复杂文档信息处理中的分析及测试

XX

(XXXXXXXXXXXXXX，XX XXXXXX)

**摘 要** 分析了当前文档信息处理中复杂数据的处理问题，提出了解决方案并在此基础上构建一个原型系统。

**关键词** 检索, 复杂文档信息处理, 测试集

Analysis and Test in the Complex Document Information Processing

X X

(Beijing CUTECH Technology Development Co.,Ltd. Beijing 100101, China)

**Abstract** Research the complex data processing in documents information analysis, propose solutions and implement a prototype system.

**Keywords** Retrieval, Complex document information processing, Test collection

# 1.复杂文档信息处理问题

复杂文档信息处理是基于多种组合的分析，包括手写体，表格，图片以及可打印的格式化文本。X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X 。这些数据集可能是由军方，强力部门或者公司其他部门所收集，只有对这些数据进行扫描之后才能在短时间内理解和利用它们的全部价值。

这些X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X一类文档。

因而从这些数据中提取有用的信息很大程度上依赖于人工分类，多种分析工具的组合使用以及结果校对。X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X XX X X X X X X X X X X X X X X X X ；其次，当处理大量文档时，人工校验，交叉检查和持续检查缓慢而且易错，这极大的降低了所得信息的价值。

因此，无论在研究还是商业领域，复杂文档信息处理的操作问题是由统一化工具的缺失造成的。X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X。因此，创建一个复杂文档信息处理测试集是该领域不可缺少的一环。

# 2.构造一个复杂文档信息处理系统

设计一个完整的复杂文档信息处理系统，它实现以下功能：

X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X 。目标是允许用户访问所有包含在复杂文档中的信息，包括图片，文本，手写体，图像和结构化数据。

## 2.1．图片分析

图片分析是复杂文档信息处理系统的一个关键部分因为数据集通常是以纸模和扫描的形式获得。X X X X X X X X X X X X X X X 。可打印的文本将被从手写体和图像[1], [2]中分离出来，可打印和手写文本中的文字将被从彼此中分离出来并转换为合适的字符集，描述这些不同组件的特性将在更高级别的处理中抽取。

## 2.2．文本/符号分析

文本/符号分析以类标签和结构化说明的方式解析文本文档。X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X。相同的，图表集中化[4]在某些情况下能够为文档中的相似图形提供鉴别标签。

## 2.3．数据挖掘完整性

从不同的数据源中获取完整的数据是一个关键的中间功能---系统支持众多传统数据源，诸如非结构化文本，半结构化的XML文本数据也包括结构化的数据库查询[10]。X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X 。这个基础功能的目标是使要查询的信息从复杂文档中提炼出来。

## 2.4．数据挖掘

数据挖掘X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X和非结构化数据组合上。

# 3.建立复杂文档信息处理评价框架

一个使得组件X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X 数据类型。

由加利福尼亚大学主办X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X文字。

一个测试集不仅仅是一个数据集。X X X X X X X X X X X X X X X X X X X X X X X X X X X。为了支持组件调优以及系统效率的评估，所选的例子必须被人工注解以符合测试平台的检索，分析和数据挖掘任务。

使用指标方式评估系统中独立的组件。X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X。定量指标，定性分析和跟踪误差的成分来源都将用来评估最初的原型和定位未来在复杂文档信息处理方面可能的问题。

# 4.当前状态

当前系统是一个完整的针对以架构化格式描述文档数据的检索系统原型。X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X 。复杂文档信息将被整合成一个较低层次的处理模块而达到成熟。

符号分析关注基于文本分类的主题和风格。X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X。正在开发一个新的语言分析提取器，能够获得更精确的文本分类[5]。

文档图像处理的研究当前集中在文本/手写体分离，文本分段，笔划分割以及特征提取上。X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X 。正在研究诸如参数选择及尺度恒定的问题。

**参考文献**

[1] G. Agam and I. Dinstein. Adaptive directional morphology with application to document analysis. In P. Maragos, R. W. Schafer, and M. A. Butt, editors, *Mathematical Morphology and its Applications* *to Image and Signal Processing*, pages 401–408. Kluwer Academic Publishers, 1996.

[2] G. Agam and I. Dinstein. Regulated morphological operations. *Pattern Recognition*, 32(6):947–971, 1999.

[3] Rakesh Agrawal, Hiekki Mannila, Ramakrishnan Srikant, Hannu Toivonen, and A. Inkeri Verkamo. Fast discovery of association rules. Pages 307–328, 1996.

[4] S. Aksoy and R. M. Haralick. Graph-theoretic clustering for image grouping and retrieval. In *Proc. IEEE Conf. on Computer Vision and* *Pattern Recognition*, pages 63–69, 1999.

[5] S. Argamon and J. Dodick. Linking rhetoric and methodology in formal scientific writing. In *Proc. 26th Annual Meeting of the Cognitive Science Society*, August 2004.

[6] S. Argamon, M. Koppel, J. Fine, and A. R. Shimony. Gender, genre, and writing style in formal written texts. *Text*, 23(3), 2003.

[7] S. Argamon, M. ˇ Sari´c, and S. S. Stein. Style mining of electronic messages for multiple author discrimination. In *Proc. ACM Conference* *on Knowledge Discovery and Data Mining*, 2003.

[8] S. Beitzel, E. Jensen, A. Chowdhury, D. Grossman, O. Frieder, and N. Goharian. On fusion of effective retrieval strategies in the same information retrieval system. *Journal of the American Society of Information Science and Technology*, 55(10), July 2004.

[9] A. Genkin, D. D. Lewis, and D. Madigan. Large-scale Bayesian logistic regression for text categorization. *Journal of Machine Learning Research*, 2004. submitted.

[10] D. Grossman, S. Beitzel, E. Jensen, and O. Frieder. IIT intranet mediator: Bringing data together on a corporate intranet. *IEEE IT PRO*, January/February 2002.

[11] D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, April 2004.