OpenOffice.org

Conference 2008 Beijing

世界开源大会

Learn more about office users

-- Feature usage study by document element statistics

Rui SuYing

IBM Lotus Symphony

Agenda

- **Why we need analyse office feature usage**
- **Feature usage study by document element statistics**
  - Introduction on methodology and tool
- **Statistic result sharing**
- **Future work**
- **Q&A**

Why we need analyse office features usage

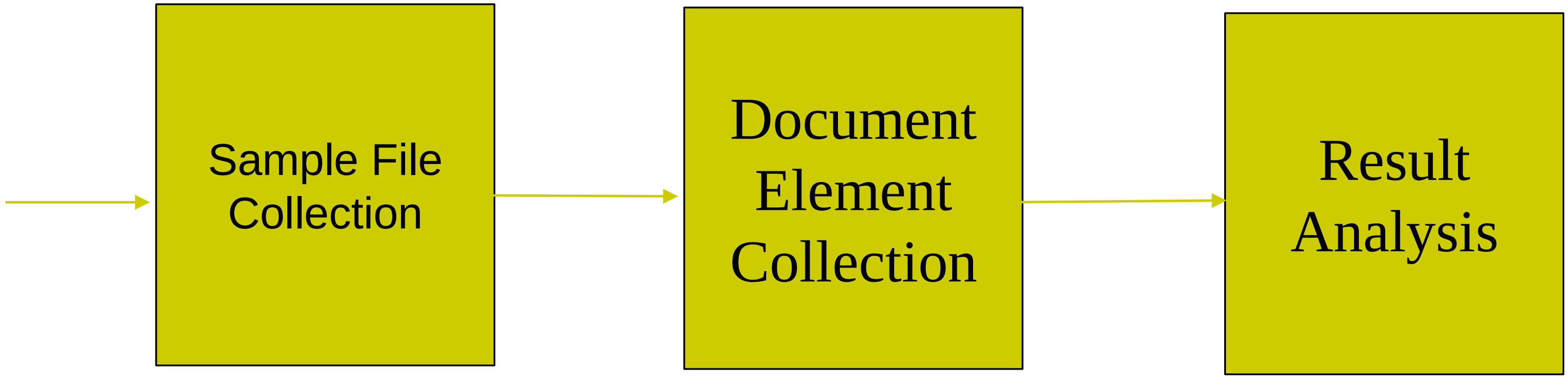
- **Thousands of features in office application**
  - About 270 menu items in Office 2003, more features in 2007
  - 400+ subsections in ODF spec used to describe office features
- **Large quality of features brings challenges to office product**
  - UI design sometimes depends on feature usage
  - Task prioritization
  - Limited dev resource vs. endless requirements

Some approaches

- **User Survey**
  - Questionnaire Survey
  - Customer evaluation
  - Can get special requirement from special user group
- **User behaviour collection in office application**
  - User action recording when using office application
  - Focusing on UE improving
  - Can get accurate user data
  - Not all users are willing to join for privacy concern
  - Cross network framework needed

Feature usage study
by document element statistics

Feature usage study
by document element statistics



- Large quantity of files were collected for analysis use
- We detached document elements usage from the sample files statically
- Result analysis convert raw data to visual result

Feature usage study by document element statistics

-- Sample File collection

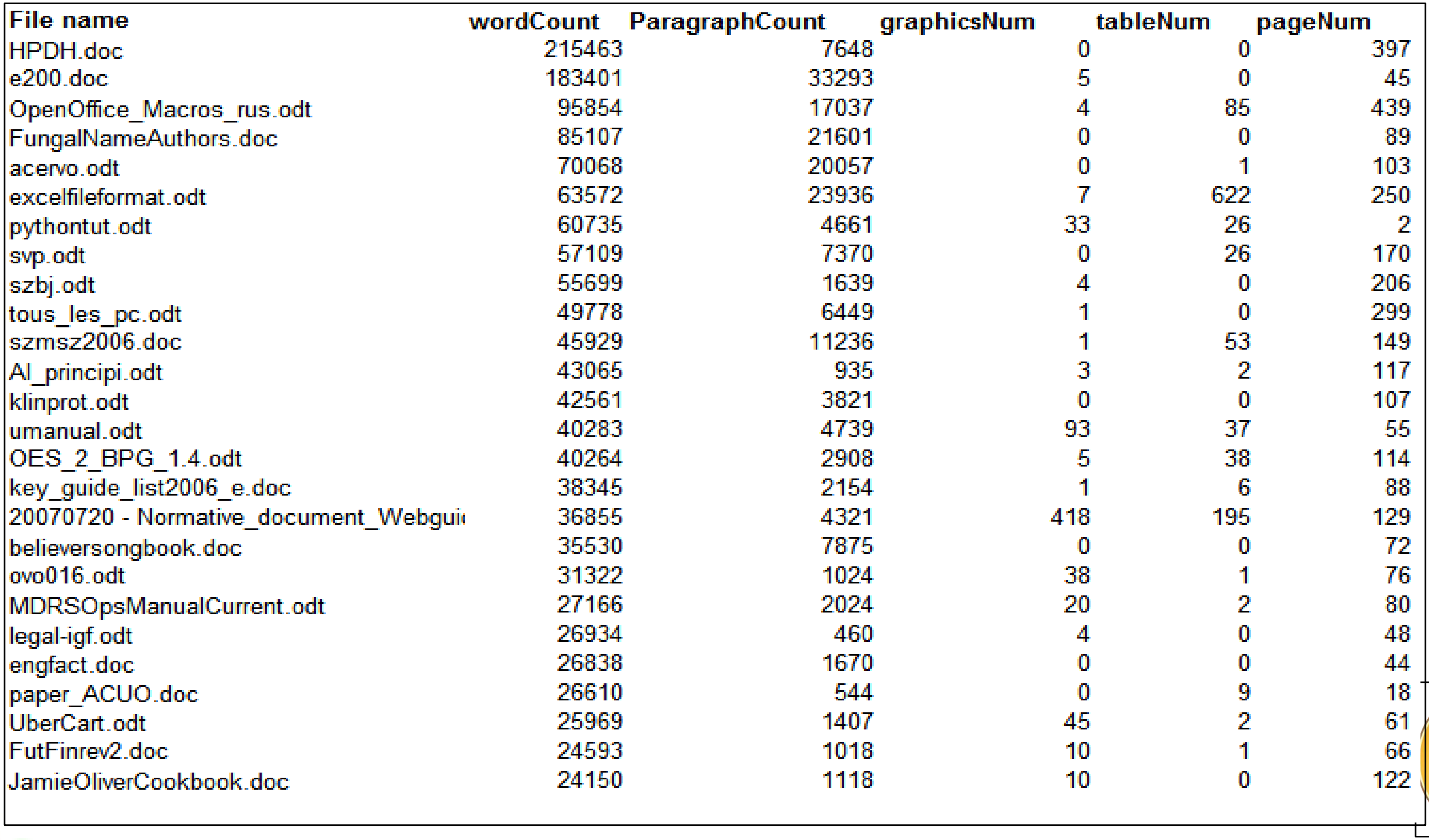
- Two key points
  - Large Quantity
  - As random as we can
- Methods
  - Google search with only file extension name as key word
  - Web download one by one
- Sample File Coverage
  - 1400+ spreadsheet files(xls,ods, 123)
  - 1600+ document files(doc, odt, lwp)
  - 400+ presentation files(ppt, odp, prz)(to be added)
  - 90%+ written in English, covering multiple language(Chinese, French, Japanese, etc)

Document element collection -- Methodology

- We need to analyse document formats
  - ODF
  - MS Binary
  - Lotus SmartSuite
- Parse and load sample files with different filters in IBM Lotus Symphony/OpenOffice
- Document element collection with UNO call after document loading
- *Why not work on disk file than collecting after file loading?*
  - *XML parser can handle ODF format, but cannot deal with MS and Lotus SS format*
  - *Some information can not be collected before document formatting*

Statistic Result Analysis

- Raw result – document element usage per file



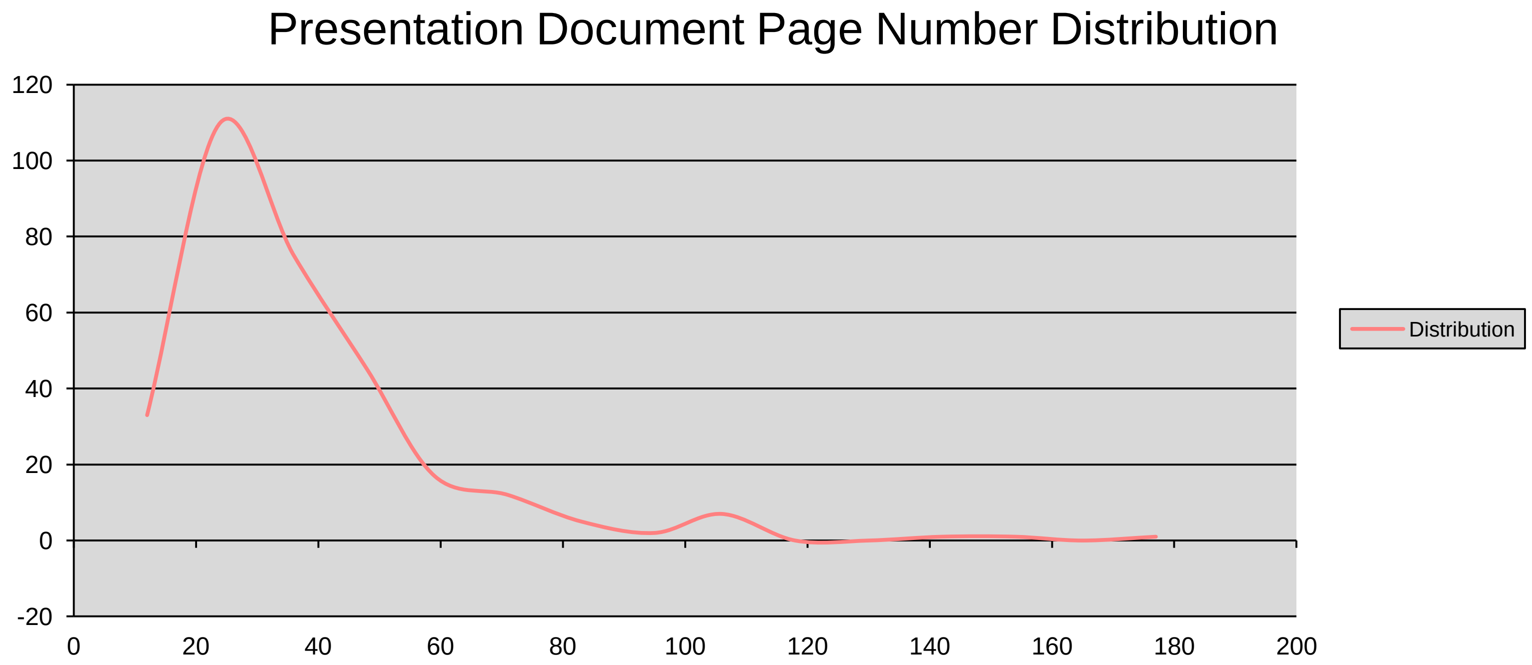
File name	wordCount	ParagraphCount	graphicsNum	tableNum	pageNum
HPDH.doc	215463	7648	0	0	397
e200.doc	183401	33293	5	0	45
OpenOffice_Macros_rus.odt	95854	17037	4	85	439
FungalNameAuthors.doc	85107	21601	0	0	89
acervo.odt	70068	20057	0	1	103
excelfileformat.odt	63572	23936	7	622	250
pythontut.odt	60735	4661	33	26	2
svp.odt	57109	7370	0	26	170
szbj.odt	55699	1639	4	0	206
tous_les_pc.odt	49778	6449	1	0	299
szmsz2006.doc	45929	11236	1	53	149
AI_principi.odt	43065	935	3	2	117
klinprot.odt	42561	3821	0	0	107
umanual.odt	40283	4739	93	37	55
OES_2_BPG_1.4.odt	40264	2908	5	38	114
key_guide_list2006_e.doc	38345	2154	1	6	88
20070720 - Normative_document_Webgui	36855	4321	418	195	129
believersongbook.doc	35530	7875	0	0	72
ovo016.odt	31322	1024	38	1	76
MDRSOpsManualCurrent.odt	27166	2024	20	2	80
legal-igf.odt	26934	460	4	0	48
engfact.doc	26838	1670	0	0	44
paper_ACUO.doc	26610	544	0	9	18
UberCart.odt	25969	1407	45	2	61
FutFinrev2.doc	24593	1018	10	1	66
JamieOliverCookbook.doc	24150	1118	10	0	122

Statistic Result Analysis

- Average value, maximum value, minimum value
- Element use frequency distribution analysis
- We leveraged D.Scott's method
  - Find a proper bin width, get the number of document files whose element usage is in the bin
  - The number combined with the bin composes distribution
  - Bin width = 3.49 * Standard deviation of sample data * the quantity of sample data ^(-1/3)
-

Statistic Result Sharing

Presentation Documents(odp+ppt files)

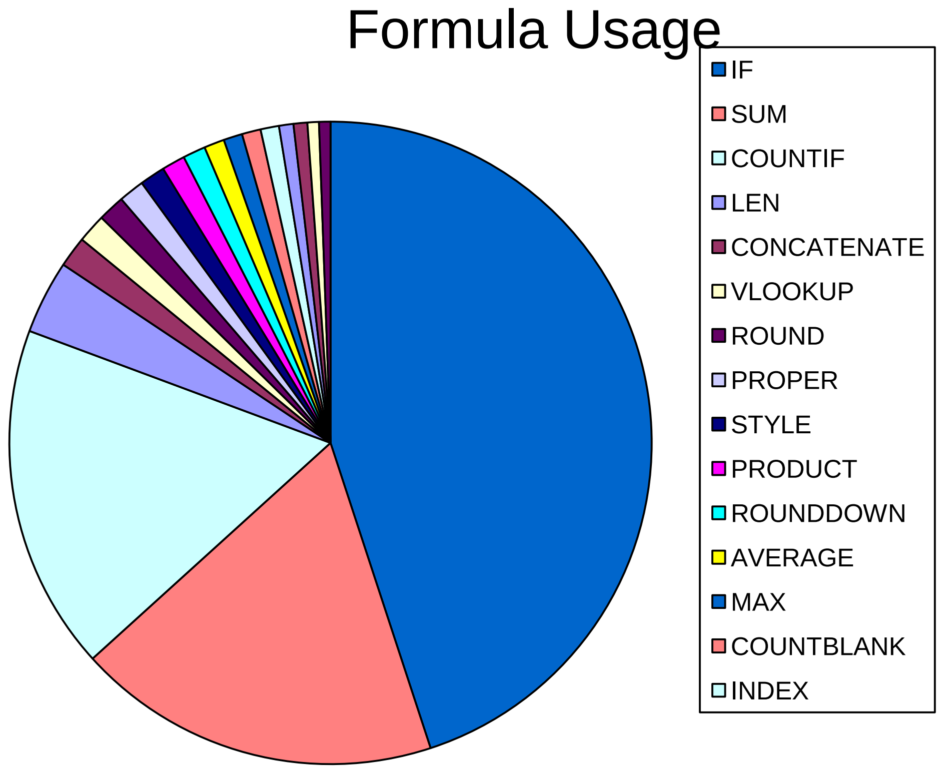


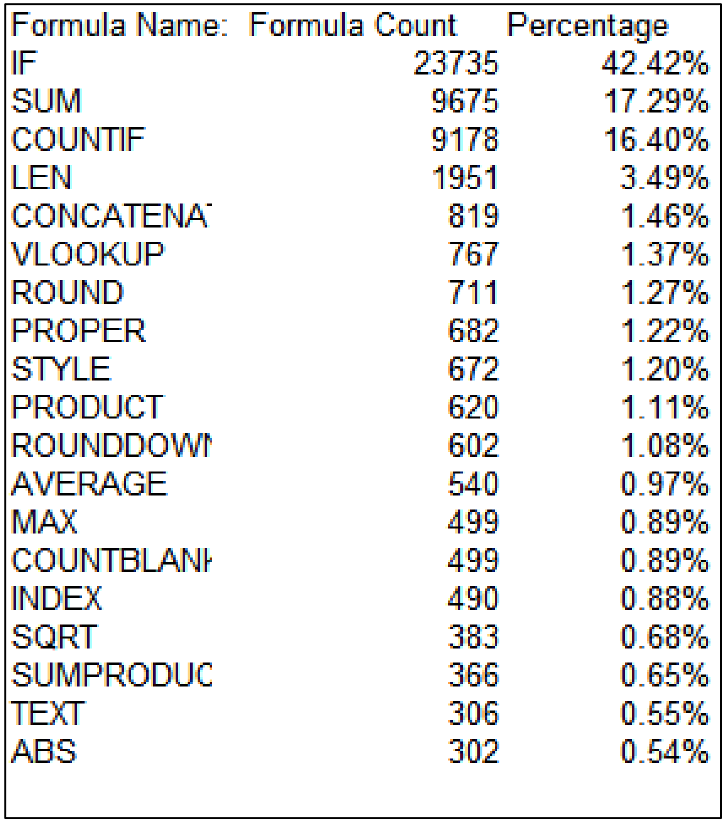
- **412 sample files**
- **30.71 slides as average**
- **Presentation files with less than 30 slides covers more than 90% usage**

What presentation slides number tells us

- Load/save performance evaluation
  - 90% coverage when page number is less than 30
  - 95% coverage when page number is less than 70
- Page Slider Design
  - Why we need a page slider in presentation
  - A reference for page slider design -- 6 pages shown in page slider as default in Symphony/7 pages shown as default in MS PPT 2003

Spreadsheet Documents(xls+ods file)





Formula Name:	Formula Count	Percentage
IF	23735	42.42%
SUM	9675	17.29%
COUNTIF	9178	16.40%
LEN	1951	3.49%
CONCATENA	819	1.46%
VLOOKUP	767	1.37%
ROUND	711	1.27%
PROPER	682	1.22%
STYLE	672	1.20%
PRODUCT	620	1.11%
ROUNDDOW	602	1.08%
AVERAGE	540	0.97%
MAX	499	0.89%
COUNTBLAN	499	0.89%
INDEX	490	0.88%
SQRT	383	0.68%
SUMPRODUC	366	0.65%
TEXT	306	0.55%
ABS	302	0.54%

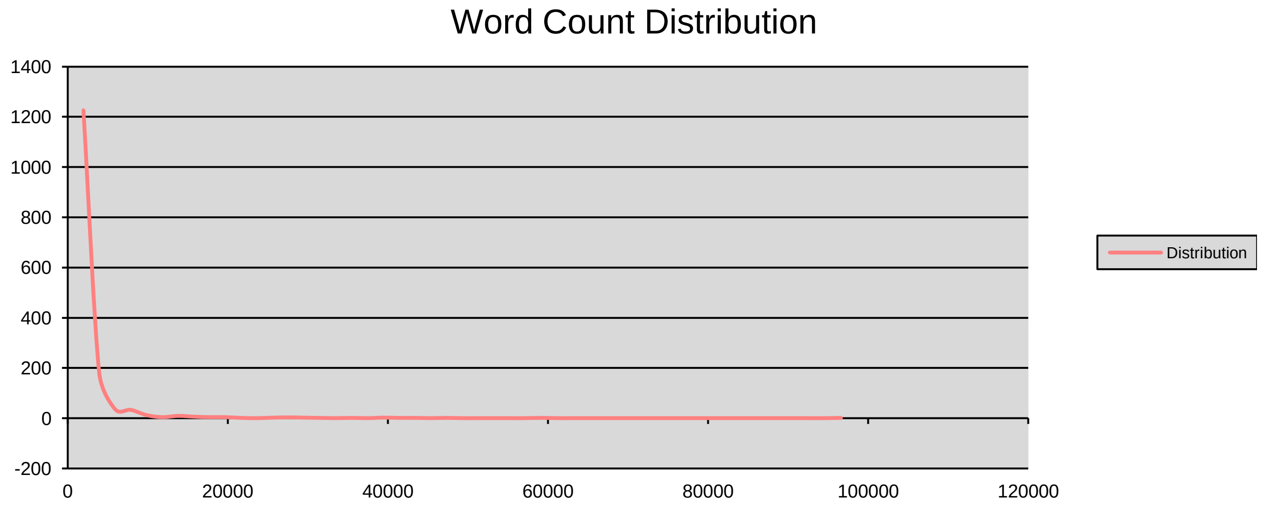
- **Top 10 formulas covers 88.31% usage**
- **Total 129 formula used in 1531 sample files**

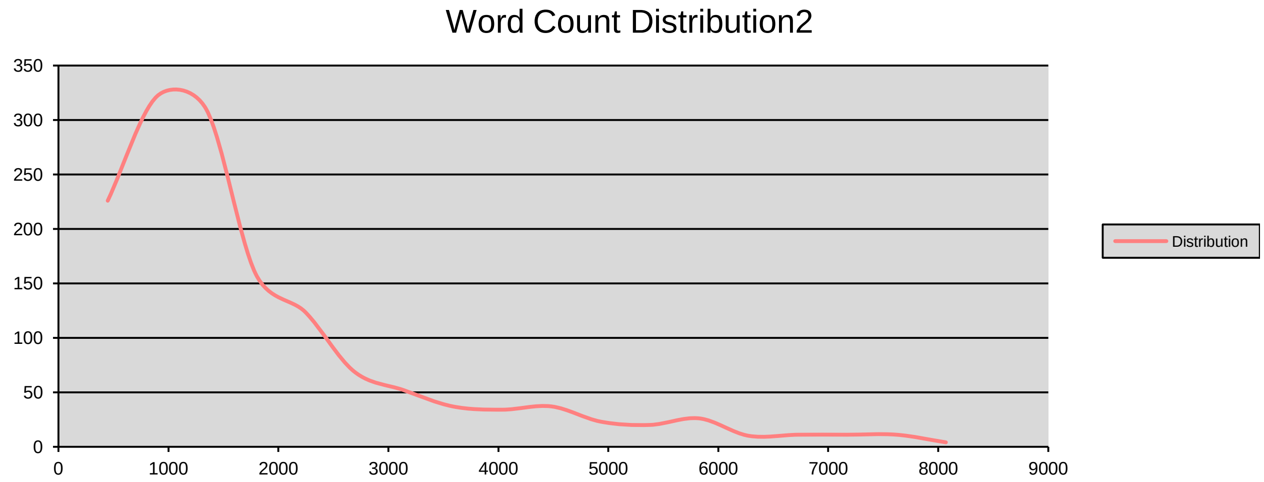
What Formula Usage tells us

- Assumption:
  - The spreadsheet file collected from web indicates normal users behavior
- Only 129 formulas used in more than 1500 sample files
  - OpenOffice supports 371, Symphony supports 377
  - A reference when we develop a light-weight spreadsheet(web spreadsheet)
- Formula testing focus finding
- *Thinking...*
  - *If we can get enterprise user's sample file, perhaps we can get a different result.*

Word Processor Document

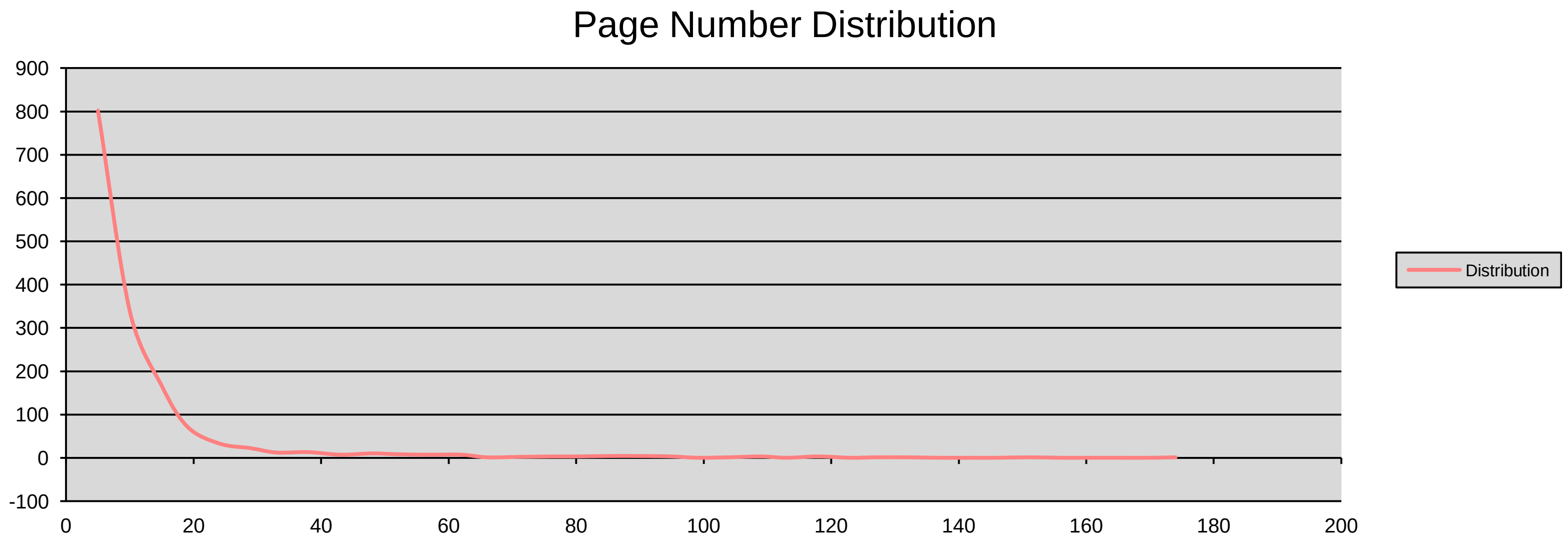
- Word Count Distribution & Analysis





Word Processor Document

- **Page Number Distribution & Analysis**



- Average Page Number: 10.15 pages
- Short Documents published in web

Word Processor Document

- **Table usage in sample document**
  - **Table used in 44.58% of sample documents**
  - **Most of them are middle size**
- **Graphic usage in sample document**
  - **Graphic usage in 43.41% of sample documents**

Limitation of document element analysis by file sampling

- Issues in file sampling
  - Coverage
  - Randomicity
  - Lack of files in enterprise environment
  -
- Limitation in document element collection
  - Limitation of filter capability of Symphony and OpenOffice
  - UNO Call quality

Future Work

Future Work

- We will go deeper in this work
  - Animation usage statistic – For development priority and UI design
  - Chart usage - Chart type & Chart property usage
  - Paragraph statistic – Reference for collaboration writing and paragraph sharing
- Document element statistic for sample files
  - documents for different industries and different language
  - Issues: Document categorisation for industries
  -
- ***A more smart way to collect sample file***

OpenOffice.org Conference 2008 Beijing

Q & A

Reference

- MS CEIP - http://www.microsoft.com/products/ceip/EN-US/default.mspx
- D. Scott, “On Optimal and Data-based Histograms,” Biometrika, vol. 66, no. 3, pp. 605–610, 1979.

Thanks！

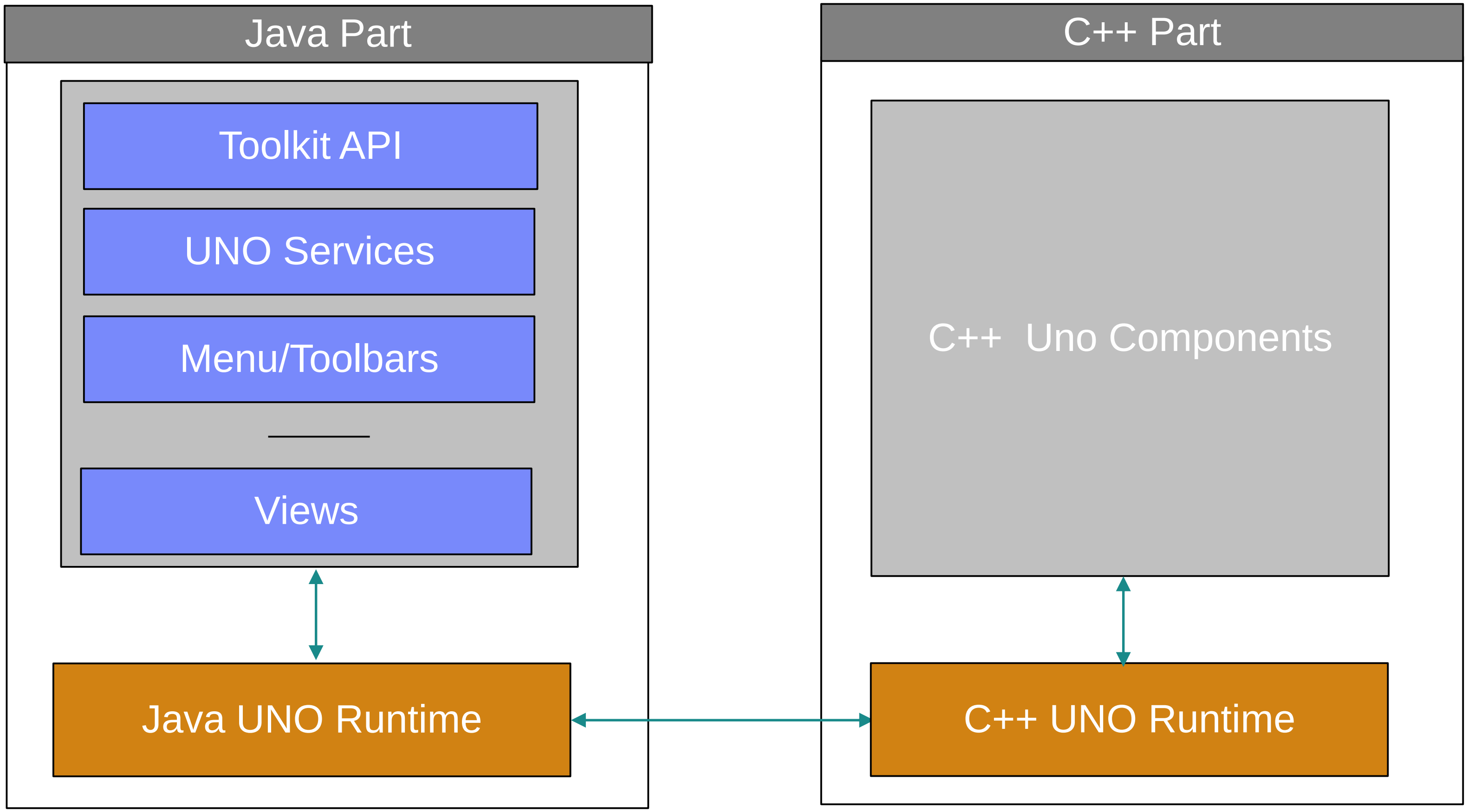
凝聚全球力量 绽放开源梦想

www.OOobeijing2008.com

Feature usage study
by document element statistics

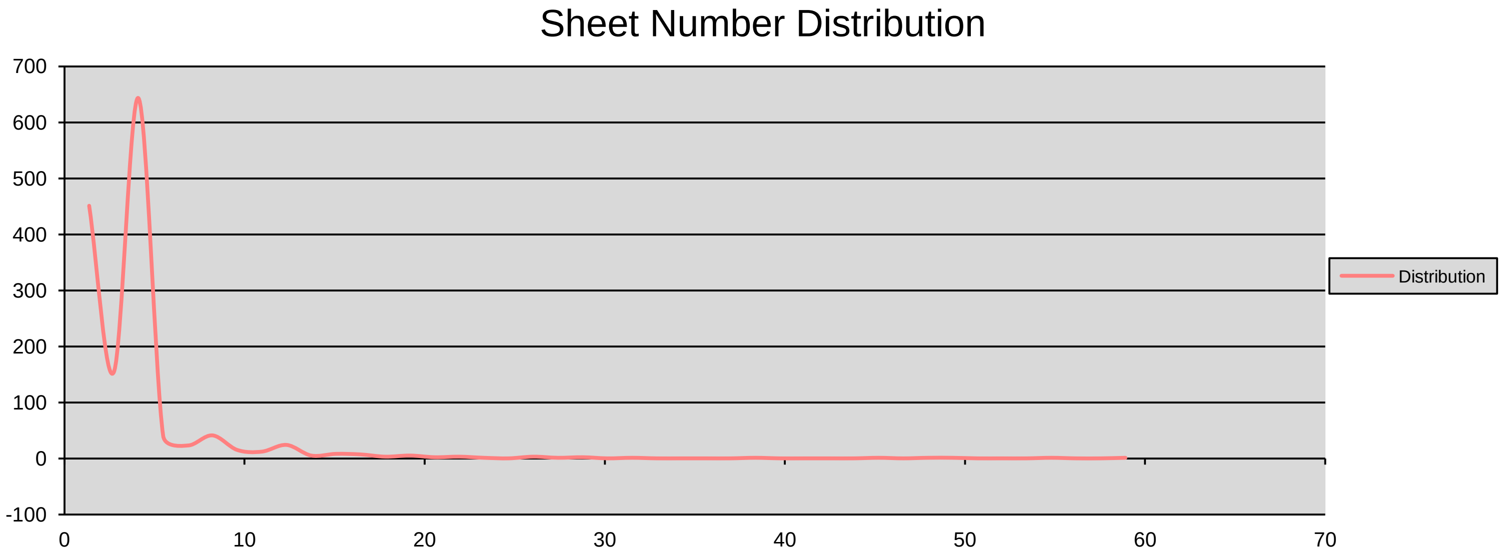
- Sample files in actual use are resource for feature usage study
  - Document element usage information are stored in those files
  - Large quantity of sample files will tell us something
- We can happen to find large quality of files from web
  - Assumption: most of documents in web are for actual use
- We have existing tool to be reused for the feature analysis
  - IBM Lotus Symphohy/OpenOffice have ability to open multiple types of documents
  - IBM Lotus Symphony/OpenOffice can recognize most of document elements

Document element collection – Symphony plugin



Spreadsheet Documents(xls+ods file)

- **Sheet number distribution show**



- Spreadsheet Document Sampling issues
  - Different usage between enterprise users and individual users