PENNSTATE

COLLEGE OF INFORMATION SCIENCES AND TECHNOLOGY

TableSeer: Automatic Table Metadata Extraction and Searching in Digital Libraries

Ying Liu, Prasenjit Mitra, C. Lee Giles
Information Sciences and Technology
Pennsylvania State University
University Park, PA 16802, USA

Outline

- Motivation - $Chem_X$Seer
  - Importance of tables for subsequent research
- Related work
- TableSeer - extracting and indexing tables
- TableRank – table ranking algorithm
- Experimental results
- Conclusions and future work

Motivation

- Tables are an important data resource for information retrieval
- The condensed information in tables enables users to quickly find important information without having to read the entire document

Table 1 Temperature effect on resistance change (ΔR) and response time of tin oxide thin film with 1% CCl_4

Temperature/ °C	$\Delta R^a/\Omega$	$\frac{\Delta R}{(R, O_2)}$ (%)	Response time	Reproducibiliy
100	223	5	~22 min	Yes
200	270	9	~7-8 min	Yes
300	1027	21	<20 s	Yes
400	993	31	~10 s	No

[a] $\Delta R = (R, CCl_4) - (R, O_2)$.

Table 4 The separation of ephedrine enantiomers on computationally designed MIPs under optimized conditions. Ten microlitres of sample (concentration, 1 mg ml^{-1}) were injected for analysis. The flow rate was 1 ml min^{-1}

Polymer	Eluent	$k'(-)$	$k'(+)$	α
P1 (IA)	10% acetic acid in chloroform	3.25	2.76	1.18
P2 (MA)	1% acetic acid in chloroform	6.48	4.82	1.34
P3 (HEM)	0.1% HMDA in chloroform	1.09	0.77	1.42
P4 (AA)	0.1% HMDA in chloroform	1.1	0.92	1.2
P5 (2-VP)	0.1% HMDA in chloroform	0.1	0.1	1

- Interest in and use of past data necessitates table indexing and search
  - *Table data is now manually extracted from documents!*
- Existing search engines do not support table search and no table search engine exists

Chem$_X$Seer

- NSF funded portal for researchers in environmental chemistry integrating the scientific literature with experimental, analytical and simulation results and tools

- Provides unique metadata extraction, indexing and searching pertinent to the chemical literature.
  - **Tables (TableSeer)**
  - Chemical names and formulae
  - Figures

- After extraction, data is stored in API accessed xml databases

- Hybrid repository *(Not fully open):* Serves as a federated information interoperable system
  - Searchable and indexed scientific papers crawled from the web
  - User submitted papers and datasets (e.g. excel worksheets, Gaussian and CHARMM toolkit outputs)
  - Scientific documents and metadata from publishers (e.g. Royal Society of Chemistry)

- Takes advantage of developments in other funded cyberinfrastructure projects and open source software
  - CiteSeer^X, PlanetLab, Lucene, Fedora, etc.

http://chemxseer.ist.psu.edu

Research Issues

- Crawling documents
  - Filtering out the documents with tables
- Extracting tables from a document
  - Table Boundary Detection
  - Table Structure Analysis
  - Table Information Collection
- Diverse medium types, press layouts, cell types, affiliate table elements
- No standard table representation
- Table Indexing and Ranking
  - Current ranking schemes are inadequate and not designed for table search
- Result Interface

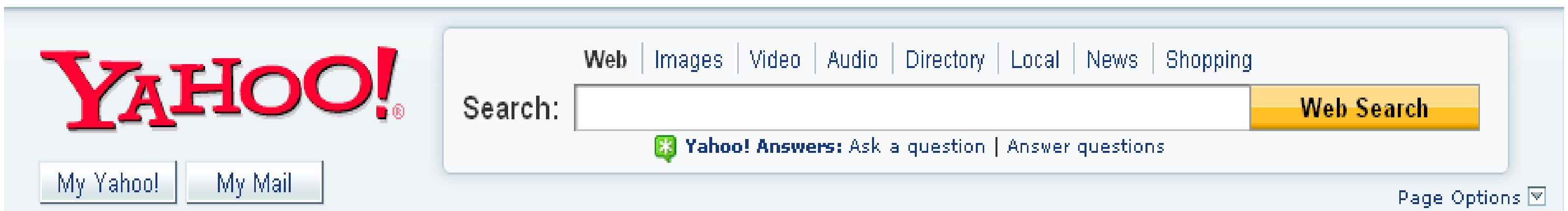
Our Approach

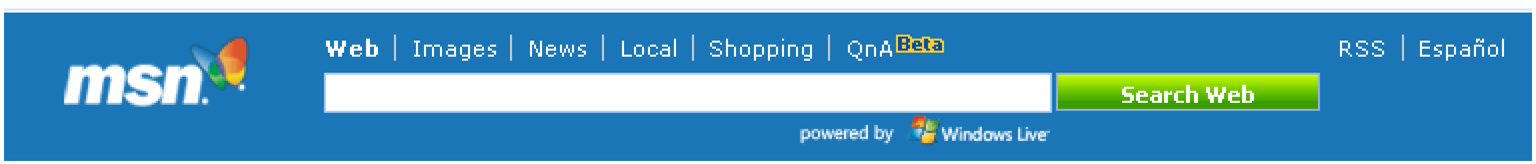
- Use machine learning methods (SVMs) and heuristics to automatically …. tables.
  - Identify
  - Extract
  - Represent
  - Index
  - Rank
- Take advantage of innate cell structure of tables for effective extraction
- Use standard open sources tools (Lucene) for indexing and extraction (pdfbox)
- Modular design

Our Contribution

- Developed a table search engine – TableSeer
- Designed a set of universal medium-independent metadata specifications for tables
- Created an novel first time table ranking algorithm, TableRank
  - Has a tailored term vector space and a novel term weighting scheme
- TableSeer processes an important document medium – PDF
- Developed a novel page box-cutting method to speed up the table detection
- Index table referenced text and captions in documents

Table search?

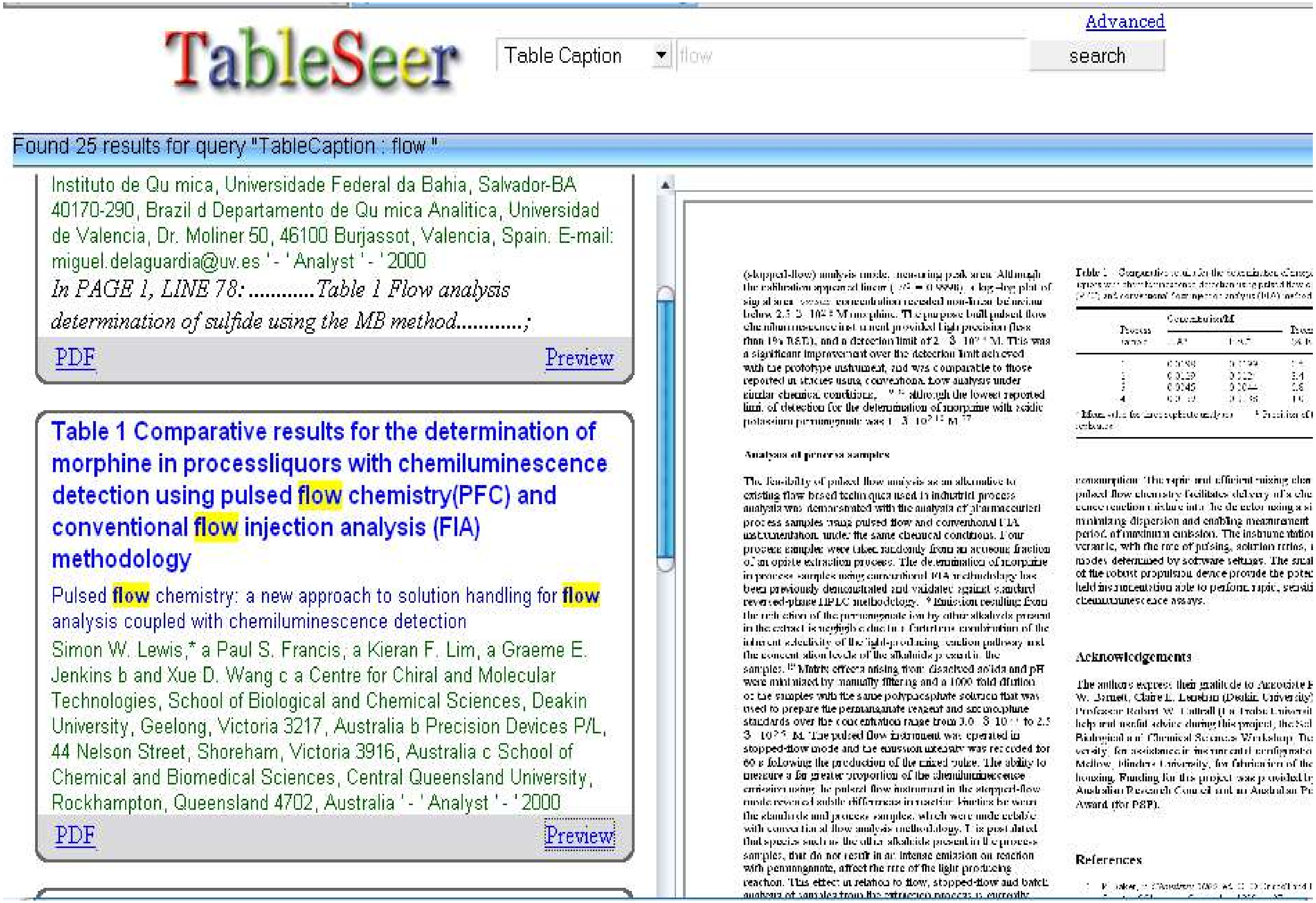




User can search for text, images, local news, maps, articles, etc., ***but not tables***

TableSeer

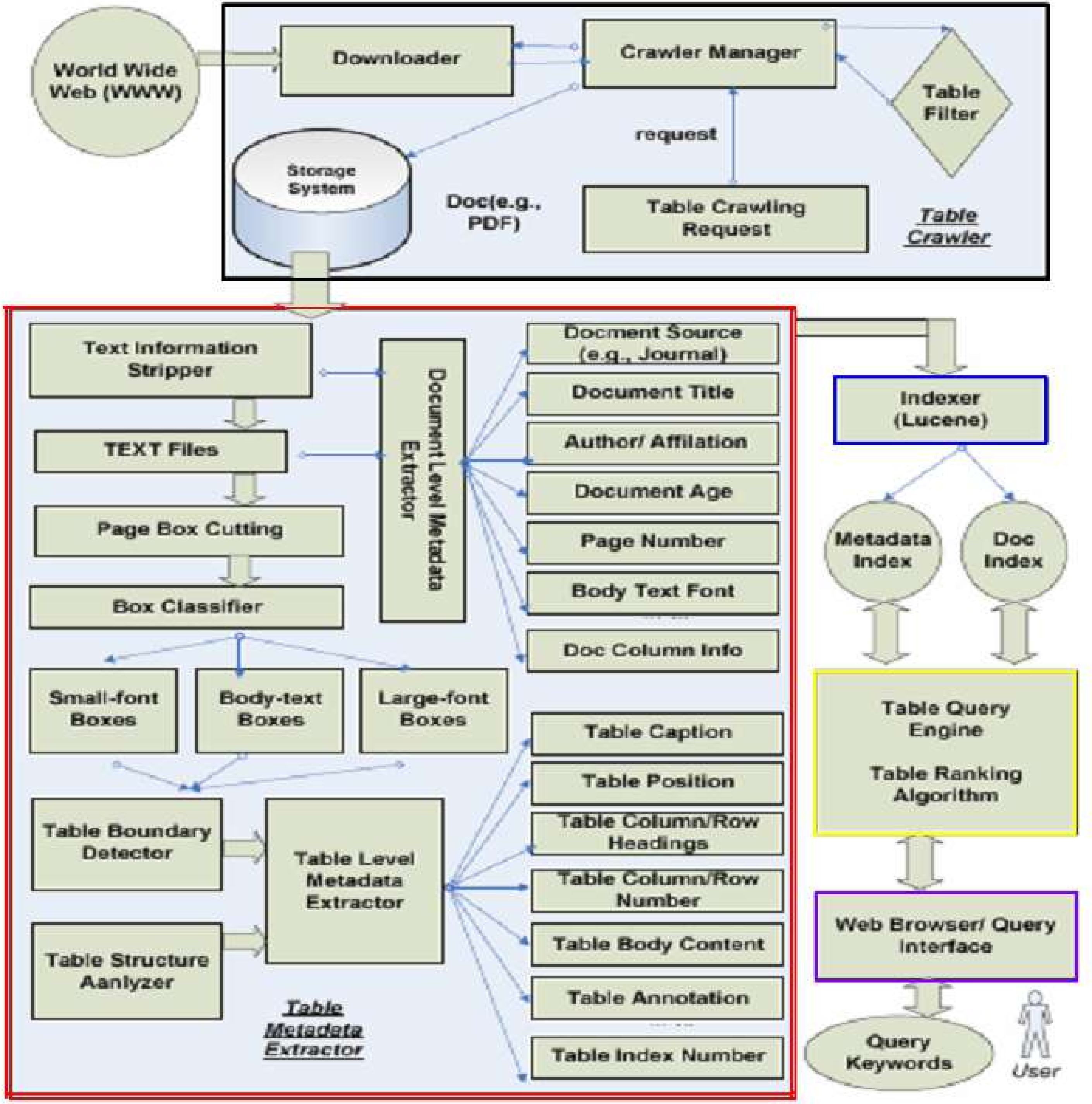
Beta online working design of a table search engine



Related Work

- Search html table content
  - TINTIN system [1]: table caption and table entries
  - Hu et. Al [2]: man-machine dialog to access the table data
  - Pyreddy et. Al [3]: associates tables with QA
- Table representation
  - Xinxin Wang [4]: conceptual model describing the table structure
  - Table markup: XHTML, OASIS
  - (our contribution) Integrating table structure and layout information, as well as the table-related information, and the document background information
- Table Extraction
  - Previous focus primarily on HTML documents or Images
  - (our contribution) Focus on untagged documents, e.g., PDF documents

TableSeer System Architecture



Related Works on Table Detection

- Zanibbi [28] provides a survey paper
  - Previous focus primarily on HTML documents or Images
    - Chen et al. [3] used heuristic rules and cell similarities to identify tables from web pages
    - Penn et al. [18] identify genuinely tabular information and news links in HTML documents
    - Yoshida et al. proposed a method to integrate WWW tables according to the category of objects presented in each table [27]
    - Chao et al. [2] reported their work on extract the layout and content from PDF documents.
    - Hadjar et al. developed a tool for extracting the structures from PDF documents.
  - Our contribution
    - Process the untagged documents, PDF documents, in the text level using the machine learning methods
    - A novel pre-process step based an interesting observation

Related works on table analysis with machine learning approaches

- Hurst mentioned in [5] that a Naive Bayes classifier algorithm produced adequate results
  - no detailed algorithm and experimental information
- Wang et. al. tried both the decision tree classifier and SVM to classify each given table entity as either *genuine* or *non-genuine* table
  - started with the detected tables
  - all features are only related to the table itself
- The most related work: Pinto et al. [19]
  - extracted table from plain-text government reports
  - adopted special labels and corresponding features
  - features focus on white space, text, and separator instead of the coordinate features
  - No detail about the table locating

Research Issues

- Table boundary in our problem
  - the table data rows without the table caption and the table footnote
- Table boundary detection problem contains four main sub-problems
  - Construct the lines in a document page
  - Identify and remove all the non-sparse lines from the line set
  - Identify and remove all the noisy sparse lines
  - Label table lines by considering the keywords

Page Box-Cutting Algorithm

- Improves the table detection performance by excluding more than 93.6% document content in the beginning

The Sparse-line Property of Tables

- Different lines in the same document page have
  - Different widths, text densities, spaces between words
- A document line is a *sparse line* if any of the condition is satisfied
  - The minimum space gap between a pair of consecutive words within the line is > a threshold *sg*.
  - The length of the line is < a threshold *ll*;
- Classifying document lines into sparse/non-sparse categories
  - The majority of the lines in a document belong to the non-sparse category
  - Narrowing down the search for table boundaries to sparse lines can save substantial time and effort

Machine Learning techniques

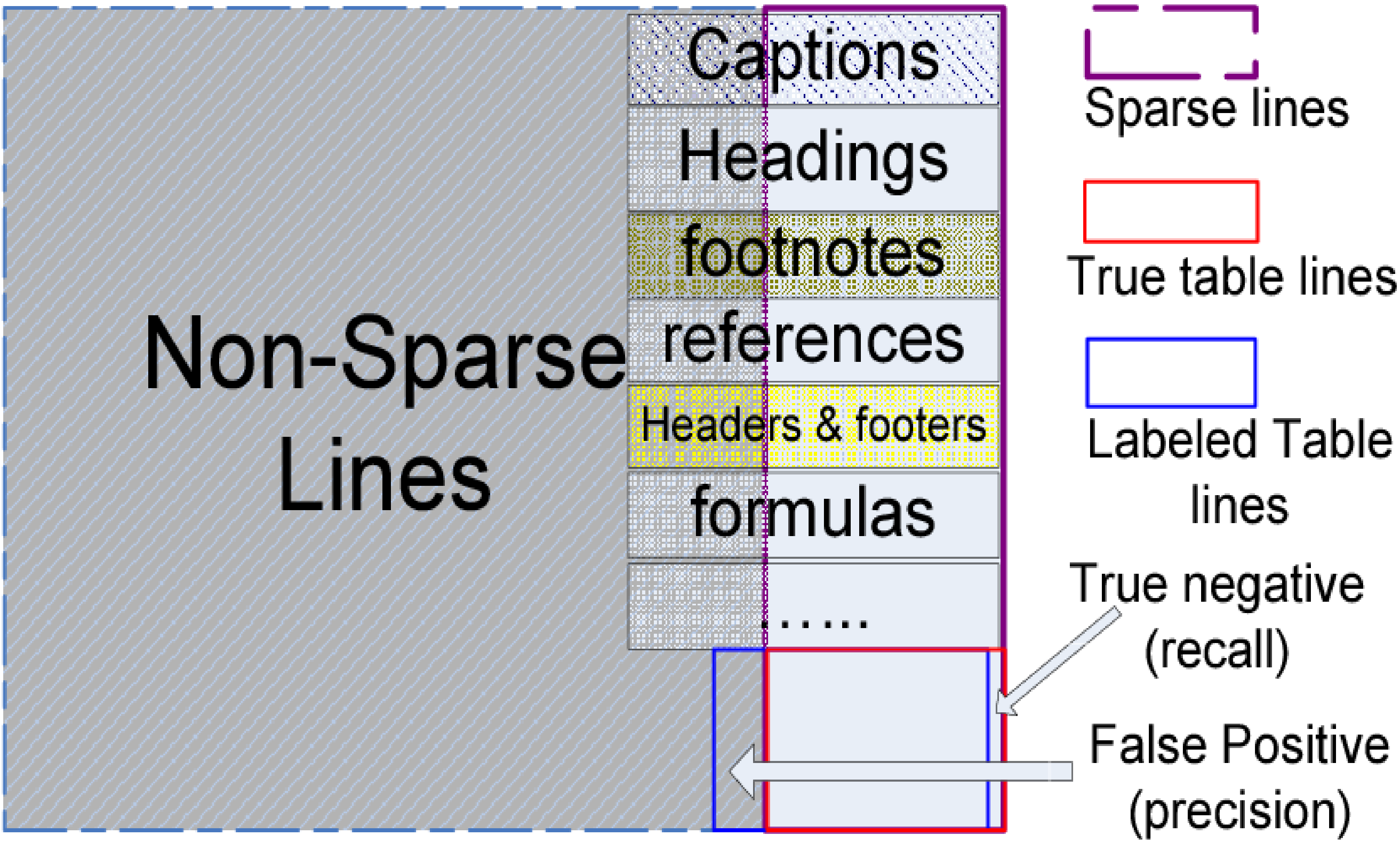
- Support Vector Machines
  - A binary classification method
  - Finding an optimal separating hyperplane *x* : *wx* + *b* = 0 to maximize the margin between two classes
- Conditional Random Fields (Lafferty, et al., [11])
  - Probabilistic model to segment and label sequence data

$$P(s|\mathbf{o}) = \frac{1}{Z_o} exp(\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_j f_j(s_{i-1}, s_i, \mathbf{o}, i)) \qquad Z_o = \sum_{s\in S} exp(\sum_{i=1}^{n}\sum_{j}\lambda_j f_j(s_{i-1}, s_i, \mathbf{o}, i))$$

PENNSTATE

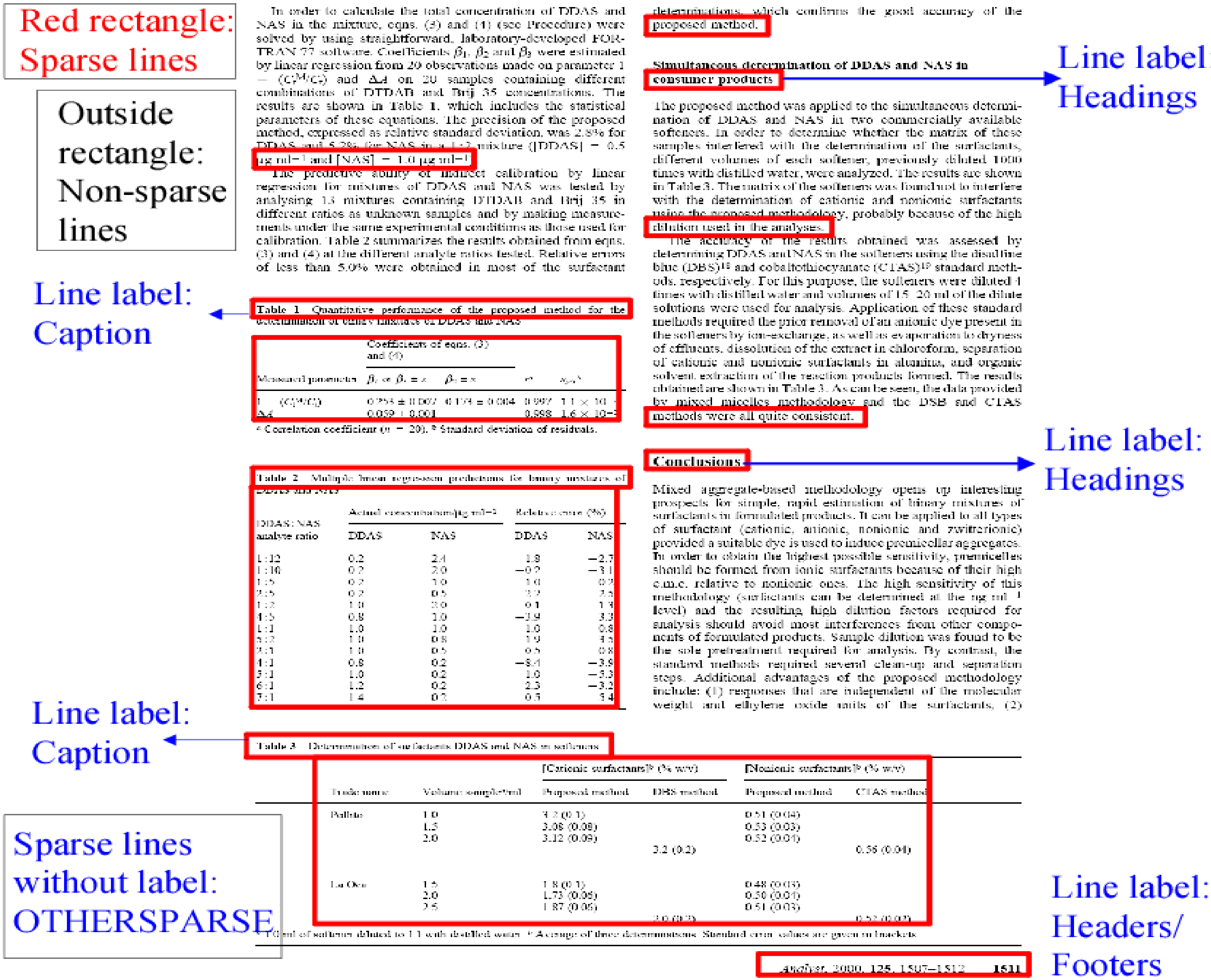
COLLEGE OF INFORMATION SCIENCES AND TECHNOLOGY

Line Labels



Line Labels

- Each line will be initially labeled as
  - either *SPARSE* or *NONSPARSE*
- *NONSPARSE* lines usually cover:
  - Most document titles, abstracts, paragraphs, etc
- *SPARSE* lines cover other specific document
- components entirely/partially:
  - tables, mathematical formulas, texts in figures, short headings, affiliations, short document headers and footers, references, etc.



In order to calculate the total concentration of DDAS and NAS in the mixture, eqns. (3) and (4) (see Procedure) were solved by using straightforward, laboratory-developed FORTRAN 77 software. Coefficients β_1, β_2 and β_3 were estimated by linear regression from 20 observations made on parameter 1 − (C_i^M/C_i) and ΔA on 20 samples containing different combinations of DTDAB and Brij 35 concentrations. The results are shown in Table 1, which includes the statistical parameters of these equations. The precision of the proposed method, expressed as relative standard deviation, was 2.8% for DDAS and 5.2% for NAS in a 1:2 mixture ([DDAS] = 0.5 μg ml^{-1} and [NAS] = 1.0 μg ml^{-1}).

The predictive ability of indirect calibration by linear regression for mixtures of DDAS and NAS was tested by analysing 13 mixtures containing DTDAB and Brij 35 in different ratios as unknown samples and by making measurements under the same experimental conditions as those used for calibration. Table 2 summarizes the results obtained from eqns. (3) and (4) at the different analyte ratios tested. Relative errors of less than 5.0% were obtained in most of the surfactant determinations, which confirms the good accuracy of the proposed method.

Table 1 Quantitative performance of the proposed method for the determination of binary mixtures of DDAS and NAS

	Coefficients of eqns. (3) and (4)			
Measured parameter	β_1 or $\beta_3 \pm s$	$\beta_2 \pm s$	r^a	$s_{y/x}^b$
1 − (C_i^M/C_i)	0.253 ± 0.007	0.173 ± 0.004	0.997	1.1 × 10^{-}[illegible]
ΔA	0.039 ± 0.001		0.998	1.6 × 10^{-}[illegible]

^{a} Correlation coefficient (n = 20). ^{b} Standard deviation of residuals.

Table 2 Multiple linear regression predictions for binary mixtures of DDAS and NAS

	Actual concentration/μg ml^{-1}		Relative error (%)	
DDAS : NAS analyte ratio	DDAS	NAS	DDAS	NAS
1 : 12	0.2	2.4	1.8	−2.7
1 : 10	0.2	2.0	−0.2	−3.1
1 : 5	0.2	1.0	1.0	0.2
2 : 5	0.2	0.5	3.2	2.5
1 : 2	1.0	2.0	0.1	1.3
4 : 5	0.8	1.0	−3.9	3.3
1 : 1	1.0	1.0	1.0	0.8
5 : 2	1.0	0.8	1.9	4.5
2 : 1	1.0	0.5	0.5	0.8
4 : 1	0.8	0.2	−8.4	−3.9
5 : 1	1.0	0.2	1.0	−5.3
6 : 1	1.2	0.2	2.3	−3.2
7 : 1	1.4	0.2	0.5	5.4

Simultaneous determination of DDAS and NAS in consumer products

The proposed method was applied to the simultaneous determination of DDAS and NAS in two commercially available softeners. In order to determine whether the matrix of these samples interfered with the determination of the surfactants, different volumes of each softener, previously diluted 1000 times with distilled water, were analyzed. The results are shown in Table 3. The matrix of the softeners was found not to interfere with the determination of cationic and nonionic surfactants using the proposed methodology, probably because of the high dilution used in the analyses.

The accuracy of the results obtained was assessed by determining DDAS and NAS in the softeners using the disulfine blue (DBS)[18] and cobaltothiocyanate (CTAS)[19] standard methods, respectively. For this purpose, the softeners were diluted 4 times with distilled water and volumes of 15–20 ml of the dilute solutions were used for analysis. Application of these standard methods required the prior removal of an anionic dye present in the softeners by ion-exchange, as well as evaporation to dryness of effluents, dissolution of the extract in chloroform, separation of cationic and nonionic surfactants in alumina, and organic solvent extraction of the reaction products formed. The results obtained are shown in Table 3. As can be seen, the data provided by mixed micelles methodology and the DSB and CTAS methods were all quite consistent.

Conclusions

Mixed aggregate-based methodology opens up interesting prospects for simple, rapid estimation of binary mixtures of surfactants in formulated products. It can be applied to all types of surfactant (cationic, anionic, nonionic and zwitterionic) provided a suitable dye is used to induce premicellar aggregates. In order to obtain the highest possible sensitivity, premicelles should be formed from ionic surfactants because of their high c.m.c. relative to nonionic ones. The high sensitivity of this methodology (surfactants can be determined at the ng ml^{-1} level) and the resulting high dilution factors required for analysis should avoid most interferences from other components of formulated products. Sample dilution was found to be the sole pretreatment required for analysis. By contrast, the standard methods required several clean-up and separation steps. Additional advantages of the proposed methodology include: (1) responses that are independent of the molecular weight and ethylene oxide units of the surfactants, (2)

Table 3 Determination of surfactants DDAS and NAS in softeners

		[Cationic surfactants]^b (% w/v)		[Nonionic surfactants]^b (% w/v)	
Trade name	Volume sample^a/ml	Proposed method	DBS method	Proposed method	CTAS method
Pallas	1.0	3.2 (0.1)		0.51 (0.04)	
	1.5	3.08 (0.08)		0.53 (0.03)	
	2.0	3.12 (0.09)		0.52 (0.01)	
			3.2 (0.2)		0.56 (0.04)
La Oca	1.5	1.8 (0.1)		0.48 (0.03)	
	2.0	1.73 (0.06)		0.50 (0.01)	
	2.5	1.87 (0.06)		0.51 (0.03)	
			3.0 (0.3)		0.52 (0.02)

^{a} 10 ml of softener diluted to 1 l with distilled water. ^{b} Average of three determinations. Standard error values are given in brackets.

Analyst, 2000, **125**, 1507–1512 **1511**

Feature sets

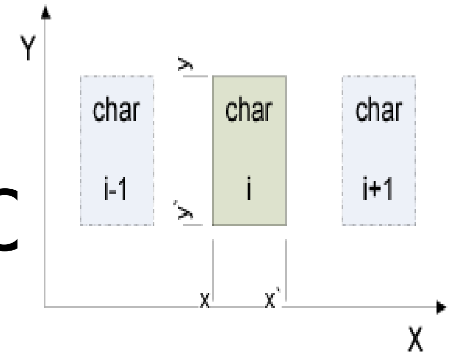
- Orthographic features
  - Vocabulary: the simplest and most obvious feature set
  - InitialCaptical, AllCaptical, FontSize, Font-type, BoldOrNot, HasDot, HasDigital, AllDigital, etc.
- Lexical features
  - TableKwdBeginning, FigureKwdBeginning, ReferenceKwdBeginning, AbstractKwdBeginning, SpecialCharBeginning, DigitalBeginning, SuperscriptBeginning, SubscriptBeginning, LineItself, etc.
- Layout features
  - The most important features
  - *LineNumFromDocTop, LineNumToDocBottom, NumOfTextPieces, LineWidth, CharacterDensity, LargestSpaceInLine, LeftX, rightX, MiddleX, DisToPrevLine, DisToNextLine, etc.*
- Conjunction features
  - window size of features: *-1, 0, 1*

Line Construction

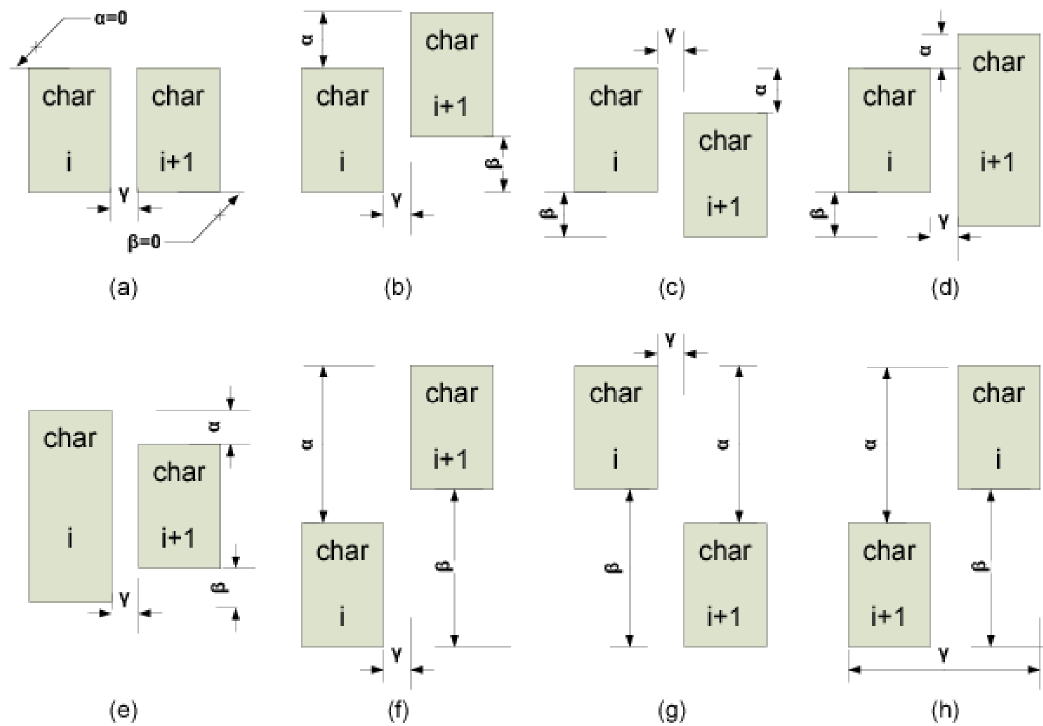
- The unit is a document line
  - instead of the word in the word tagging problem
- Deal with the characters and the related glyph information of PDF files through analyzing the *text operators*
  - Adobe's Acrobat word-finder
  - The PDFlib Text Extraction Toolkit (TET)
    - extracts the text in different levels (character, word, line, paragraph,etc.)
    - only provides the content instead of other style information in all the levels except the character level
- Similar to Xpdf library, we adopt a *bottom-up* approach to reconstruct these characters into words then lines
- Only analyze the coordinate information
- *Font* information is not used to merge texts

The parameter thresholds

- A document $D = U_{k=1}^{n}(P_k)$
- P_k = an aggregation of characters C
- C: {[X, X'], [Y, Y'], W, H, F, T}



P	Definition
α	the vertical distance between two top Y-axis values: $alpha = Y_{i+1} - Y_i$
β	the vertical distance between two bottom Y-axis values: $beta = Y'_{i+1} - Y'_i$
γ	the horizontal distance between these two characters: $\gamma = X_{i+1} - X'_i$
δ	the vertical distance of two characters
θ	the maximal width of the space with a word
η	the maximum vertical distance between two characters in a same line



PENNSTATE

COLLEGE OF INFORMATION SCIENCES AND TECHNOLOGY

Character → Word

Table 1 Quantitative performance of the proposed method for the determination of binary mixtures of DDAS and NAS

	Coefficients of eqns. (3) and (4)			
Measured parameter	β_1 or $\beta_3 \pm s$	$\beta_2 \pm s$	r^a	$s_{y/x}{}^b$
$1 - (C_t^M/C_t)$	0.253 ± 0.007	0.173 ± 0.004	0.997	1.1×10^{-2}
ΔA	0.059 ± 0.001		0.998	1.6×10^{-3}

[a] Correlation coefficient (n = 20). [b] Standard deviation of residuals.

Word → Line

- **The number of text pieces in the 8 lines**
  - ***1, 1, 1, 1,5, 5, 4,1.***

Table 1 Quantitative performance of the proposed method for the determination of binary mixtures of DDAS and NAS

	Coefficients of eqns. (3) and (4)			
Measured parameter	β_1 or $\beta_3 \pm s$	$\beta_2 \pm s$	r^a	$s_{y/x}{}^b$
$1 - (C_t^M/C_t)$	0.253 ± 0.007	0.173 ± 0.004	0.997	1.1×10^{-2}
ΔA	0.059 ± 0.001		0.998	1.6×10^{-3}

[a] Correlation coefficient ($n = 20$). [b] Standard deviation of residuals.

Detecting The Table Boundary Based on The Keywords

- **After sparse line detection and noisy line removal**
- **Combine the *OTHERSPARSE* lines with the table keyword list**
- **Keyword is very useful to separate the consecutive tables**
- **Vertical distance is the key feature to**
  - Construct the sparse areas based on the sparse lines
  - Filter out the noisy sparse lines

– ***Recall* more important than the *precision***

- **Retrieve the long table lines (labeled as non-sparse lines) back to improve the *recall***

Observed Table Metadata Types

- Six mutually exclusive categories:
  - Table environment/typography metadata (document level)
    - Document title, author, etc.
  - Table frame metadata
    - Left, right, top, bottom, all, none, top and bottom, left and right
  - Table affiliated metadata
    - Table caption, footnote, reference text, etc.
  - Table layout metadata
    - Table width, length, number of rows, stub separator, horizontal alignment, etc.
  - Table cell-content metadata
  - Table type metadata
    - *Numerical* and/or *symbolic*

Sample Table Metadata Extracted File

Table 1 Temperature effect on resistance change (ΔR) and response time of tin oxide thin film with 1% CCl_4

Temperature/ °C	$\Delta R^a/\Omega$	$\frac{\Delta R}{(R,\ O_2)}$ (%)	Response time	Reproducibiliy
100	223	5	~ 22 min	Yes
200	270	9	~ 7-8 min	Yes
300	1027	21	< 20 s	Yes
400	993	31	~ 10 s	No

[a] $\Delta R = (R,\ CCl_4) - (R,\ O_2)$.

- <Table>
- <DocumentOrigin>Analyst</DocumentOrigin>
- <DocumentName>b006011i.pdf</DocumentName>
- <Year>2001</Year>
- <DocumentTitle>Detection of chlorinated methanes by tin oxide gas sensors </DocumentTitle>
- <Author>Sang Hyun Park, a ? Young-Chan Son, a Brenda R . Shaw, a Kenneth E. Creasy,* b and Steven L. Suib* acd a Department of Chemistry, U-60, University of Connecticut, Storrs, C T 06269-3060</Author>
- <TheNumOfCiters></TheNumOfCiters>
- <Citers></Citers>
- <TableCaption>Table 1 Temperature effect o n r esistance change (D R) and response timeof tin oxide thin film with 1 % C Cl 4</TableCaption>
- <TableColumnHeading>D R Temperature/ ¡ã C D R a / W (R ,O 2) (%) R esponse time Reproducibiliy </TableColumnHeading>
- <TableContent>100 223 5 ~ 22 min Yes 200 270 9 ~ 7-8 min Yes 300 1027 21 < 2 0 s Yes 400 993 31 ~ 1 0 s No </TableContent>
- <TableFootnote> a D R =(R , CCl 4) - (R ,O 2). </TableFootnote>
- <ColumnNum>5</ColumnNum>
- <TableReferenceText>In page 3, line 11, … Film responses to 1% CCl4 at different temperatures are summarized in Table 1……</TableReferenceText>
- <PageNumOfTable>3</PageNumOfTable>
- <Snapshot>b006011i/b006011i_t1.jpg</Snapshot>
- </Table>

Research Issues

- Crawling documents with tables
- Extracting tables from a document
- No standard table representation
- <u>Table metadata Indexing</u>
- <u>Table ranking</u>
  - Current ranking schemes are inadequate and not designed for table search
- <u>Result interface</u>

Our Approach

- Design and evaluate a novel table ranking algorithm
- Rank tables by rating the <query, table> pairs, instead of the <query, document> pairs
  - prevents a lot of false positive hits for table search, which frequently occur in current web search engines
- Use machine learning methods and heuristics to automatically …. tables
  - Identify, Extract, Represent, Index
- Use standard open sources tools for indexing (Lucene) and extraction (pdfbox)
- Modular design

Our Contribution

- An novel first time table ranking algorithm -- TableRank
- A tailored table term vector space
- An innovative table term weighting scheme – TTF-ITTF
  - Aggregating impact factors from three levels: the term, the table, and the document
- Consider and index table referenced texts, term locations, and document backgrounds
- Design and implement a table search engine, TableSeer, to evaluate the TableRank and compare with popular web search engines

Related Work

- Search table content
  - TINTIN system [1]: table caption and table entries
  - Hu et. Al [2]: man-machine dialog to access the table data
  - Pyreddy et. Al [3]: associates tables with QA
- Table representation
  - Xinxin Wang [4]: conceptual model describing the table structure
  - Table markup: XHTML, OASIS
  - (our contribution) Integrating table structure and layout information, as well as the table-related information, and the document background information
- Table Extraction
  - Automata extraction of table ontologies
  - Previous focus primarily on HTML documents or Images
  - (our contribution) Focus on untagged documents, e.g., PDF documents

Table ranking

To our knowledge, no existing work on table ranking

- Existing ranking schemes are not designed for table search. Typical techniques includes …
  - the similarity of a query and a whole PAGE, as well as the overall page quality
  - Term weighting: vector space model (Baeza-Yates & Ribeiro-Neto 1999) and TFIDF (G. Salton 1988)
  - PageRank (Sergey Brin 1999)
- Our contribution: TableRank algorithm
  - Considering features of both the table and the document it appears in
  - Uses some important but ignored features, e.g,: the referenced text of tables
  - Aggregates features to determine the final rank

Observed Table Metadata Types

- Six mutually exclusive categories:
  - Table environment/typography metadata (document level)
    - Document title, author, etc.
  - Table frame metadata
    - Left, right, top, bottom, all, none, top and bottom, left and right
  - Table affiliated metadata
    - Table caption, footnote, reference text, etc.
  - Table layout metadata
    - Table width, length, number of rows, stub separator, horizontal alignment, etc.
  - Table cell-content metadata
  - Table type metadata
    - *Numerical* and/or *symbolic*

Table Metadata Used

- For indexing, use the following metadata
  - DocumentOrigin
  - DocumentName
  - Year
  - DocumentTitle
  - DocumentAuthor
  - TheNumOfCiters
  - Citations (replaces citers)
  - TableCaption
  - TableCaptionHeading
  - TableContent
  - TableFootnote
  - ColumnNum
  - TableReferenceText
  - PageNumOfTable
  - Snapshot
  - ……

TableRank

- **The similarity between a <table, query> pair: the cosine of the angle between vectors**

$$cos(tb_j, Q) = \frac{\sum_{i=1}^{s} w_{i,j,k} w_{i,q,k}}{|tb_j||Q|}$$

- **Tailored term vector space => table vectors:**
- Query vectors and table vectors, instead of document vectors
- **Novel term weighting schemes:**
- *TTF – ITTF:* (Table Term Frequency-Inverse Table Term Frequency)
- Efficiently prevent false positive hits
- Consider the term position and document background features

$$w_{i,j,k} = w_{i,j,k,TermLevel} * TLB_{i,j} * DLB_j$$

- *TLB: Table Level Boost* Factors (e.g., table frequency)
- DLB: *Document Level Boost* factors (e.g., journal/proceeding order, document citation)

Table Vector Space Representation

Table 1: The Vector Space for Tables and Queries

	$m_1(MW_1)$			$m_2(MW_2)$			...	$m_k(MW_k)$			TLB	DLB
	$t_{1,1}$	...	$t_{x,1}$	$t_{1,2}$	...	$t_{y,2}$	...	$t_{1,k}$	...	$t_{z,k}$	...	...
$tb1$	$w_{1,1,1}$	...	$w_{x,1,1}$	$w_{1,1,2}$	...	$w_{y,1,2}$	...	$w_{1,1,k}$	...	$w_{z,1,k}$	...	...
$tb2$	$w_{1,2,1}$	...	$w_{x,2,1}$	$w_{1,2,2}$	...	$w_{y,2,2}$	...	$w_{1,2,k}$	...	$w_{z,2,k}$	...	...
$\vdots$	...	$\vdots$	...	...	$\vdots$	...	$\vdots$	...	$\vdots$	...	...	...
tb_b	$w_{1,b,1}$	...	$w_{x,b,1}$	$w_{1,b,2}$	...	$w_{y,b,2}$	...	$w_{1,b,k}$	...	$w_{z,b,k}$	...	...
Q	$w_{1,q,1}$			$w_{1,q,2}$			...	$w_{1,q,k}$			...	...
$ITTF$	...	...	...		...	...	...	...	...	...	...	...

The Term Level

$$w_{i,j,k}^{TermLevel} = TTFITTF_{i,j,k} = TTF_{i,j,k} * ITTF_{i,j,k}$$

$$TTF_{i,j,k} = (p + (1-p) * \frac{tf_{i,j,k}}{tf_{i\_k}}) * MW_k$$

$$ITTF_{i,j,k} = \log_2(\frac{b}{IDF_{i,j,k}}) + 1$$

- **A term occurring in a few tables is likely to be a better discriminator than a term appearing in most or all tables**
- **Similar to document abstract, table metadata and table query should be treated as semi-structured text**
- Not complete sentences and express a summary
- P = 0.5 (G. Salton 1988)
- **b is the total number of tables**
- **IDF(ijk): the number of tables that term t(i) occurs in the matadata m(k)**

Table Level Boost and Document Level Boost

$$TLB_{i,j} = B_{tbf} + B_{trt} + (r * B_{tp})$$

$$TLB_{i,j} = \begin{cases} \frac{tbf_{i,j}}{tbf_j} * (log_2 \frac{b}{b_i} + 1) + nlr_j + B_{tp} & \text{if } r = 1 \\ \frac{tbf_{i,j}}{tbf_j} * (log_2 \frac{b}{b_i} + 1) + nlr_j & \text{if } r{=}0 \end{cases}$$

B_{tbf} is the boost value of the *table frequency*
B_{trt} is the boost value of the *table reference text* (e.g., the normalized length), and B_{tp} is the boost value of the *table position*. *r* is a parameter, which is 1 if users specify the table position in the query. Otherwise, *r* = 0.

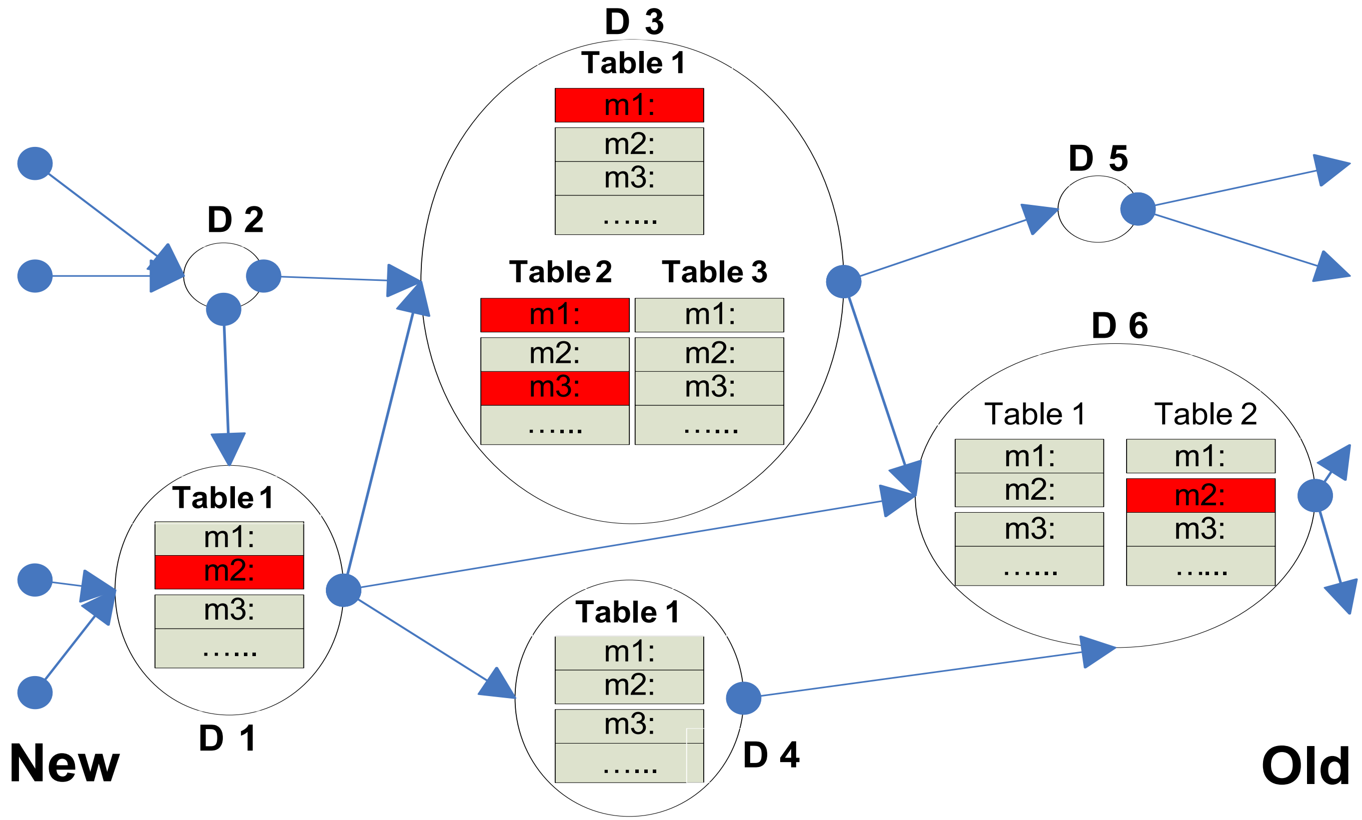
$$DLB_j = IV_j = IC_j * DO_j * DF_j = (\frac{\sum_{v=1}^{x} IV_v}{x}) * DO_j * DF_j$$

$IV_{j:}$ document *Importance Value (IV).* If a table comes from a document with a high *IV* , all the table terms of this document should get a high document level boost.
IC_j: the inherited citation value (*ICj*)
$DO_{j:}$ source value (the rank of the journal/conference proceeding)
DF_j: document freshness

Table citation network

- Similar to the *PageRank* network
  - Documents construct a network from the citations
  - The “incoming links” – the documents that cite *the document win which the table is located*
  - Exponential decay used to deal with the impact of the propagated importance
- Unlike the *PageRank* network
  - *Directed Acyclic Graph*
  - *Importance Value (IV)* of a document not decreased as the number of citations increases
  - *IV* not divided by the number of outbound links
- A document may have multiple, one, or no tables
- Each table is consisted as a set of metadata
- Same keywords may appear in different metadata in different tables

An Example of the Citation Network



Parameter Settings

- Metadata Weight (MW_k)
  - proportional to the occurrence frequency of the meaningful keywords in the metadata
- Meaningful keywords: representative terms which are always selected to construct queries
- Determining each metadata's weight based on a statistical study of the keyword distribution over different metadata

Metadata Names	Metadata Weight
Document Title	4.60
Table Column Heading	4.40
Table Caption	4.00
Table Reference Text	1.75
Table Cells	1.25
Author	1.00
Table Footnote	1.00

Setting the Metadata Weights based on Keyword Distribution

- Initially randomly select sets of sample tables
- For each table, a *Term Dictionary* is generated without the *stop list words*
  - The *Term Dictionary* is created in descending order according to term-occurrence frequency
- Identify the top *k* meaningful terms from the ordered *Term Dictionary* in order to construct a *Popular Term List*
  - Although different tables have different *Popular Term Lists*, the term distributions over the metadata should be similar
- For each metadata, a summation is made of the term-occurrence frequency of all the terms in the *Popular Term List*
  - The larger the summation, the higher the weight is given to the metadata

Parameters for Document Origination (DO)

- In each research field, scholarly journals or conferences are scored and ranked based on the opinions of domain experts
- *CiteSeer* gives an estimation of the impact ranking for the computer science publications
- *Wikipedia* estimates for chemistry papers
  - http://en.wikipedia.org/wiki/List of scientific journals inchemistry/
- A comprehensive journal impact factor list spanning the years 2002-2004 for all the fields can be found in *CNCSIS*
  - http://www.cncsis.ro/PDF/IF 2004.pdf

Parameter Setting of Document Freshness (DF)

- Age of a document
- More weight credits are assigned to fresher documents
  - Recently published documents always reflect the latest research status and results
  - Limits “The Rich Get Richer” phenomenon caused by the boosting of citation frequency
    - Issues with the bias to “old and famous” documents

TableRank - ranking tables in search

•**The similarity between a <table, query> pair: the cosine of the angle between vectors**

$$cos(tb_j, Q) = \frac{\sum_{i=1}^{s} w_{i,j,k} w_{i,q,k}}{|tb_j||Q|}$$

•**Tailored term vector space => table vectors:**

•Query vectors and table vectors, instead of document vectors

•**Novel term weighting schemes:**

•*TTF – ITTF:* (Table Term Frequency-Inverse Table Term Frequency)

$$w_{i,j,k} = w_{i,j,k,TermLevel} * TLB_{i,j} * DLB_j$$

•*TLB: Table Level Boost* Factors (e.g., table frequency)

•DLB: *Document Level Boost* factors (e.g., journal/proceeding order, document citation)

Table Vector Space Representation

Table 1: The Vector Space for Tables and Queries

	$m_1(MW_1)$			$m_2(MW_2)$			...	$m_k(MW_k)$			TLB	DLB
	$t_{1,1}$	...	$t_{x,1}$	$t_{1,2}$	...	$t_{y,2}$	...	$t_{1,k}$	...	$t_{z,k}$	...	...
$tb1$	$w_{1,1,1}$	...	$w_{x,1,1}$	$w_{1,1,2}$	...	$w_{y,1,2}$	...	$w_{1,1,k}$	...	$w_{z,1,k}$	...	...
$tb2$	$w_{1,2,1}$	...	$w_{x,2,1}$	$w_{1,2,2}$	...	$w_{y,2,2}$	...	$w_{1,2,k}$	...	$w_{z,2,k}$	...	...
$\vdots$	...	$\vdots$	...	...	$\vdots$	...	$\vdots$	...	$\vdots$	...	...	...
tb_b	$w_{1,b,1}$	...	$w_{x,b,1}$	$w_{1,b,2}$	...	$w_{y,b,2}$	...	$w_{1,b,k}$	...	$w_{z,b,k}$	...	...
Q	$w_{1,q,1}$			$w_{1,q,2}$			...	$w_{1,q,k}$			...	...
$ITTF$	...	...	...		...	...	...	...	...	...	...	...

Query Interface Design

Many design issues

- How are tables presented
- What in tables should be presented
- Ranking based on what attributes
- Links to actual tables and documents

Data set

- Tables PDF scientific documents
  - Wang's table ground truth database focuses on the web tables
  - No benchmark dataset exists in PDF tables
- Diverse journals and proceedings
  - Chemical scientific digital libraries (RCS) (H)
  - Computer science proceedings (S)
  - Archeology journals (A)
- 300 randomly selected pages in three fields
  - Line numbers: 10177, 13151, and 9641
  - Hold-out method to do the training and testing

Text extraction from PDFs

- PDF document content stream contains
  - Texts, graphics, images, etc.
  - Object overlapping problem happens frequently
  - Identified objects/structures are still too high level for our problem
- Most table-related application focuses on the text, instead of the borderlines
- Most tables are text tables
- PDF converters
  - Xpdf, PDF2TEXT, PDFBOX, Text extracting tool (TET), PDFTEXTSTREAM

Performance of line construction

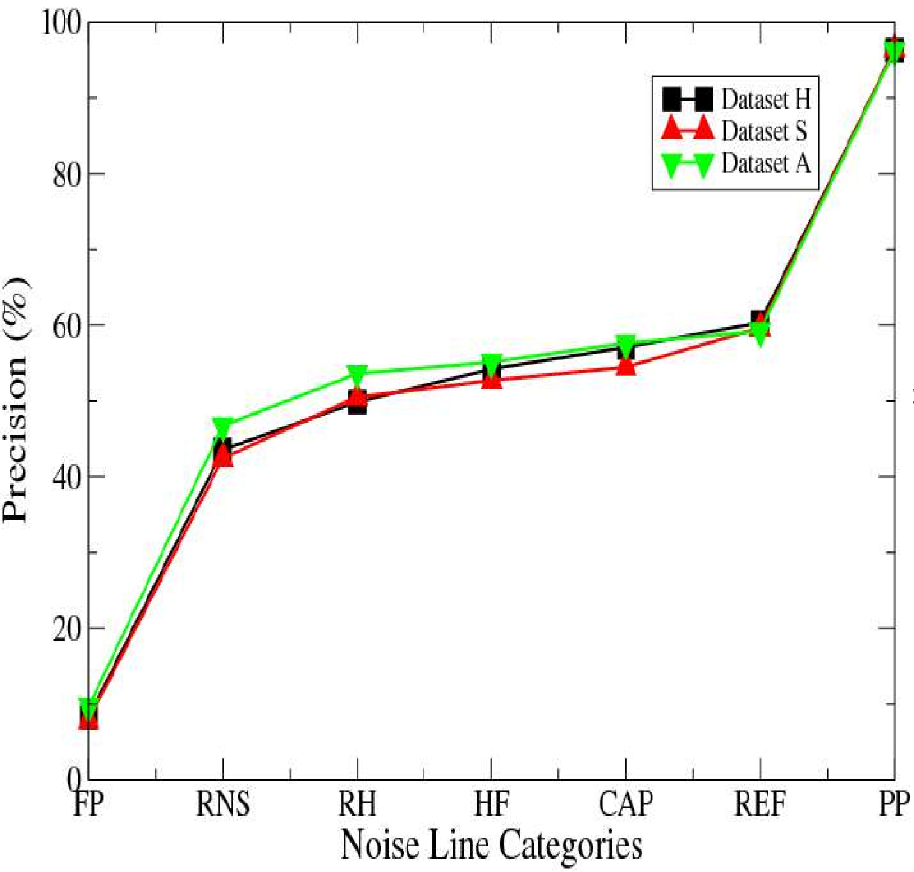
- *T*: the number of the total lines
- *C*: the number of constructed lines that do not have any error
- Error lines
  - Include texts that should not belong to it
  - Miss a part of the text
- Accurately constructed 99.057% lines
- Error reasons:
  - the inherited coordinate error from the text extractors
  - Superscripts/subscripts

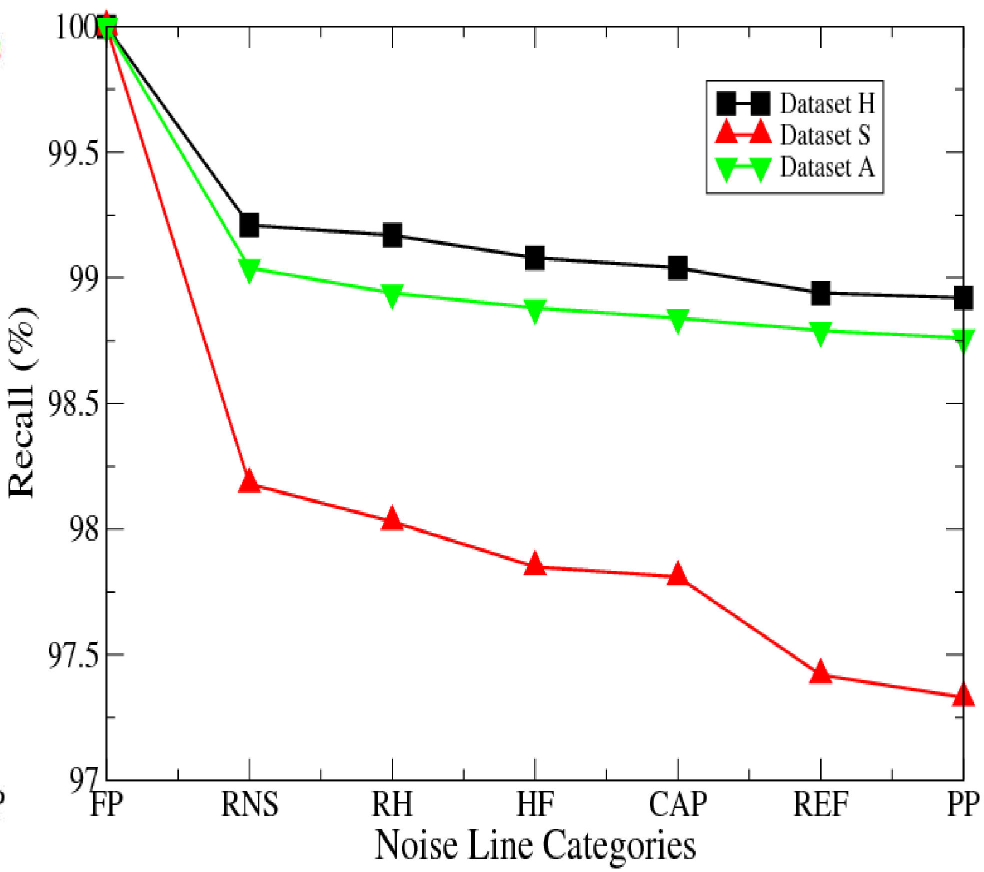
Performance of sparse line detection

- User study on 20 PDF pages
- Recall: ts/(ts + tn); Precision: ts/(ts+tp)
- Two goals of our method
  - Removing non-sparse lines as much as possible
  - Keeping true table lines as much as possible
- In dataset H, A, S
  - 84.63% are labeled as non-sparse lines
  - 44.23% sparse lines are real table lines
  - 95.35% table lines are in the sparse line set
- Reasons
  - Long cross-column table cells
  - The inherited test missing problem

Performance of noise removal

- Precision: tl/sp; recall: tl/(tl+to)





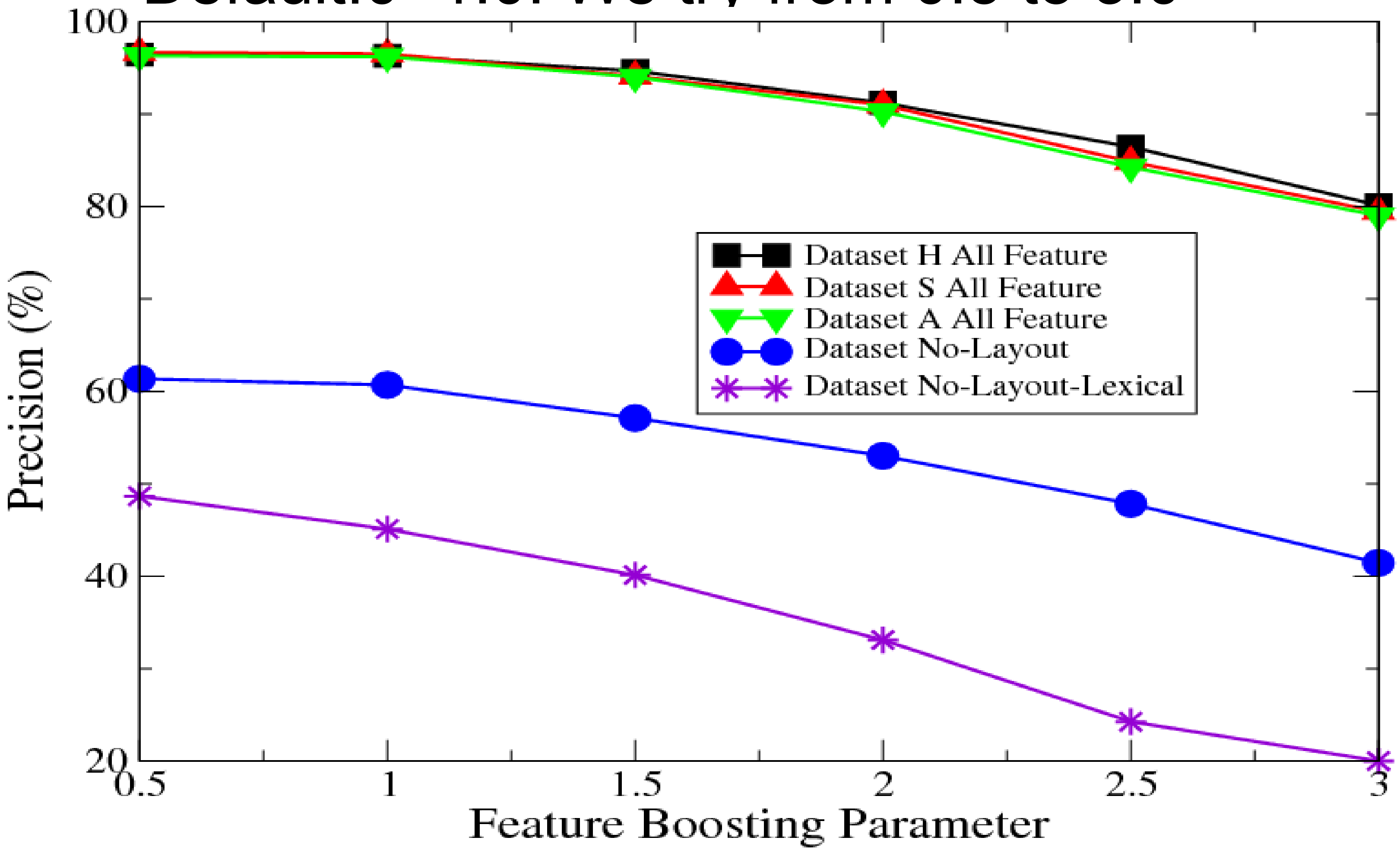
Impact effects of feature sets

- Precision: A/ (A+C);
- Recall: A/ (A+B)
- F-measure: (2*Recall*Precision)/(Recall+Precision)

feature sets, and datasets	Recall	Precision
CRF, Orthographic, *H*	42.18%	44.96%
CRF, Orthographic, *S*	41.66%	45.14%
CRF, Orthographic, *A*	40.89%	45.16%
CRF, Orthographic+Lexical, *H*	61.22%	61.66%
CRF, Orthographic+Lexical, *S*	59.30%	59.81%
CRF, Orthographic+Lexical, *A*	59.98%	60.58%
CRF, Orthographic+Lexical+Layout, *H*	**98.92%**	**96.28%**
CRF, Orthographic+Lexical+Layout, *S*	97.33%	96.49%
CRF, Orthographic+Lexical+Layout, *A*	98.76%	96.20%

Impact effect of parameters

- Feature boosting parameter
  - Default:θ=1.0. We try from 0.5 to 3.0



Impact effect of different techniques

- Compare with our rule-based method
  - SVM improves the performance by 30.36%
  - CRF improves the performance by 54.90%
  - Ng et. Al achieved the best results with C4.5

Method and datasets	F-measure
Rule-based method, $H + S + A$	91.93%
CRF, $H{+}S{+}A$	**96.36%**
SVM linear, $H{+}S{+}A$	**94.38%**
Max Ent in [19]	88.7%
CRF Binary in [19]	91.2%
CRF Continuous in [19]	91.8%
C4.5 in [16]	< 95%
Bp in [16]	< 91%
Det in [16]	< 70%

Document Data

- Current focus: tables in scientific documents in PDF format
- Three sources
  - the scientific digital libraries
    - e.g., Royal Society of Chemistry
  - the web pages of research scientists
    - http://www.chem.ucla.edu/VL/Academic.html
  - the CiteSeer archive

- For experiments number of collected PDF docs: 10,000
  - More than 20 journals and conferences
  - A variety of research fields
    - chemistry, biology, computer science, etc
  - Years 1990 to 2006
  - More than 70% of the papers have tables
  - Most of them have multiple tables

Experimental Results

- Table Detection
  - Five-user study on 200 PDF documents: precision -- 100%, recall -- 93.5%
- Table Metadata Extraction Results
  - 371 tables, >95% in both precision and recall
- Table Ranking Results
  - a "*gold standard*" to define the "*correct*" ranking based on human judgment
  - *pairwise accuracy* to evaluate the ranking quality
  - two methods to set up a common testbed:
    - manual "bottom-up" method
      - Query from search engines yield pdfs with tables
      - Tableseer processes pdfs and compares ranking
    - custom search engine method
      - Google custom search for same seed
      - Same as above

Document Data

- Current focus: tables in scientific documents in PDF format
- Three sources
  - scientific digital libraries
    - e.g., Royal Society of Chemistry
  - web pages of research scientists
    - http://www.chem.ucla.edu/VL/Academic.html
  - CiteSeer archive

- For experiments number of collected PDF docs: 10,000
  - More than 20 journals and conferences
  - A variety of research fields
    - chemistry, biology, computer science, etc
  - Years 1990 to 2006
  - More than 70% of the papers have tables
  - Most of them have multiple tables

Experimental Results

- Table Ranking Results
  - A "*gold standard*" to define the "*correct*" ranking based on human judgment
  - *Pairwise accuracy* to evaluate the ranking quality
  - Two methods to set up a common test bed:
    - Manual "bottom-up" method
      - Query from search engines yield pdfs with tables
      - Tableseer processes pdfs and compares ranking
    - Custom search engine method
      - Google custom search for same seed
      - Same as above

Ranking	The Method to set-up the test-bed	Accuracy (%)
Google	Custom search engine	51.8
Google Scholar	bottom-up method	52.72
CiteSeer	bottom-up method	55.35
TableSeer	Both methods	69.61

Experimental Results

- Factor Influence in TableRank
  - How well each impact factor performs and how heavily each of them influence the final ranking?
  - Implementing TableRank algorithm on …
    - Each factor independently
    - Varied combination by incrementally adding one factor

Impact Factors	TFIDF	TTFITTF without MW	TTFITTF with MW	TLB	DLB	All
Accuracy(%)	50.19	61.46	63.55	29.60	40.33	69.61

Conclusions

- Designed and built a unique table search engine, TableSeer
  - Define the unit of TableSeer search as a table, not a document
  - Use machine learning methods for metadata extraction
- Devised a unique table ranking algorithm
  - Different tables in a same document may have different rankings
- Observations:
  - The quality of the table metadata extraction is crucial to the table searching performance
  - Term frequency is still the most significant impact factor for ranking
    - Metadata weight is also an important impact factor
  - The number of hits for a query will most likely not be comparable to the number of hits for generic web search

Future Work

- Enhanced table structure analysis and classification
- Design and implement a Dublin Core table metadata ontology
- Improve the performance of the metadata extraction
- Improve the ranking algorithm
- Design and improve the usability of the search engine
- Quantitative study of tables
- Extend to other document formats and search implementations such as CiteSeer^X

References

- [1] P. Pyreddy and W. Croft. Tintin: A system for retrieval in text tables.In *In Proceedings of the Second International Conference on Digital Libraries*, pages 193–200, 1997.
- [2] J. Wang and J. Hu. A machine learning based approach for table detection on the web. In *Proceedings of the 11th Int'l Conf. on World Wide Web (WWW'02)*, pages 242–250, Nov 2002.
- [3] P. Pyreddy and W. Croft. Tintin: A system for retrieval in text tables.In *In Proceedings of the Second International Conference on Digital Libraries*, pages 193–200, 1997.
- [4] X. Wang. Tabular abstraction, editing, and formatting. In *Ph.D. Thesis, Dept. of Computer Science, University of Waterloo*, 1996.

TableSeer beta available at:

http://chemxseer.ist.psu.edu

Comments most welcomed:

yliu@ist.psu.edu
pmitra@ist.psu.edu
giles@ist.psu.edu