Better Information Extraction and Search with Users

C. Lee Giles[1], Prasenjit Mitra[1], Rebecca Passonneau[1], Jian Wu[1], Cornelia Caragea[2]

Richard Zanibbi[3]

Pennsylvania State University[1]

Kansas State University[2]

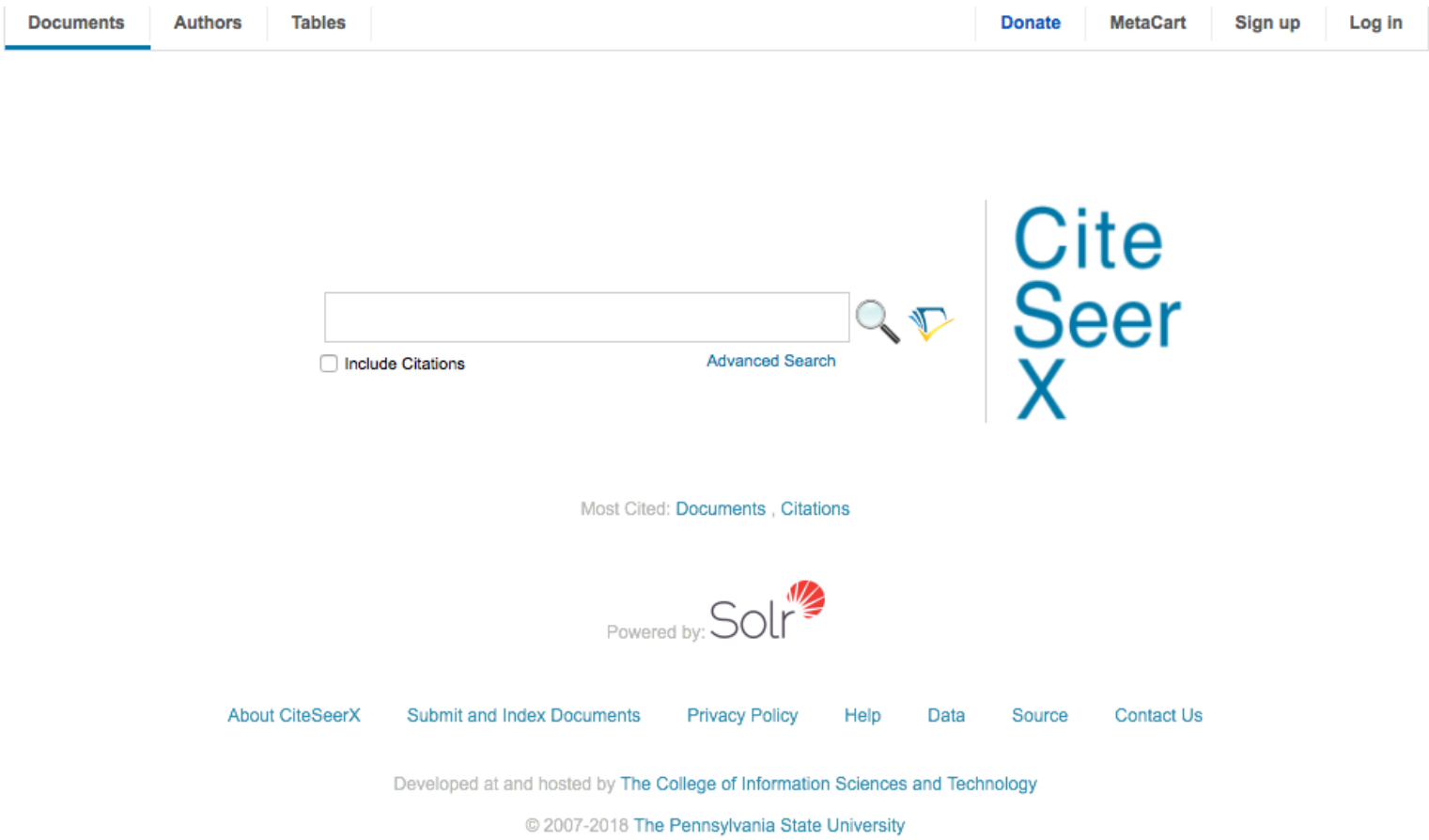
Rochester Institute of Technology[3]

giles@ist.psu.edu

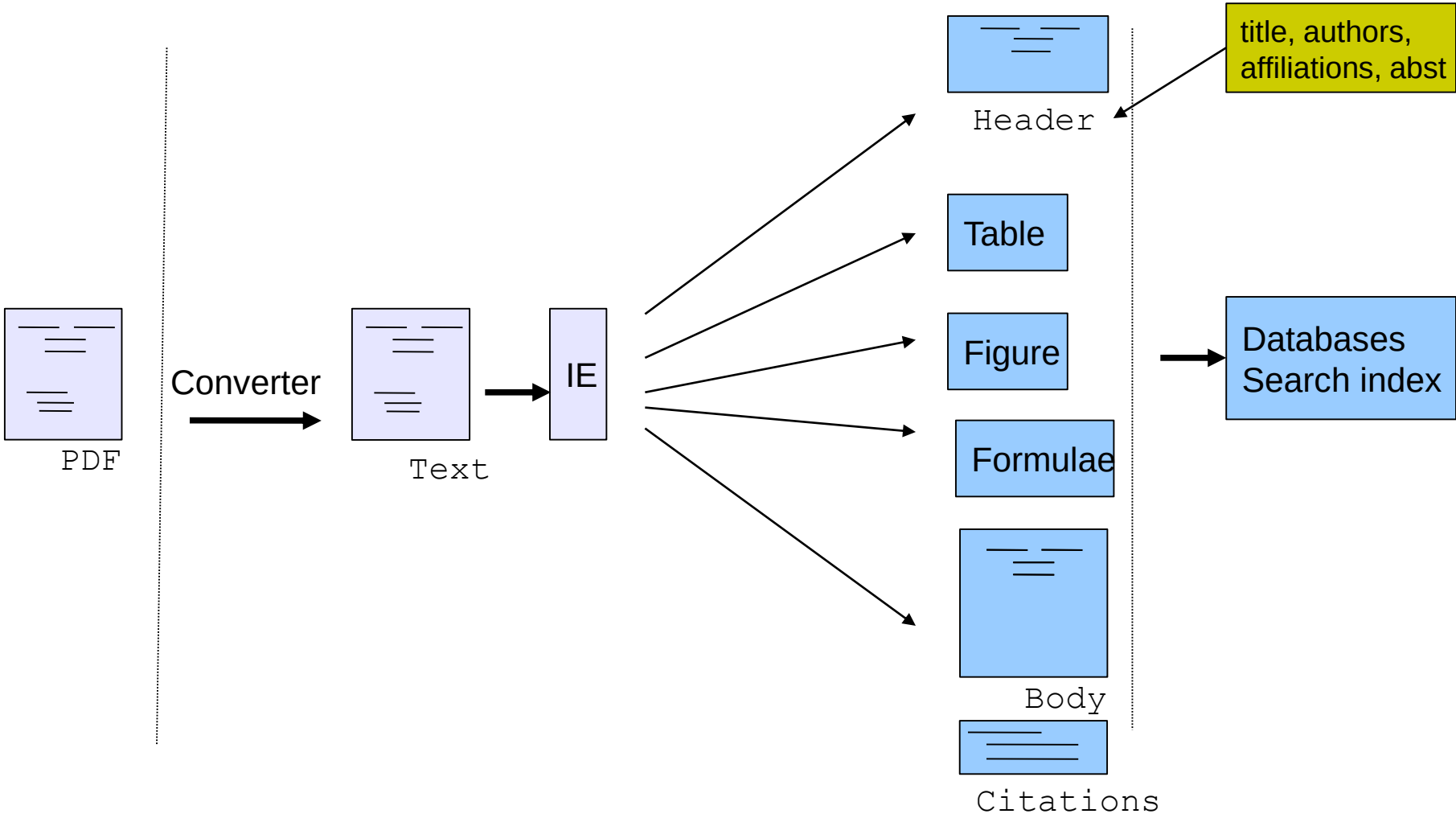
http://clgiles.ist.psu.edu

gitbhub.com/SeerLabs

Improve semantic extraction methods in CiteSeerX for better extraction



Example of Automatic Metadata Extraction (IE) – CiteSeerX (some language independent)



Many other open source academic document metadata extractors available – recent JCDL workshop, metadata hackathon, JCDL tutorial 2016

如果我们的决策面方程能够完全正确地对图2中的样本点进行分类，就会满足下面的公式

$$\begin{cases} \boldsymbol{\omega}^T \boldsymbol{x}_i + \gamma > 0 & \text{for}\sim y_i = 1 \\ \boldsymbol{\omega}^T \boldsymbol{x}_i + \gamma < 0 & \text{for}\sim y_i = -1 \end{cases} \tag{2.8}$$

SVM的全称是Support Vector Machine，即支持向量机，主要用于解决模式识别领域中的数据分类问题，属于有监督学习算法的一种。SVM要解决的问题可以用一个经典的二分类问题加以描述。如图1所示，红色和蓝色的二维数据点显然是可以被一条直线分开的，在模式识别领域称为线性可分问题。然而将两类数据点分开的直线显然不止一条。图1(b)和(c)分别给出了A、B两种不同的分类方案，其中黑色实线为分界线，术语称为“决策面”。每个决策面对应了一个线性分类器。虽然在目前的数据上看，这两个分类器的分类结果是一样的，但如果考虑潜在的其他数据，则两者的分类性能是有差别的。

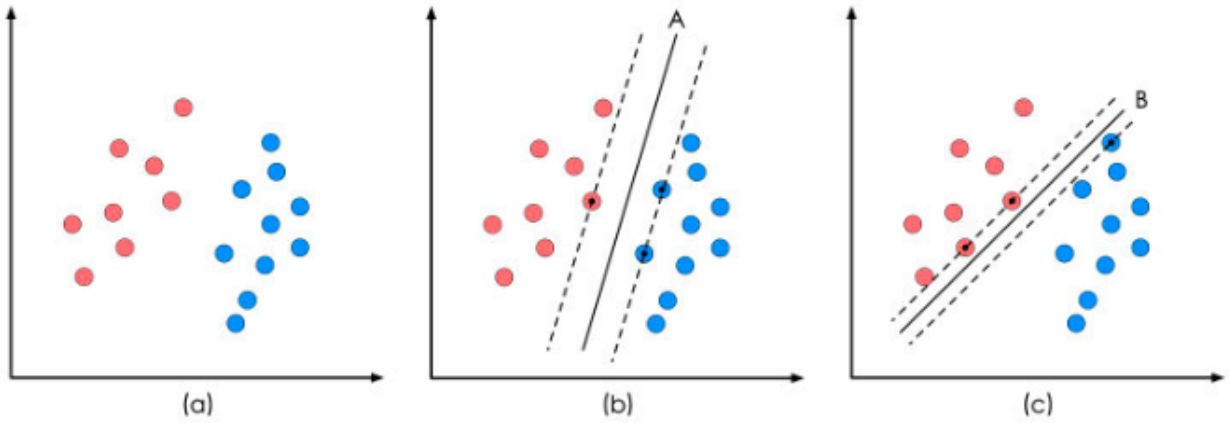
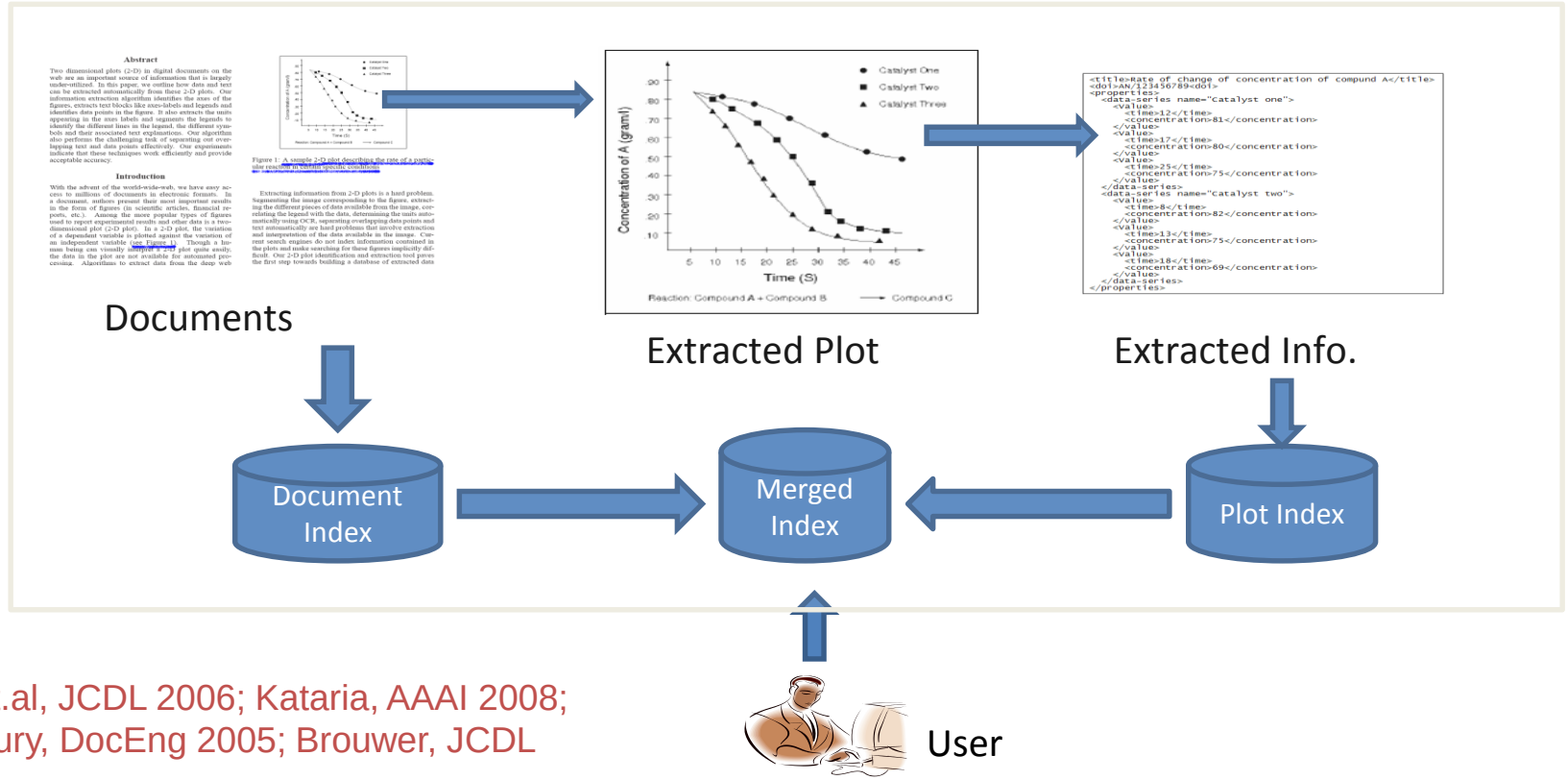


图1 二分类问题描述

Automated Figure Data Extraction and Search

- Large amount of results in digital documents are recorded in figures, time series, experimental results (eg., NMR spectra, income growth)
- Extraction for purposes of:
  - Further modeling using presented data
  - Indexing, meta-data creation for storage & search on figures for data reuse
- *Current extraction done manually!*



X. Lu, et.al, JCDL 2006; Kataria, AAAI 2008; Choudhury, DocEng 2005; Brouwer, JCDL

Bar Chart Data and Semantics Extraction

Figure Extraction – Bar Chart

"User traffic increases significantly then really drops off"

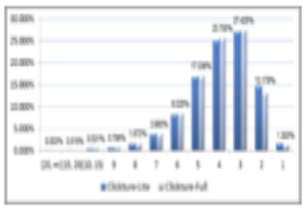


Figure 1: Query length distribution.

Bar Chart

Chart Data Extraction

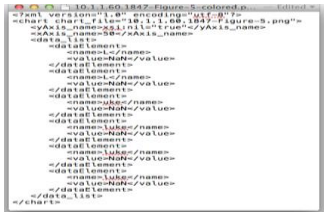


Chart Data Values

Data Feature Extraction

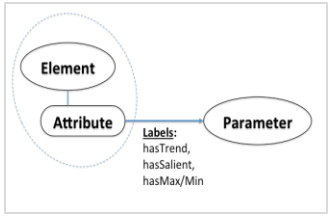


Chart structured as semantic graph

Indexed text

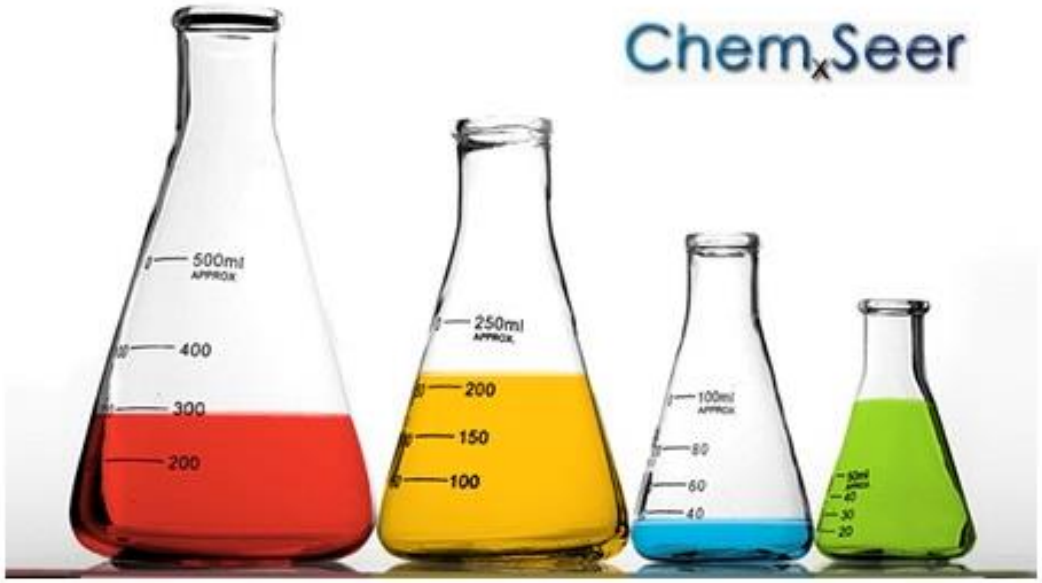
Text summaries

User queries

Al-Zaidy, AAAI 2016

ChemXSeer

Search | Papers | Authors | Tables | Figures | Formula | Extract Tables | CollabSeer



Search

Eg : Methanol, CO2, Adam Smith

B. Sun, WWW'07, WWW'08, TOIS'11
D. Yuan, ICDE'12

Challenges in Formula Search

How to identify a formula in scientific documents?

Non-Formula

*"… This work was funded under **NIH** grants …"*

*" … YSI 5301, Yellow Springs, **OH**, USA …"*

*"… action and disease. **He** has published over …"*

Formula

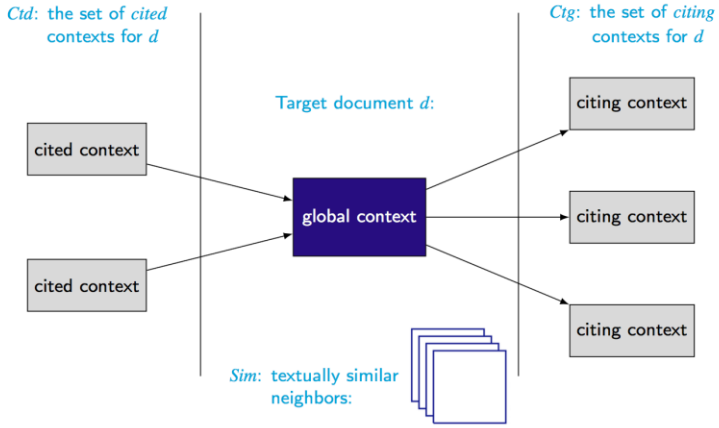
*"… such as hydroxyl radical **OH**, superoxide O2- …"*

*" and the other **He** emissions scarcely changed …"*

Machine learning algorithms (SVM + CRF) yield high accuracies for correct formula identification.

Keyphrase Extraction Approaches for IE

Using multiple sources of evidence

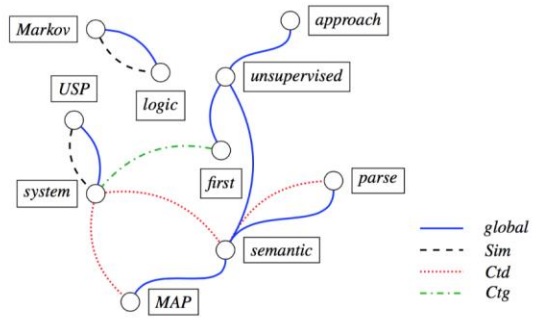


- $T = \{Ctd, Ctg, Sim, g\}$ represents the types of available contexts for d.

Unsupervised Semantic Parsing

We present the first unsupervised approach to the problem of learning a semantic parser, using Markov logic . Our USP system transforms dependency trees into quasi-logical forms, recursively induces lambda forms from these, and clusters them to abstract away syntactic variations of the same meaning. The MAP semantic parse of a sentence is obtained by recursively assigning its parts to lambda-form clusters and composing them. We evaluate our approach by using it to extract a knowledge base from biomedical abstracts and answer questions. USP substantially outperforms TextRunner, DIRT and an informed baseline on both precision and recall on this task.

$w = 2$:



Title: A Unified Approach for Schema Matching, Coreference and Canonicalization by Wick et al.

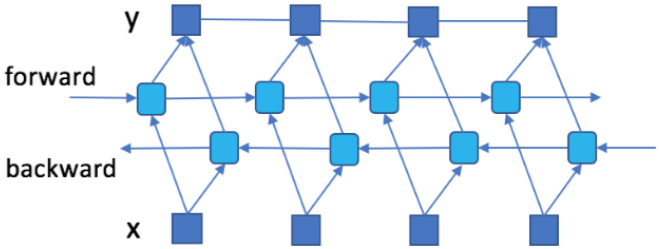
ABSTRACT

The automatic consolidation of database records from many heterogeneous sources into a single repository requires solving several information integration tasks. Although tasks such as coreference, schema matching, and canonicalization are closely related, they are most commonly studied in isolation. Systems that do tackle multiple integration problems traditionally solve each independently, allowing errors to propagate from one task to another. In this paper, we describe a discriminatively-trained model that reasons about schema matching, coreference, and canonicalization jointly. We evaluate our model on a real-world data set of people and demonstrate that simultaneously solving these tasks reduces errors over a cascaded or isolated approach. Our experiments show that a joint model is able to improve substantially over systems that either solve each task in isolation or with the conventional cascade. We demonstrate nearly a 50% error reduction for coreference and a 40% error reduction for schema matching.

Keywords

Data Integration, Coreference, Schema Matching, Canonicalization, Conditional Random Field, Weighted Logic

Taking into account deep semantics hidden in text



Bi-LSTM-CRF

[Al-Zaidy, Caragea, Giles, 2018 (under submission)]

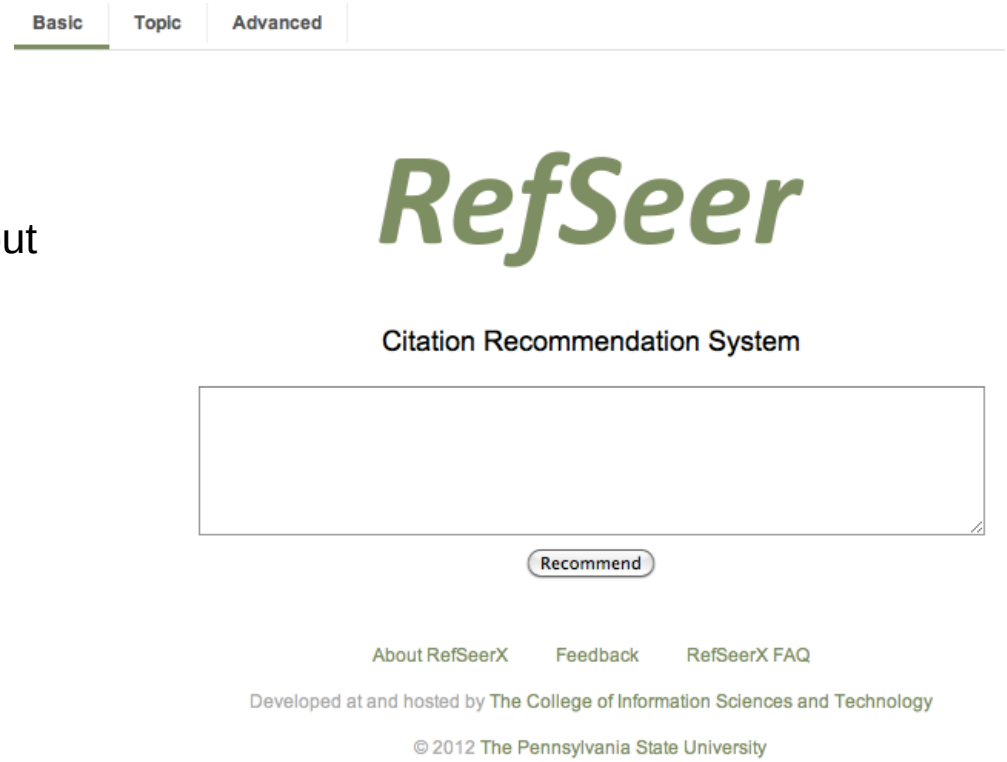
Helping you write: Automatic Citation (or paper) Recommendation

Built on millions of papers

Never miss a citation and know about the latest work

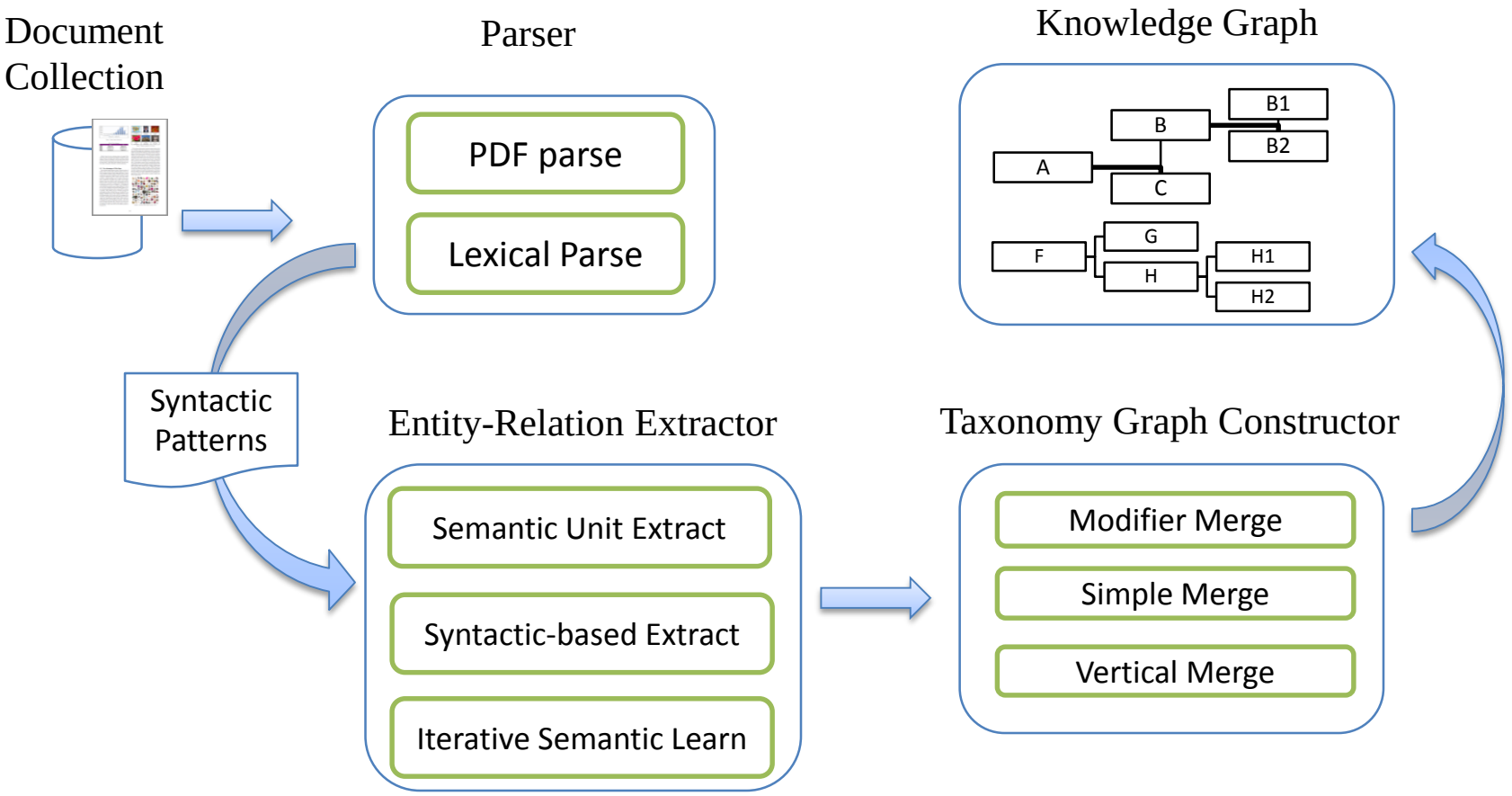
Several recommendations models

Future directions
- Where to place a citation
- What to say

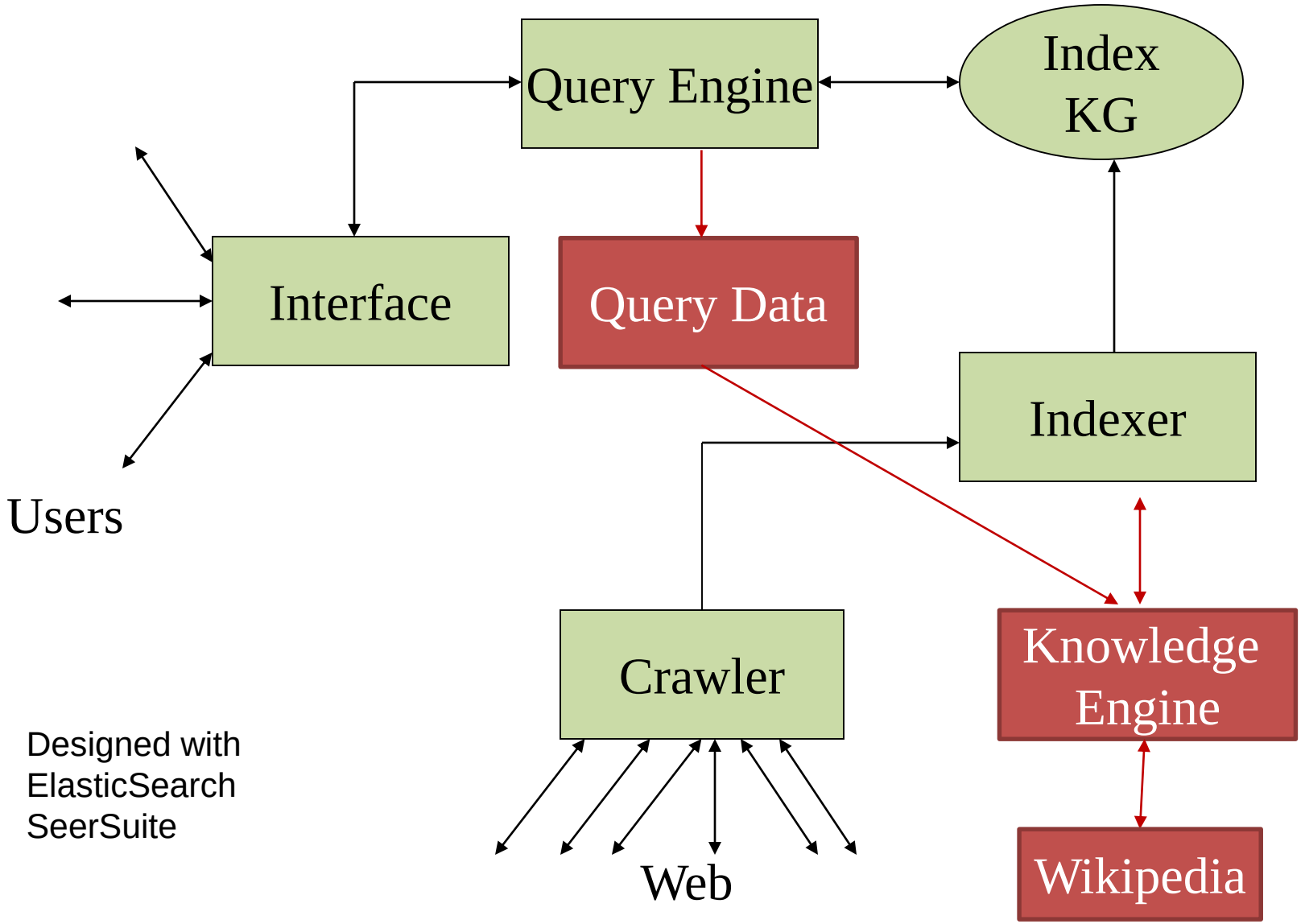


Huang, AAAI 2015
Huang, CIKM 2013
He, WWW 2010

Knowledge Graph (KG) Construction Pipeline



Al-Zaidy, Doceng 2017



Knowledge Graph Web Search Engine