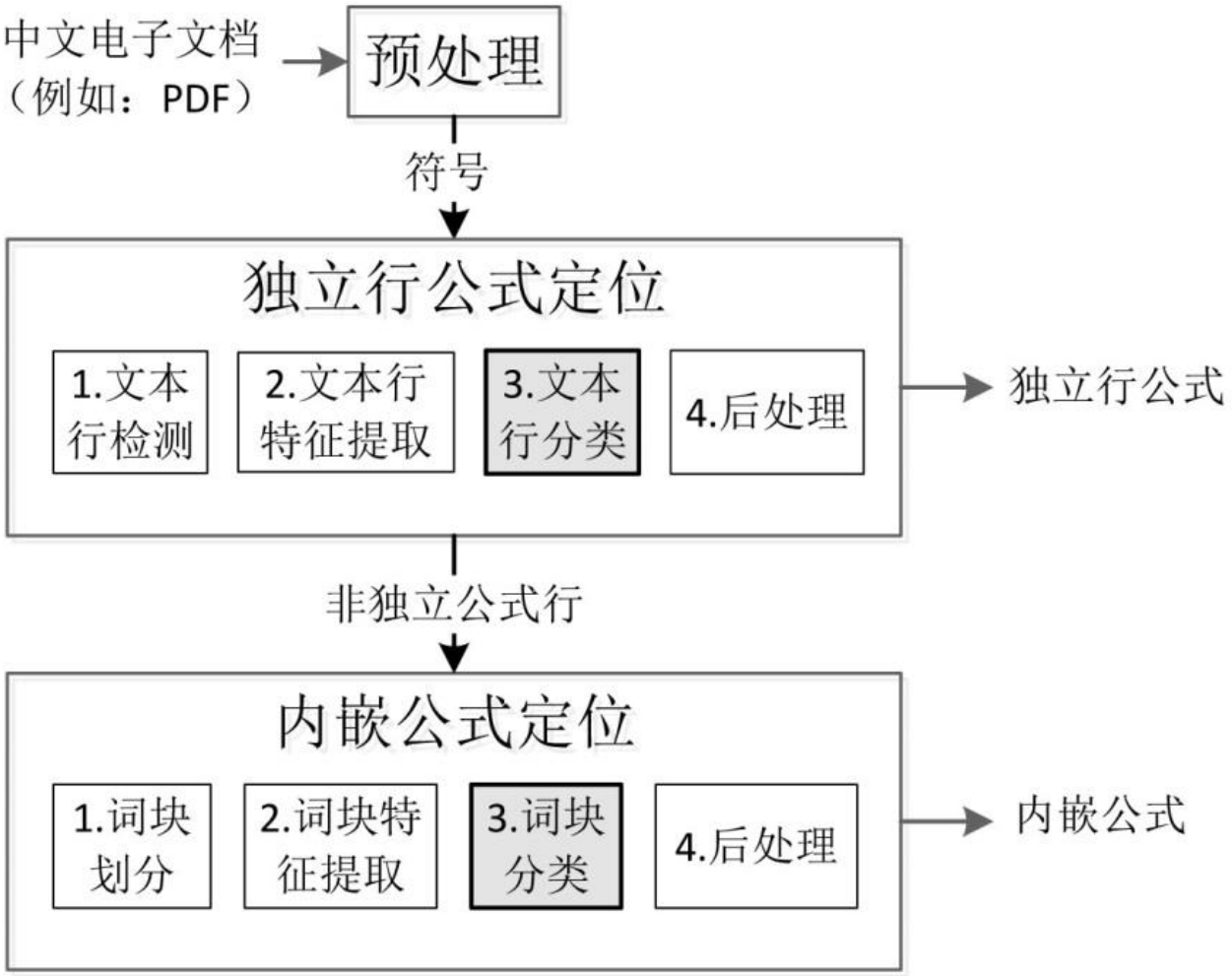
- 问题
  - 西文单词与中文词定义上的差异
  - 文本行/词在内容、布局上的差异
  - 公式行被过度分割 ➔ 部分定位

中文文档公式定位的分类单元、特征、后处理？

目录

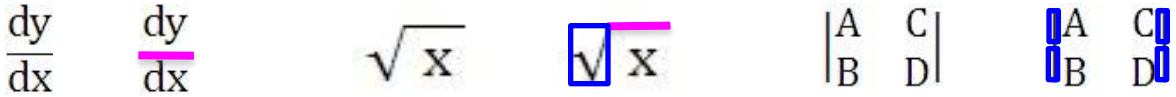
- 背景
- 相关工作
- *数学公式定位*
- 实验结果
- 结论

总体流程


中文电子文档
（例如：PDF）
预处理
符号
独立行公式定位
1.文本行检测
2.文本行特征提取
3.文本行分类
4.后处理
独立行公式
非独立公式行
内嵌公式定位
1.词块划分
2.词块特征提取
3.词块分类
4.后处理
内嵌公式

预处理

- 目标：
  - 提取电子文档的数学符号及其属性
- 方法：
  - 大部分数学符号：解析电子文档，文字对象
  - 复合符号：构建文字、图形、图像对象组合规则



- 效果：
  - 符号信息：边框位置、编码、基线、字体等

独立公式定位 – 1.文本行检测

- X-Y切割
- 过分割：文本行合并策略
  - 非纯中文文本行 & 水平交叠
  - 高度比、分数线

$$V = \frac{2r^2(\rho_1 - \rho_2)g}{g\eta}$$

式中：V 为微粒沉降速度，cm/s。

$$ROV_2 = \sum_{t=}^{5} AR_t / AR_0 \qquad (3-12)$$

计算了这两组股票的过度反应程度，分别为－30.56%%和

$$V = \frac{2r^2(\rho_1 - \rho_2)g}{g\eta}$$

式中：V 为微粒沉降速度，cm/s。

$$ROV_2 = \sum_{t=1}^{5} AR_t / AR_0 \qquad (3-12)$$

计算了这两组股票的过度反应程度，分别为－30.56%%和

独立公式定位 – 2. 特征提取

- 文本行特征

类别	特征名称	特征定义
布局特征	居中	行中心与分栏中心的相对距离。
	左空白/右空白	行左/右方空白（相对分栏）与分栏宽度比值。
	上间距/下间距	行与其上/下邻接行的间距与主体行间距的比值。
	行高	行高与页面主体中文字高度的比值。
	稀疏度	行内字符总面积与行面积的比值。
	字号/基线/高度方差	行内字符的字体大小、基线位置、高度的方差。
	Delaunay角度均值	对行内所有字符中心点构成的点集，使用算法[18]构建三角剖分（Delaunay triangulation），计算其中不跨越其他字符的边与水平方向锐角夹角的均值。
内容特征	数学符号比例	行内数字符号（预定义的数学函数、希腊字符、操作符等）比例。
	中文字符比例	行内中文字符比例。
	行末公式编号	行末是否包含常见公式编号。

独立公式定位 – 3&4 分类、后处理

- 文本行分类
  - 支持向量机（SVM）、装袋（Bagging）
- 后处理
  - 连续文本行
  - 文本行以二元运算符（如：+，=，≤等）结束或开始

而按主要设备的造价作为一次投资的组成部分来进行比较，也是允许的。故总投资

$$K = K_b + K_P + K_T$$

$$= K_b + (n_1 \cdot K_{P1} + n_2 \cdot K_{P2} + \cdots) + K_T \qquad (3.19)$$

K_b、K_T分别为变压器投资、占地及土石方投资；K_P为配电装置投资；K_{P1}、K_{P2}…分

而按主要设备的造价作为一次投资的组成部分来进行比较，也是允许的。故总投资

$$K = K_b + K_P + K_T$$

$$= K_b + (n_1 \cdot K_{P1} + n_2 \cdot K_{P2} + \cdots) + K_T \qquad (3.19)$$

K_b、K_T分别为变压器投资、占地及土石方投资；K_P为配电装置投资；K_{P1}、K_{P2}…分

内嵌公式定位 – 1. 词块划分

- 西文分词
  - 字符间距、标点
- 本文分词方法
  - 遇到分隔符，得到一个词块，分隔符自身也形成一个词块
  - 分隔符1：与主体字字体大小相同的中文字符
  - 分隔符2：标点符号

内嵌公式定位 – 1. 词块划分

式中：d——单位工作量折旧额；

W——预计使用期限内可以完成的工作量；

ω——年实际完成工作量。

[例 2－2] 某机床原值为 200 000 元，预计净残值为 50 000 元，规定可使用 12 000 个工作小时，每年实际使用 1 000 个工作小时，试计算年折旧额。

解：已知 $V_K = 200\ 000$ 元，$V_L = 50\ 000$ 元，$W = 12\ 000$ 工作小时，则：

式中：d——单位工作量折旧额；

W——预计使用期限内可以完成的工作量；

ω——年实际完成工作量。

[例 2－2] 某机床原值为 200 000 元，预计净残值为 50 000 元，规定可使用 12 000 个工作小时，每年实际使用 1 000 个工作小时，试计算年折旧额。

解：已知 $V_K = 200\ 000$ 元，$V_L = 50\ 000$ 元，$W = 12\ 000$ 工作小时，则：

内嵌公式定位 – 2. 特征提取

	名称	定义
布局特征	词高比	词块高度与页面主体字高度的比值。
	宽高比	词块宽度与高度的比值。
	字体大小/基线/间距/宽度/高度方差	词块内字符的字体大小、基线位置、间距、宽度、高度方差。
	稀疏度	词块内字符面积和与词块面积的比例。
	Delaunay角度均值	以词块内所有字符中心点为点集，计算方法与表1中Delaunay角度均值相同。
内容特征	中文字符比例	词块内中文字符比例。
	数学符号比例	词块内的数字符号（预定义数学函数、希腊字符、操作符等）比例。
	纯度	词块内包含同种字符类型（数学符号、中文符号、拉丁符号）的程度。
	左边缘/右边缘字符类型	词块内部左/右边缘字符类型（中文、一般西文、数学变量、一元数学符号、二元数学符号）。
上下文特征	前一词块最右字符类型	与当前词块相邻的前一词块的最右边缘字符类型。
	后一词块最左字符类型	与当前词块相邻的后一词块的最左边缘字符类型。

内嵌公式定位 – 3&4 分类、后处理

- 词块分类
  - 支持向量机（SVM）、装袋（Bagging）
- 后处理
  - 标点符号组成的连续词块聚合成新词块
  - 前后词块标签 ➔ 是否合并

其中，a_i、b_i 是随资产不同而变化的常数，$M \sim S(\alpha, 0, 1, 0)$ 是影响所有资产的市场因子，$\varepsilon_i \sim S(\alpha, 0, \gamma_i, 0)$ 是资产 i 的扰动，相互独立且独立于 M。

在 (4—12) 下，N 种资产的收益 $R = (R_1, \cdots, R_N)'$ 有一个

其中，a_i、b_i 是随资产不同而变化的常数，$M \sim S(\alpha, 0, 1, 0)$ 是影响所有资产的市场因子，$\varepsilon_i \sim S(\alpha, 0, \gamma_i, 0)$ 是资产 i 的扰动，相互独立且独立于 M。

在 (4—12) 下，N 种资产的收益 $R = (R_1, \cdots, R_N)'$ 有一个

目录

- 背景
- 相关工作
- 数学公式定位
- *实验结果*
- 结论

实验结果

- 数据集
  - 24本中文电子书中的200页文档
  - 1166个独立公式，3022个内嵌公式
  - 基准数据：人工标注，XML
  - 下载地址：http://www.icst.pku.edu.cn/cpdp/data/marmot_data.htm

实验结果 – 分类性能评估

训练集：100页，测试集：100页

文本行分类结果

学习算法	准确率（%）	召回率（%）	F1（%）
支持向量机	93.20	98.77	95.90
装袋	97.26	98.01	97.63
装袋（测试结果）	97.04	95.40	96.21

词块分类结果

学习算法	准确率（%）	召回率（%）	F1（%）
支持向量机	91.31	92.38	91.84
装袋	94.75	93.80	94.28
装袋（测试结果）	93.25	89.11	91.13

实验结果 – 评估准则

- **完全正确**
  - 识别结果与基准集中对应公式区域完全一致
- **过分割**
  - 识别结果仅占基准集中对应公式区域的一部分
- **过合并**
  - 识别结果不仅包含了正确公式区域，还把其他公式或非公式区域也包含在内
- **误识别**
  - 识别结果为非公式区域
- **漏失别**
  - 基准集中未被识别的公式个数

实验结果

独立公式定位结果

方法	实际个数	识别个数	完全正确	过分割	过合并	误识别	漏识别
文献[6]	598	540	53.52%	40.37%	5.00%	1.11%	12.54%
文献[13]	598	710	53.94%	39.01%	3.80%	3.24%	1.17%
本文	598	543	69.98%	13.63%	14.36%	2.03%	1.34%

内嵌公式定位结果

方法	实际个数	识别个数	完全正确	过分割	过合并	误识别	漏识别
文献[13]	1483	2018	34.69%	32.46%	5.05%	27.80%	12.54%
文献[14]	1483	1573	40.37%	26.70%	14.94%	17.99%	17.46%
本文	1483	1366	61.79%	14.49%	10.32%	13.40%	19.69%

- 漏识别率：单个字符构成的公式变量（例如：A，x）未被分类为内嵌公式所造成

目录

- 背景
- 相关工作
- 数学公式定位
- 实验结果
- *结论*

结论

- 本文贡献
  - 机器学习和规则相结合的公式定位方法
  - 针对中文文档的分行和分词方法
  - 针对中文文档的行/词特征和分类器
  - 过分割：行合并和词块合并等后处理规则
  - 公开可用的中文标注数据集
- 未来工作
  - 公式语法与领域知识应用于公式定位后处理

参考文献（1/2）

1. Zanibbi, R., and Blostein, D. . Recognition and retrieval of mathematical expressions. International Journal on Document Analysis and Recognition (IJDAR), 2012, 15(4): 331–357
2. 王科俊, 王黎斌, 林桂芳. 科技文献中数学公式定位技术概述. 自动化技术与应用, 2004, 23(5)：1–4
3. 靳简明, 江红英, 王庆人. 数学公式图像处理综述. 模式识别与人工智能, 2005, 18(4)：429–440
4. Kacem, A., Bela ïd, A., Ahmed, M. B.. Automatic extraction of printed mathematical formulas using fuzzy logic and propagation of context. International Journal on Document Analysis and Recognition (IJDAR), 2001, 4(2): 97–108
5. Suzuki, M., Tamari, F., Fukuda, R., et al.. INFTY: an integrated OCR system for mathematical documents// In Proceedings of the ACM symposium on Document engineering. Grenoble, 2003: 95–104
6. Baker, J., Sexton, A. P., Sorge,V.. Towards Reverse Engineering of PDF Documents// Towards Digital Mathematics Library. Bertinoro, 2011: 65–75
7. Chowdhury, S. P., Mandal, S., Das, A. K., et al.. Automated segmentation of math-zones from document images// In Proceedings of International Conference on Document Analysis and Recognition (ICDAR). Edinburgh, 2003: 755–759
8. Chang, T. Y., Takiguchi, Y., Okada, M.. Physical structure segmentation with projection profile for mathematic formulae and graphics in academic paper images// In Proceedings of International Conference on Document Analysis and Recognition (ICDAR). Curitiba, 2007: 1193–1197
9. Garain, U.. Identification of mathematical expressions in document images// In Proceedings of International Conference on Document Analysis and Recognition (ICDAR). Barcelona, 2009: 1340–1344

参考文献（2/2）

10. 靳简明, 江红英, 王庆人. 数学公式识别系统: MatheReader. 计算机学报, 2006, 29(11): 2018–2026
11. Drake, D. M., Baird, H. S.. Distinguishing mathematics notation from English text using computational geometry// In Proceedings of International Conference on Document Analysis and Recognition (ICDAR). Seoul, 2005: 1270–1274
12. Liu, Y., Bai, K., Gao, L.. An efficient pre-processing method to identify logical components from PDF documents// In Advances in Knowledge Discovery and Data Mining. Shenzhen, 2011: 500–511
13. Lin, X., Gao, L., Tang, Z.. et al.. Mathematical formula identification in PDF documents// In Proceedings of International Conference on Document Analysis and Recognition (ICDAR). Beijing, 2011: 1419–1423
14. Lin, X., Gao, L., Tang, Z.. et al.. Identification of embedded mathematical formulas in PDF documents using SVM// In Document Recognition and Retrieval (DRR) XIX. SPIE-IS&T, Burlingame, 2012: 8297 0D 1–8
15. University of Washington UW-III English/Technical Document Image Database. [EB/OL]. (1996)[2013-08-20] http://www.science.uva.nl/research/dlia/datasets/uwash3.html
16. Suzuki, M., Uchida, S., Nomura, A.. A ground-truthed mathematical character and symbol image database// In Proceedings of International Conference on Document Analysis and Recognition (ICDAR). Seoul, 2005: 675-679
17. Gao, L., Tang, Z., Qiu, R.. A mixed approach to auto-detection of page body// In Proceedings of Conference on Document Recognition and Retrieval XV. SPIE-IS&T, San Jose, 2008: 68150 1–7
18. A two-dimensional quality mesh generator and Delaunay triangulator. [EB/OL]. (2005-07-28)[2013-08-20] http://www.cs.cmu.edu/~quake/triangle.html

谢 谢！