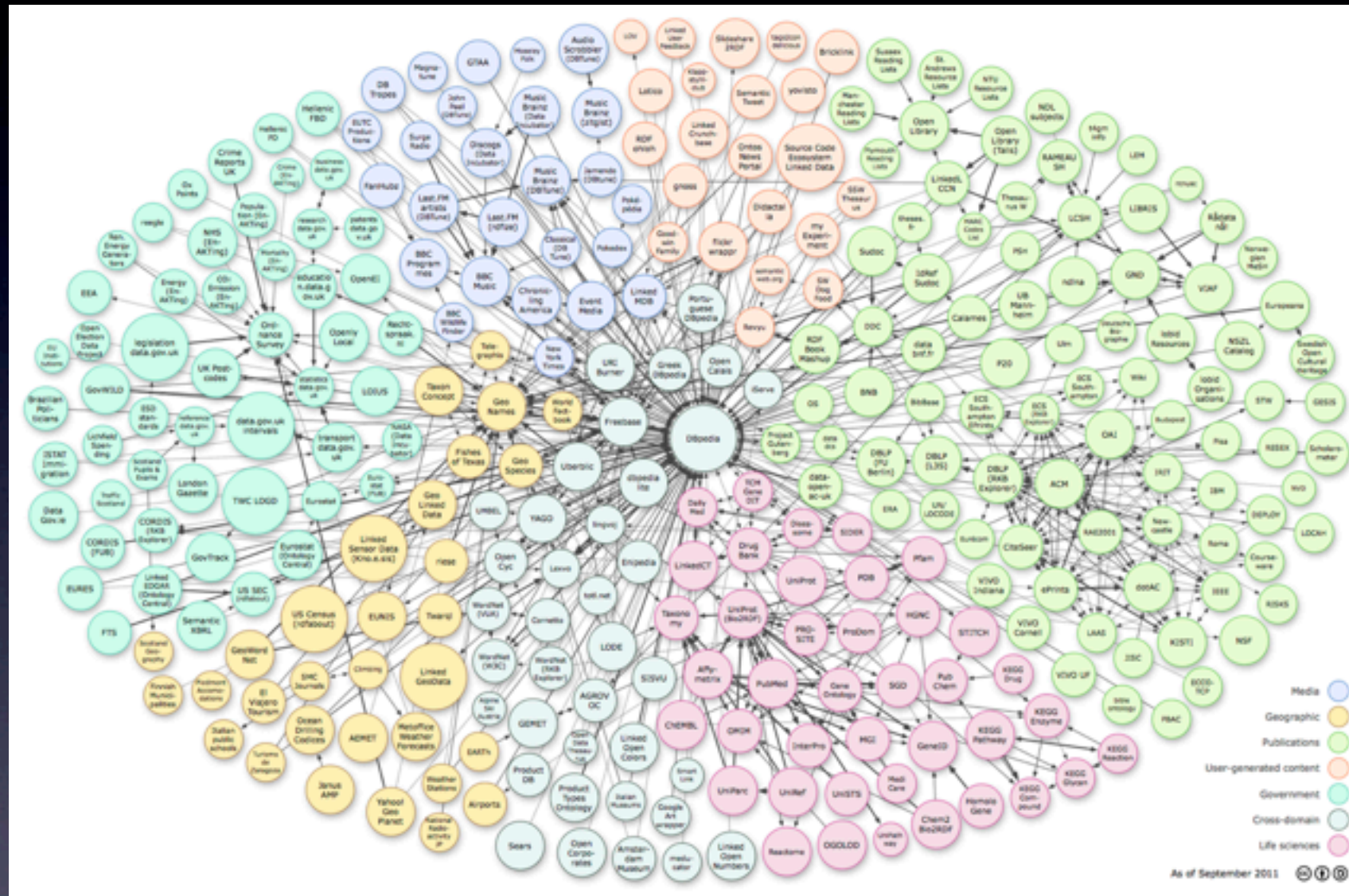


Methods for Exploring and Mining Tables on Wikipedia

Chandra Sekhar Bhagavatula, Thanapon Noraset, Doug Downey
Department of Electrical Engineering & Computer Science
Northwestern University

Web-Data: RDFs, OpenIE, Tables



Web-Data: RDFs, OpenIE, Tables



Object	Property	Value
China	populationCensus	1339724852
India	populationCensus	1210193422
United States	populationCensus	308745538
:	:	:
China	gdpPPPPerCapita	7518
United States	gdpPPPPerCapita	47123



Opportunity

- Search
- Explore
- Mine
- Make sense of large heterogenous sets of relations

Example

Example

- Electricity consumption of countries

Example

- Electricity consumption of countries
- CO₂ emissions of countries

Example

- Electricity consumption of countries
- CO₂ emissions of countries
- How are the two related ? ([WikiTables link](#))

Why Wikipedia Tables

- Datasource: Wikipedia tables
- Why Tables ?
 - Each table encapsulates a non-redundant set of facts
 - Text extractors need to put facts together
- Why Wikipedia ?
 - High Quality - easy to extract and process
 - All tables contain relational data
WebTables achieved 41% precision in extracting relational data (Cafarella et.al.)

WikiTables

WikiTables

- Relevant Join
 - Find relevant additional columns for a given table

WikiTables

- Relevant Join
 - Find relevant additional columns for a given table
- Correlation Mining
 - Automatically uncover interesting relationships

WikiTables

- Relevant Join
 - Find relevant additional columns for a given table
- Correlation Mining
 - Automatically uncover interesting relationships
- Table Search
 - Find relevant tables for a given text query

Outline

- Table Extraction
 - Method
 - Evaluation
- Relevant Join
- Correlation Mining
- Table Search
- Conclusion & Future Work

Table Extraction

Ruling date	Sport	Event	Nation	Gold	Silver	Bronze	Total
August 15, 2008	Shooting	Men's 10 metre air pistol	 North Korea (PRK)			-1	-1
			 United States (USA)			+1	+1
		Men's 50 metre pistol	 North Korea (PRK)		-1		-1
			 China (CHN)		+1	-1	
			 Russia (RUS)			+1	+1



Table Extraction

Ruling date	Sport	Event	Nation	Gold	Silver	Bronze	Total
August 15, 2008	Shooting	Men's 10 metre air pistol	 North Korea (PRK)			-1	-1
			 United States (USA)			+1	+1
		Men's 50 metre pistol	 North Korea (PRK)		-1		-1
			 China (CHN)		+1	-1	
			 Russia (RUS)			+1	+1

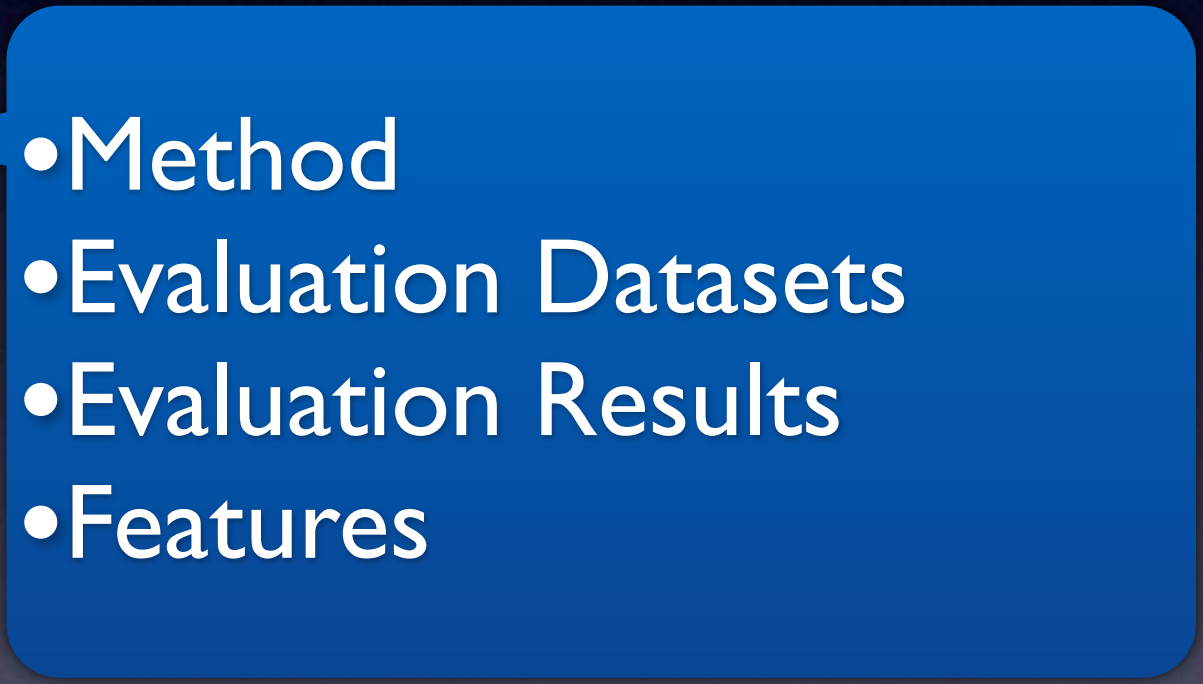


Ruling Date	Sport	Event	Nation	Gold	Silver	Bronze	Total
August 15, 2008	Shooting	Men's 10 metre air pistol	North Korea (PRK)			-1	-1
August 15, 2008	Shooting	Men's 10 metre air pistol	United States (USA)			+1	+1
August 15, 2008	Shooting	Men's 50 metre pistol	North Korea (PRK)		-1		
August 15, 2008	Shooting	Men's 50 metre pistol	China (CHN)		+1	-1	
August 15, 2008	Shooting	Men's 50 metre pistol	Russia (RUS)			+1	+1

Table Extraction Evaluation

- 1.4 Million tables on 643,000 pages
- Extraction evaluation:
 - Number of tables in test set: 111
 - Tables Extracted: 82 (recall = 0.74)
 - Number of cells in 82 tables: 6404
 - Cells extraction precision: 6383 (precision = 99.7%)

Outline

- Table Extraction
 - Relevant Join
 - Correlation Mining
 - Table Search
 - Conclusion & Future Work
- 
- Method
 - Evaluation Datasets
 - Evaluation Results
 - Features

Relevant Join

Country	GDP
United States	15,090,000
China	7,298,000
Japan	5,869,000
Germany	3,577,000
⋮	⋮
Nieu	10.01

Relevant Join

Country	GDP
United States	15,090,000
China	7,298,000
Japan	5,869,000
Germany	3,577,000
⋮	⋮
Nieu	10.01

Country	Population
China	1,347,350,000
India	1,210,193,422
United States	314,249,000
Indonesia	237,641,326
⋮	⋮
Nieu	1000

Relevant Join

Country	GDP
United States	15,090,000
China	7,298,000
Japan	5,869,000
Germany	3,577,000
⋮	⋮
Nieu	10.01



Country	Population
China	1,347,350,000
India	1,210,193,422
United States	314,249,000
Indonesia	237,641,326
⋮	⋮
Nieu	1000

Relevant Join

Country	GDP
United States	15,090,000
China	7,298,000
Japan	5,869,000
Germany	3,577,000
⋮	⋮
Nieu	10.01



Country	Population
China	1,347,350,000
India	1,210,193,422
United States	314,249,000
Indonesia	237,641,326
⋮	⋮
Nieu	1000

Country	GDP	Population
United States	15,090,000	314,249,000
China	7,298,000	1,347,350,000
Japan	5,869,000	127,570,000
Germany	3,577,000	81,844,000
⋮	⋮	⋮
Nieu	10.01	1000

Relevant Join

Country	GDP
United States	15,090,000
China	7,298,000
Japan	5,869,000
Germany	3,577,000
⋮	⋮
Nieu	10.01



Country	Population
China	1,347,350,000
India	1,210,193,422
United States	314,249,000
Indonesia	237,641,326
⋮	⋮
Nieu	1000

Country	GDP	Population
United States	15,090,000	314,249,000
China	7,298,000	1,347,350,000
Japan	5,869,000	127,570,000
Germany	3,577,000	81,844,000
⋮	⋮	⋮
Nieu	10.01	1000

Pearson's r
= 0.429

Relevant Join

- Novel task: Find *relevant* columns that can be added to a table
- Difficulties:
 - How do we find similar columns from distinct tables ?
 - How do we identify relevant columns ?

Relevant Join Method

- Finding similar columns from distinct tables
 - Use heuristics to scale:
 1. Non-numeric
 2. More than 4 rows
 3. Average string length ≥ 4
 - 340M pairs of similar columns

Relevant Join Method

- Identifying relevant columns to add:
 - Use Machine Learning methods
 - Logistic regression classifier
 - Coordinate Ascent ranker

Relevant Join Evaluation Datasets

- Training:
 - 82 Distinct tables
 - 4 users posted 593 ratings on a scale of 1 to 5
- Testing:
 - 20 Wikipedia pages selected randomly
 - 2 raters rate 4 columns per table on a scale of 1 to 5
 - Inter-Annotator Agreement: 0.73

Relevant Join Results

Model	Columns Added	Accuracy	DCG@4	nDCG@4
BASE	423	0.29	11.38	0.43
WikiTables-SR	414	0.43	13.71	0.48
WikiTables	410	0.62	30.71	0.76

Relevant Join Results

Model	Columns Added	Accuracy	DCG@4	nDCG@4
BASE	423	0.29	11.38	0.43
WikiTables-SR	414	0.43	13.71	0.48
WikiTables	410	0.62	30.71	0.76

- BASE: Rank columns in descending order of *MatchPercentage* (percentage overlap)

Relevant Join Results

Model	Columns Added	Accuracy	DCG@4	nDCG@4
BASE	423	0.29	11.38	0.43
WikiTables-SR	414	0.43	13.71	0.48
WikiTables	410	0.62	30.71	0.76

- BASE: Rank columns in descending order of *MatchPercentage* (percentage overlap)
- WikiTables-SR: Ignores the Semantic Relatedness feature in the ML models

Relevant Join Results

Model	Columns Added	Accuracy	DCG@4	nDCG@4
BASE	423	0.29	11.38	0.43
WikiTables-SR	414	0.43	13.71	0.48
WikiTables	410	0.62	30.71	0.76

- BASE: Rank columns in descending order of *MatchPercentage* (percentage overlap)
- WikiTables-SR: Ignores the Semantic Relatedness feature in the ML models
- **WikiTables:** Our system

Relevant Join Results

Model	Columns Added
BASE	423
WikiTables-SR	414
WikiTables	410

- BASE: Rank columns in descending order of *MatchPercentage* (percentage overlap)
- WikiTables-SR: Ignores the Semantic Relatedness feature in the ML models
- **WikiTables:** Our system

Relevant Join Results

Model	Columns Added	Accuracy
BASE	423	0.29
WikiTables-SR	414	0.43
WikiTables	410	0.62

- BASE: Rank columns in descending order of *MatchPercentage* (percentage overlap)
- WikiTables-SR: Ignores the Semantic Relatedness feature in the ML models
- **WikiTables:** Our system

Relevant Join Results

Model	Columns Added	Accuracy	DCG@4
BASE	423	0.29	11.38
WikiTables-SR	414	0.43	13.71
WikiTables	410	0.62	30.71

- BASE: Rank columns in descending order of *MatchPercentage* (percentage overlap)
- WikiTables-SR: Ignores the Semantic Relatedness feature in the ML models
- **WikiTables:** Our system

Relevant Join Results

Model	Columns Added	Accuracy	DCG@4	nDCG@4
BASE	423	0.29	11.38	0.43
WikiTables-SR	414	0.43	13.71	0.48
WikiTables	410	0.62	30.71	0.76

- BASE: Rank columns in descending order of *MatchPercentage* (percentage overlap)
- WikiTables-SR: Ignores the Semantic Relatedness feature in the ML models
- **WikiTables:** Our system

Relevant Join Results

Model	Columns Added	Accuracy	DCG@4	nDCG@4
BASE	423	0.29	11.38	0.43
WikiTables-SR	414	0.43	13.71	0.48
WikiTables	410	0.62	30.71	0.76

- BASE: Rank columns in descending order of *MatchPercentage* (percentage overlap)
- WikiTables-SR: Ignores the Semantic Relatedness feature in the ML models
- **WikiTables:** Our system

Semantic Relatedness

- Estimate of relatedness of Wikipedia concepts
- Similar pages share a large fraction of in-links
- Only defined for Wikipedia concepts
 - Cannot be defined for WebTables

Relevant Join Results

Model	Columns Added	Accuracy	DCG@4	nDCG@4
BASE	423	0.29	11.38	0.43
WikiTables-SR	414	0.43	13.71	0.48
WikiTables	410	0.62	30.71	0.76

- BASE: Rank columns in descending order of *MatchPercentage* (percentage overlap)
- WikiTables-SR: Ignores the Semantic Relatedness feature in the ML models
- **WikiTables:** Our system

Relevant Join Features

- Top 3 features:
 - Binary: Is the candidate column numeric
 - Numeric: Semantic Relatedness
 - Numeric: Number of values per cell in the added column

Relevant Join Features

Feature	Correlation with Rating
Target Column is numeric	0.285
Semantic Relatedness	0.253
Number of values per cell in added column	-0.231
Ratio of distinct values in source column	0.209
Index of source column from left of table	-0.202
Match Percentage	0.167
Number of out links	-0.161
Number of in-links	-0.157
Fraction of numeric columns in source table	0.134

Outline

- Table Extraction
- Relevant Join
- Correlation Mining
 - Method
 - Evaluation Datasets
 - Evaluation Results
 - Features
- Table Search
- Conclusion & Future Work

Correlation Mining

Correlation Mining

- Novel task: Find *interesting* correlations between numeric columns

Correlation Mining

- Novel task: Find *interesting* correlations between numeric columns
- Automatically uncover interesting or surprising statistical relationships

Correlation Mining

- Novel task: Find *interesting* correlations between numeric columns
- Automatically uncover interesting or surprising statistical relationships
 - Country GDP highly correlated with Imports ($r = 0.941$)

Correlation Mining

- Novel task: Find *interesting* correlations between numeric columns
- Automatically uncover interesting or surprising statistical relationships
 - Country GDP highly correlated with Imports ($r = 0.941$)
 - Total CO₂ emissions not at all related to per-capita CO₂ emissions ($r = -0.006$)

Correlation Mining

Evaluation

- Dataset
 - 100 labelled pairs of numeric columns
 - Evaluation by cross-validation
- Metrics
 - Accuracy
 - F1

Correlation Mining

Evaluation

Model	Accuracy	F1
Majority class model	66%	0.525
SVM Classifier	65%	0.52
1-Nearest Neighbor	64%	0.62
3-Nearest Neighbor	71%	0.657
Logistic Regression	71%	0.661

Correlation Mining

Features

- Correlation Features
 - Correlation Coefficient between numeric columns
 - Standard Error of the correlation coefficient
 - Sample size
 - zScore
- Relevance Features
 - Semantic Relatedness

Outline

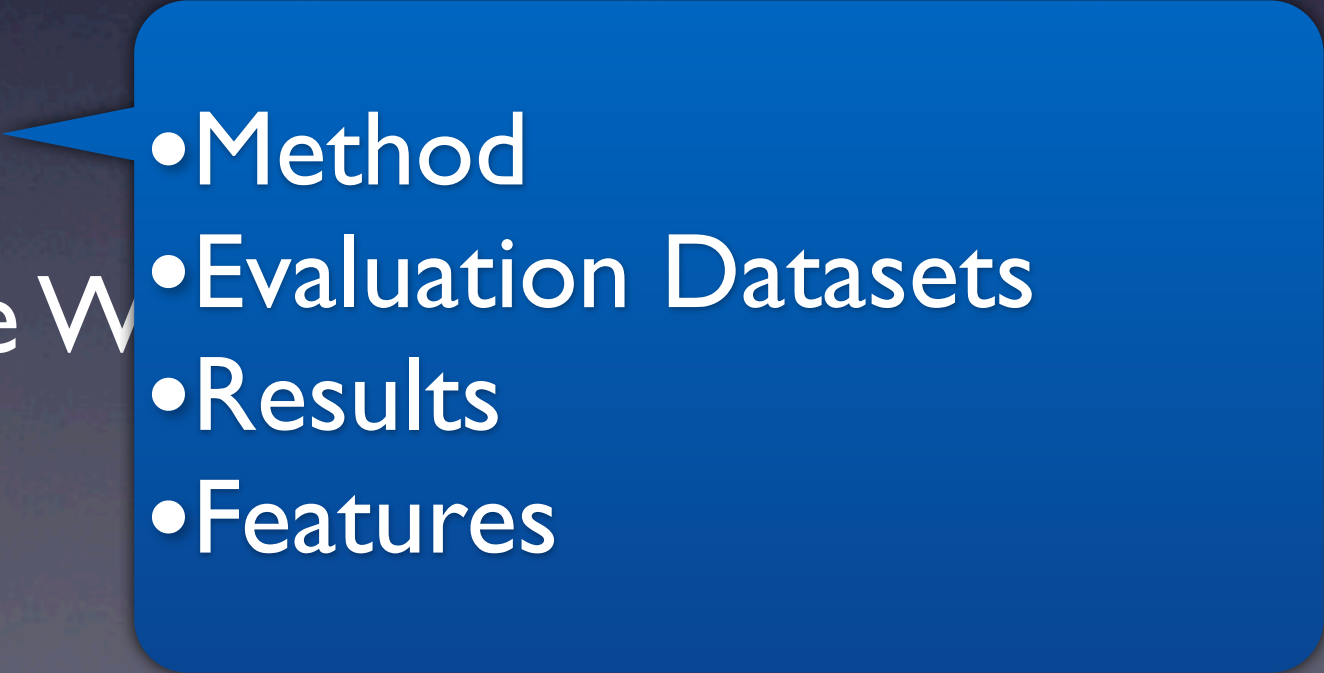
- Table Extraction
 - Relevant Join
 - Correlation Mining
 - Table Search
 - Conclusion & Future Work
- 
- Method
 - Evaluation Datasets
 - Results
 - Features

Table Search

- TableSearch
 - Input: Text query + Table corpus
 - Output: Ranked list of tables, in decreasing order of relevance to query

Table Search Method

- For a given query, leverage an existing search engine to get top 30 Wikipedia pages with tables
- Coordinate Ascent Ranking model ranks these tables

Table Search Dataset

- Training Set
 - 5 queries; ~350 labeled examples
- Test Set (Same as Venetis et.al.)
 - 100 queries of the form (C, P)
 - C: class of instances
 - P: property associated with these instances

Table Search Dataset

- Rating Scheme (as Venetis.et.al):
 - Right on
 - Relevant
 - Irrelevant
- Inter-annotator agreement: 0.71

Table Search Evaluation

- TABLE: Model from previous work on Table Search by Venetis et. al.
- Google: Results returned by google.com
- **WikiTables:** Our system

Table Search Evaluation

TABLE
Google
WikiTables

- TABLE: Model from previous work on Table Search by Venetis et. al.
- Google: Results returned by google.com
- **WikiTables:** Our system

Table Search Evaluation

	No. of Queries for which some results returned by system
TABLE	49
Google	100
WikiTables	100

- **TABLE:** Model from previous work on Table Search by Venetis et. al.
- **Google:** Results returned by google.com
- **WikiTables:** Our system

Table Search Evaluation

	No. of Queries for which some results returned by system	Right On / Relevant Tables
TABLE	49	40
Google	100	35
WikiTables	100	63

- TABLE: Model from previous work on Table Search by Venetis et. al.
- Google: Results returned by google.com
- **WikiTables:** Our system

Table Search Features

Feature	Correlation With Rating
tablePageFraction	0.41
numRows	0.40
tableCaptionContainsQuery	0.31
sectionNumber	-0.19
inlinks	-0.16
numCols	0.156

Table Search

Features

Name	Description
Page Features	
yRank	Rank in Yahoo!'s results for query
inLinks	Number of in-links to page
outLinks	Number of out-links from page
pageViews	Number of page-views of Wikipedia page
Table Features	
numRows	Number of rows in the table
numCols	Number of columns in the table
emptyCellRatio	Fraction of empty cells in the table
Table+Page Features	
sectionNumber	Wikipedia section number in which table occurs
tableImportance	Inverse of number of tables on page
tablePageFraction	Ratio: Table size to page size
QueryFeatures	
qInPgTitle	Ratio: Number of query tokens found in page title to total number of tokens
qInTableTitle	Ratio: Number of query tokens found in table title to total number of tokens

Outline

- Table Extraction
- Relevant Join
- Correlation Mining
- Table Search
- Conclusion & Future Work

Conclusion

- *WikiTables*: a system that extracts, searches and automatically joins Wikipedia tables
- Leveraging high quality tables and text from Wikipedia is valuable for mining and joining
- For search, smaller number of tables on Wikipedia is outweighed by its high quality

Future Work

- Combining with higher coverage systems like Google Fusion Tables
- Richer interactive data mining, not just 2 column correlations
- Predicting some attributes of entities using others

Thanks !

Questions ?