Learning to rank adaptively for scalable information extraction

Pablo Barrio, Columbia University
Gonçalo Simões, INESC-ID and IST, University of Lisbon
Helena Galhardas, INESC-ID and IST, University of Lisbon
Luis Gravano, Columbia University

Information Extraction (IE)

- Natural-language text **embeds** “structured” data

- Information extraction systems **extract** this data

IE is Challenging and Time Consuming

- Operates over **large sets of features**

  Bag of words, N-grams, grammar productions, dependency paths

"… A tornado swept the coast of Florida on Wednesday…" →

tornado
swept
…
wednesday
…

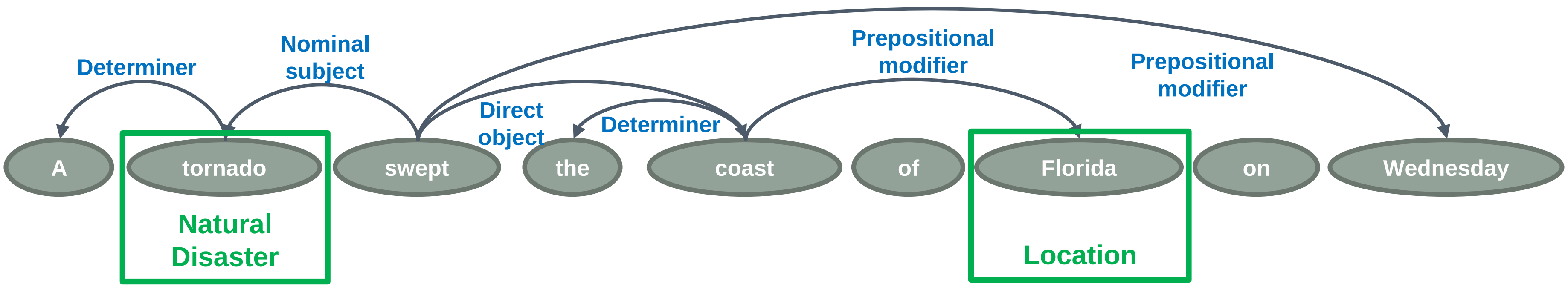
tornado swept
swept the
…
coast of
…

2-grams

May grow as large as number of **unique words** and **sequences of N words**

- Requires **complex text analysis**

  Dependency parsing, entity recognition, syntactic parsing, shallow parsing, part-of-speech tagging, semantic role labeling



May take **several seconds per document**
(e.g., with subsequence kernel extractor for Natural Disaster-Location)
Problematic over large document collections

Reducing Processing Time: Opportunities

Documents are "useful" if they produce output for a given IE task

- **Small**, **topic-specific** fraction of collection

  Only **2% of documents** in a New York Times archive, mostly **environment-related**, are useful for Natural Disaster-Location with a state-of-the-art IE system

Should focus extraction over these documents and ignore rest

- Useful documents share **distinctive words and phrases**

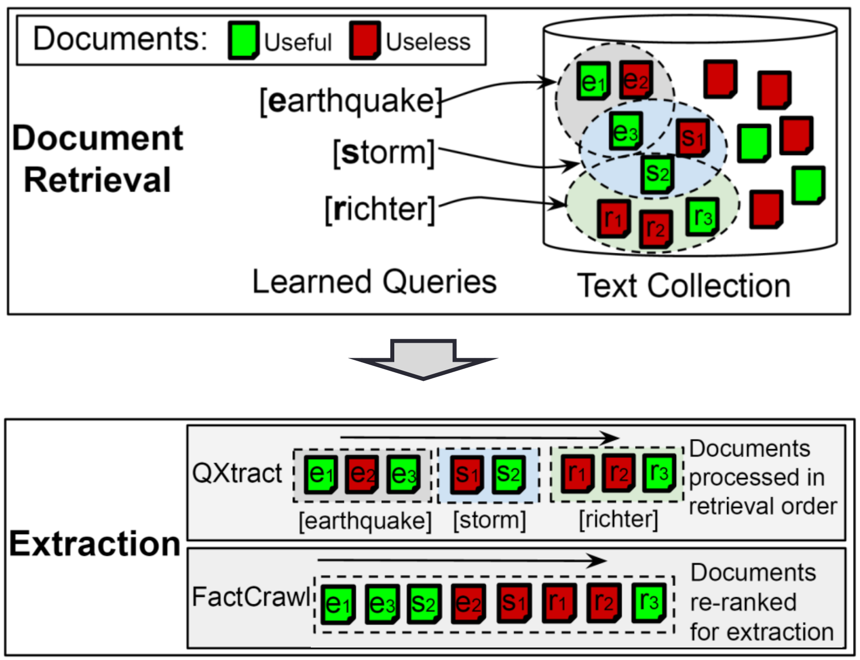
  "Earthquake," "storm," "Richter," "volcano eruption" for Natural Disaster-Location

Can learn to differentiate between useful documents for an IE task and rest

- Information extraction system **"labels"** documents as useful or not **for free**

IE process generates ever-expanding training set for learning to identify useful documents

Existing Approaches: QXtract and FactCrawl



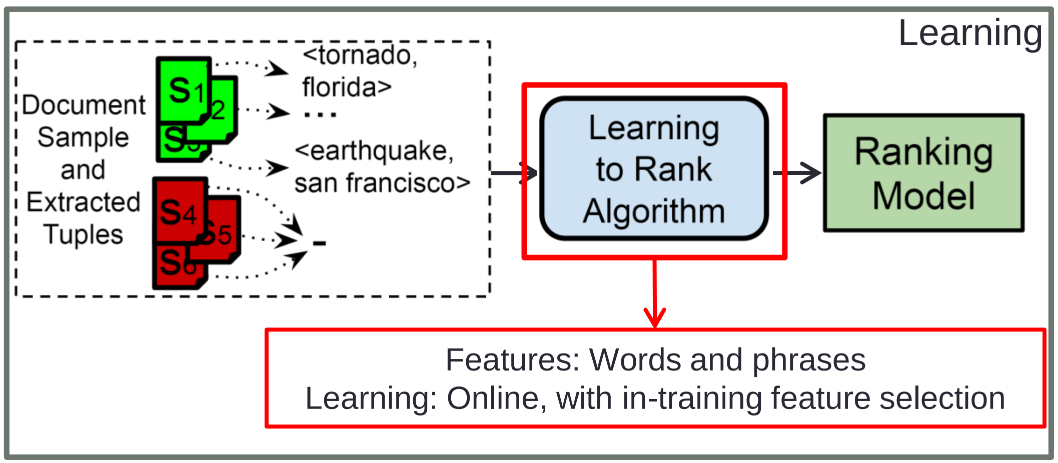
QXtract and FactCrawl learn from small document sample and exhibit far-from-perfect recall

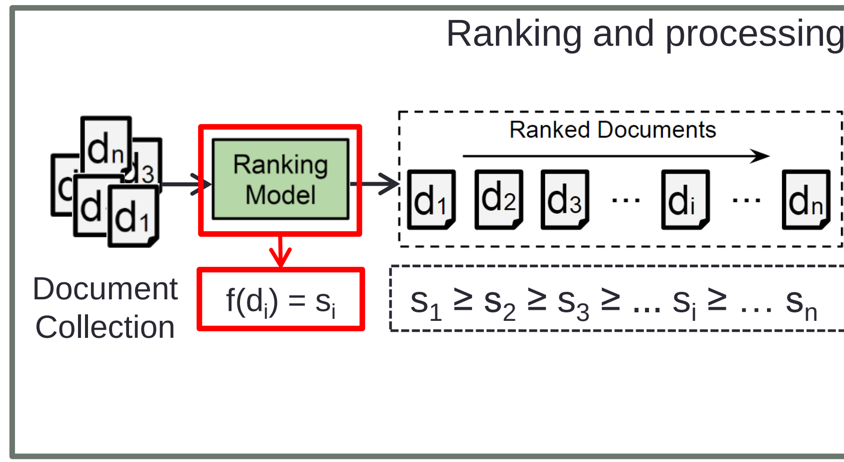
FactCrawl ranks documents using learned queries and does not adapt to new processed documents

[Eugene Agichtein and Luis Gravano, "Querying text databases for efficient information extraction." *ICDE ’03*]
[Christoph Boden et al., "FactCrawl: A fact retrieval framework for full-text indices." *WebDB* ’11]

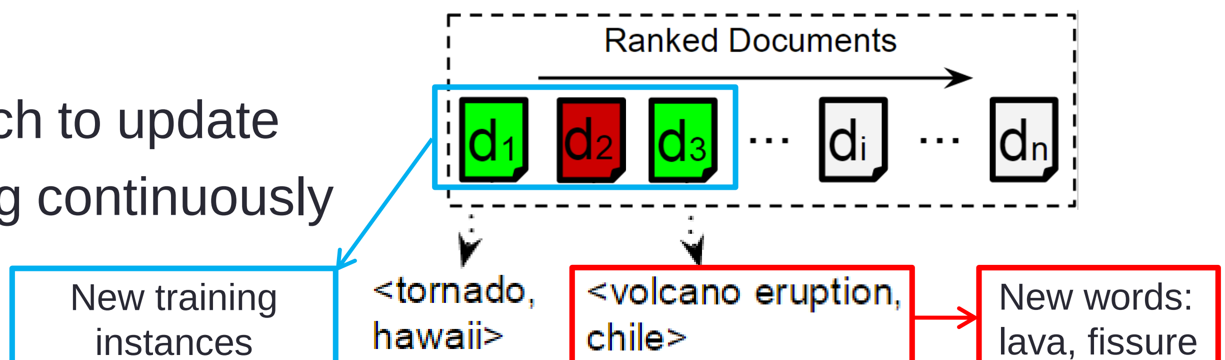
Our Approach: Key Aspects

- Document ranking needs to be **robust and efficient**
  - Learning to rank approach for document ranking

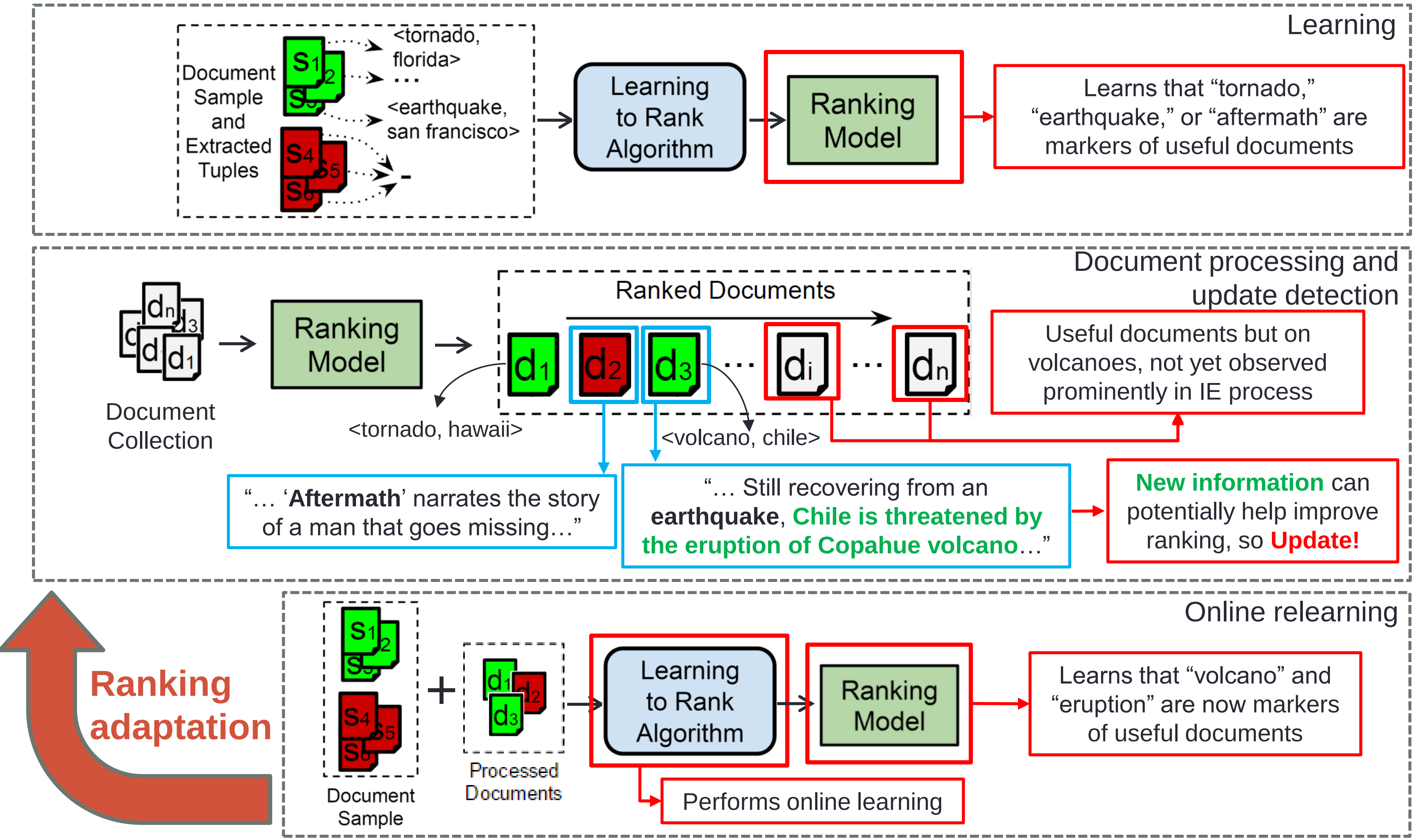




- Results of extraction process form **ever-expanding training set**
  - Adaptive approach to update document ranking continuously



Ranking Documents Adaptively for IE

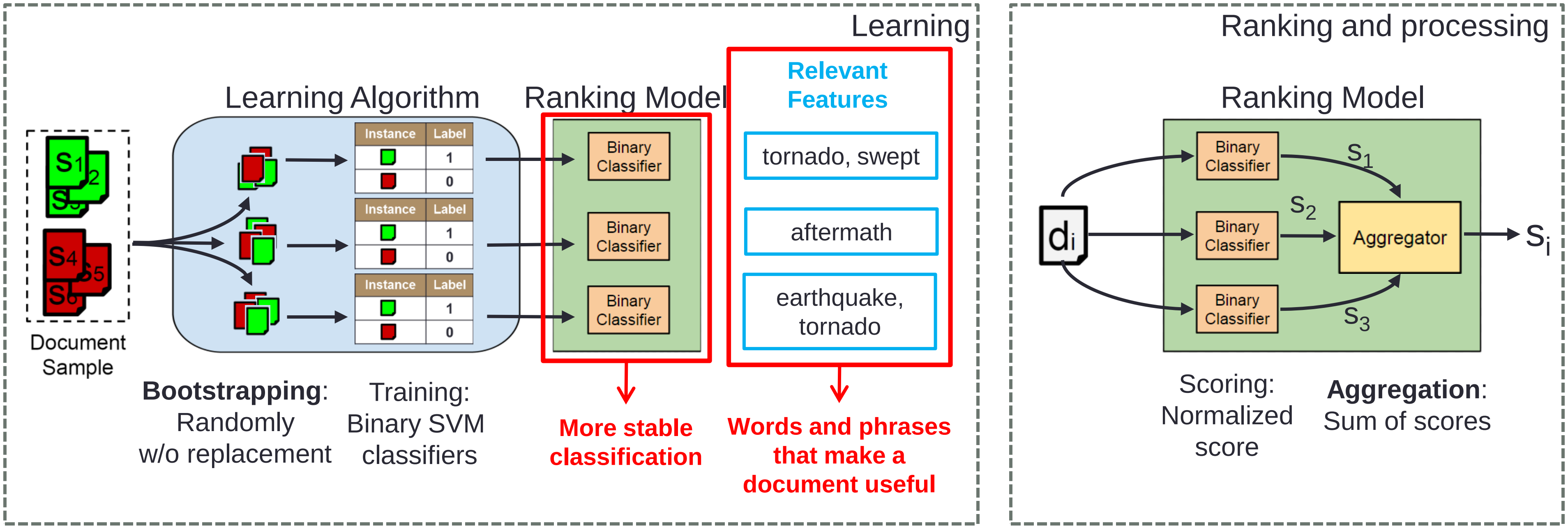


Ranking Documents Adaptively for IE: Our Alternatives

- Efficient learning-to-rank techniques for information extraction: **BAgg-IE**, **RSVM-IE**
- Update detection techniques for document ranking adaptation: **Top-*K***, **Mod-C**

Efficient Learning to Rank for IE: BAgg-IE

- Based on **bootstrapping aggregation**



All models are trained using online learning and in-training feature selection

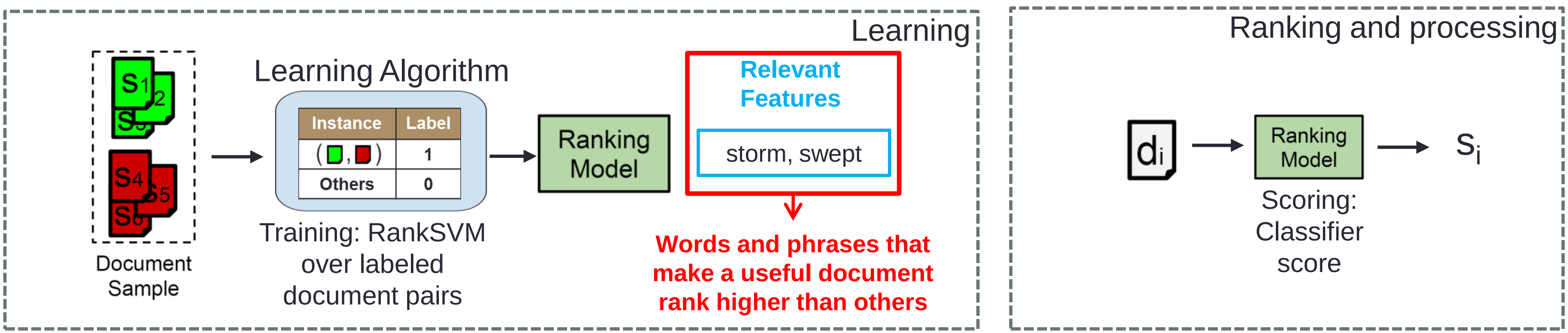
Efficient Learning to Rank for IE: RSVM-IE

- Based on **RankSVM**

  Learns SVM classifier on pairwise difference of documents

Learning model	Training instance
RankSVM	d_i - d_n
SVM	d_i

Training Label is **1** iif **d_i** is "better" than **d_n**



Model is trained using online learning and in-training feature selection

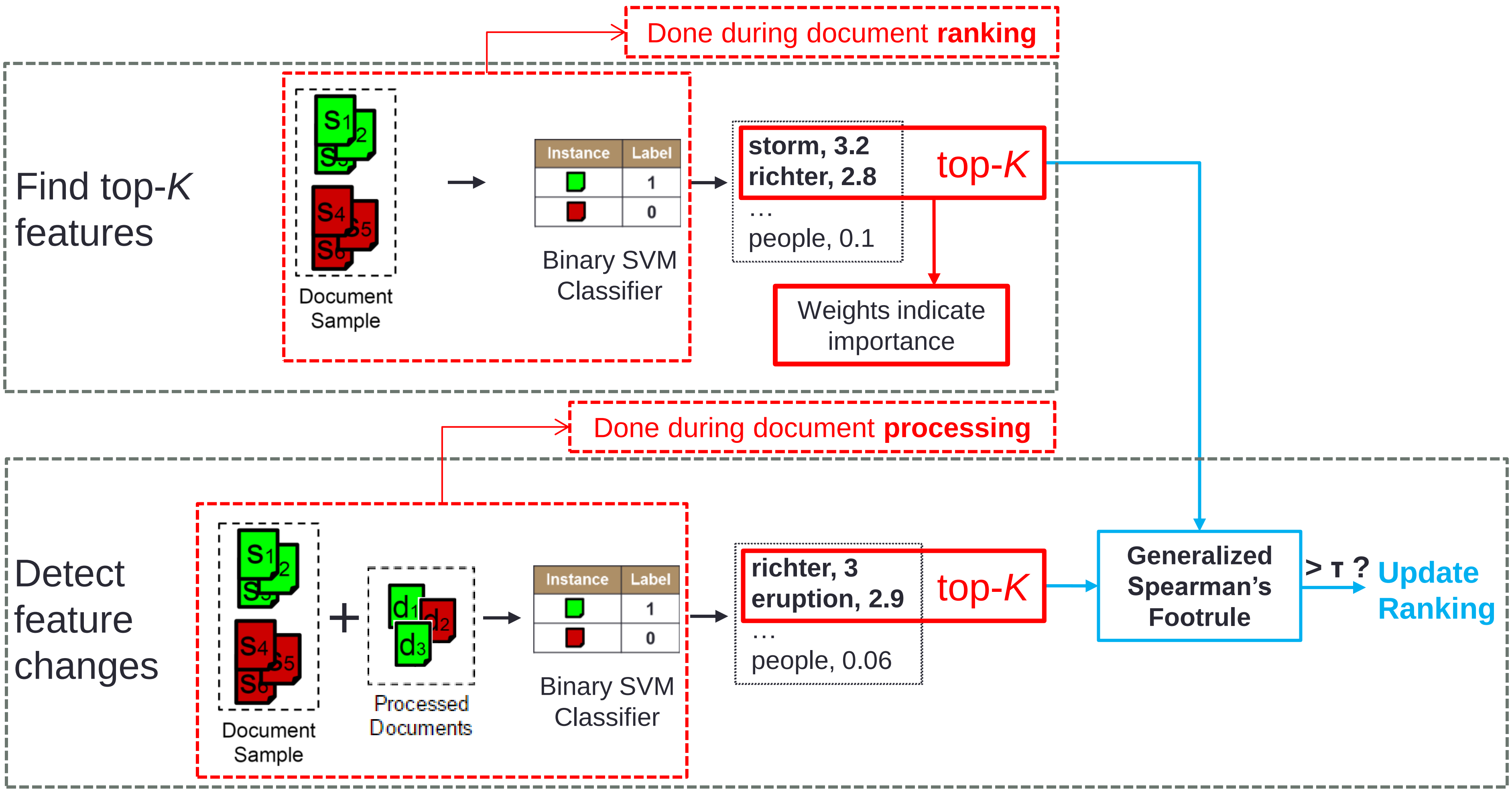
Ranking Documents Adaptively for IE: Our Alternatives

- Efficient learning-to-rank techniques for information extraction: **BAgg-IE**, **RSVM-IE**

- Update detection techniques for document ranking adaptation: **Top-*K***, **Mod-C**

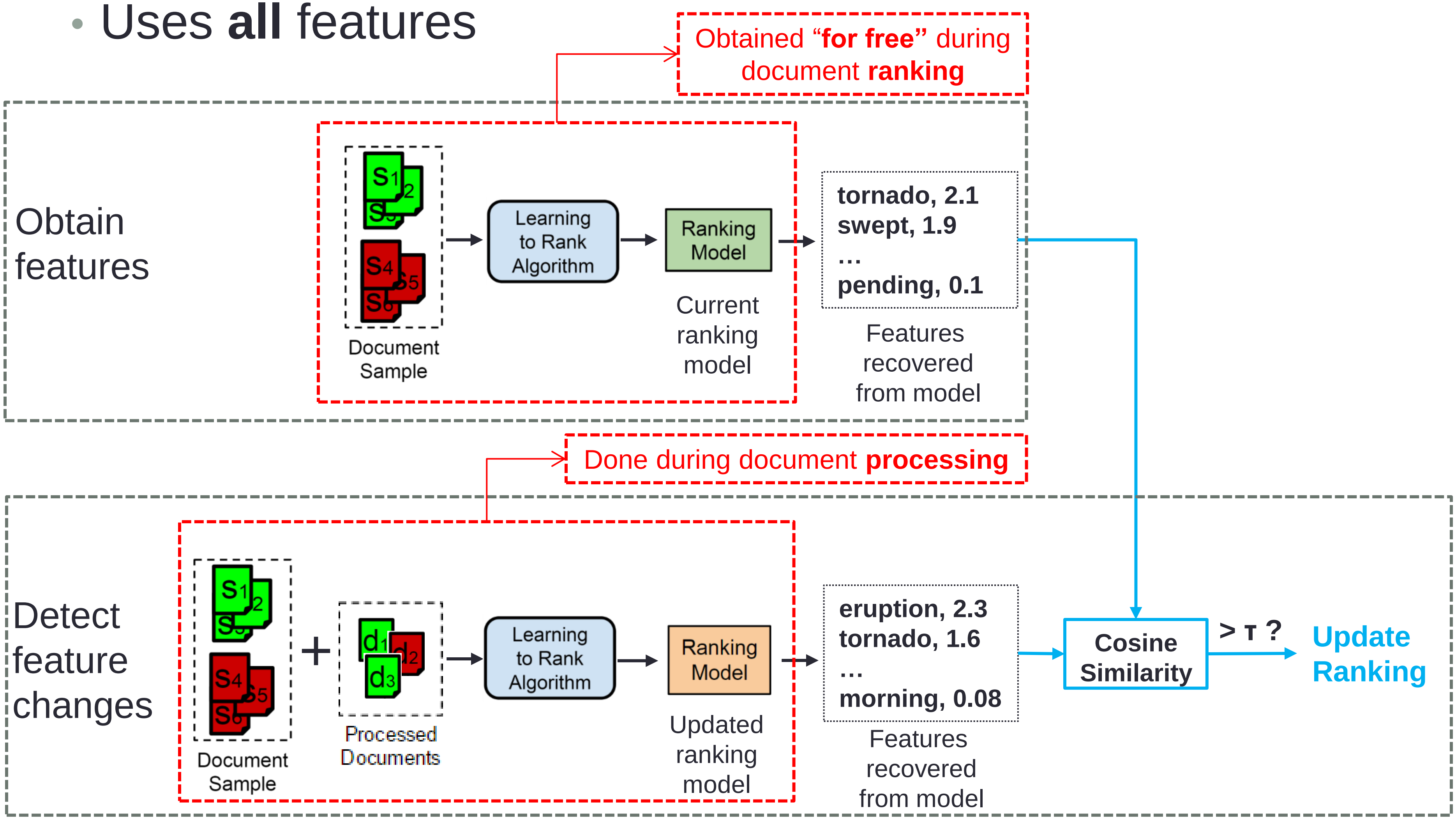
Update Detection for Document Ranking Adaptation: Top-*K*

- Uses only **most important** (top-*K*) features



Update Detection for Document Ranking Adaptation: Mod-C

- Uses **all** features



Experimental Settings

- Dataset: The New York Times archive: 1.8 million articles from 1987-2007
- Information extraction systems

Simple extraction systems:
HMMs, text patterns

Person-Organization

Google co-founders **Larry Page** and **Sergey Brin** recently sat down with billionaire venture capitalist Vinod Khosla for a lengthy interview.

Person	Organization
Larry Page	Google
Sergey Brin	Google

Disease-Outbreaks

Disease	Time Period
cholera	between 2010 and 2013

The Haiti **cholera** outbreak **between 2010 and 2013** was the worst epidemic of cholera in recent history.

Person-Career

"This is not a victimless crime," said **Jim Kendall**, **president** of the Washington Association of Internet Service Providers.

Person	Career
Jim Kendall	President

Man Made Disaster-Location

Disaster	Location
fire	Booneville

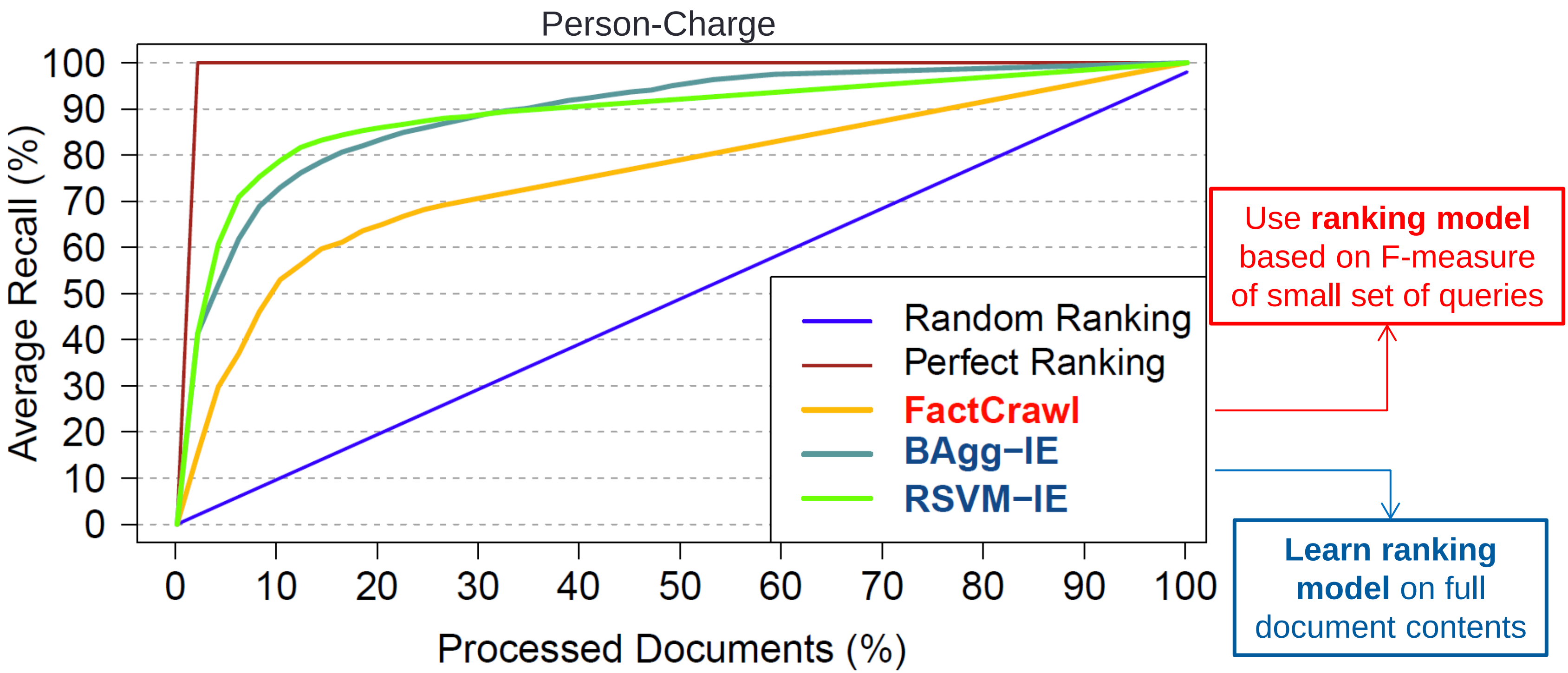
A **fire** destroyed a Cargill Meat Solutions beef processing plant in **Booneville**.

Other relations:
Person-Charge, Election-Winner, Natural Disaster-Location

Complex extraction systems:
CRFs, SVM kernels

Dense relations | Sparse relations

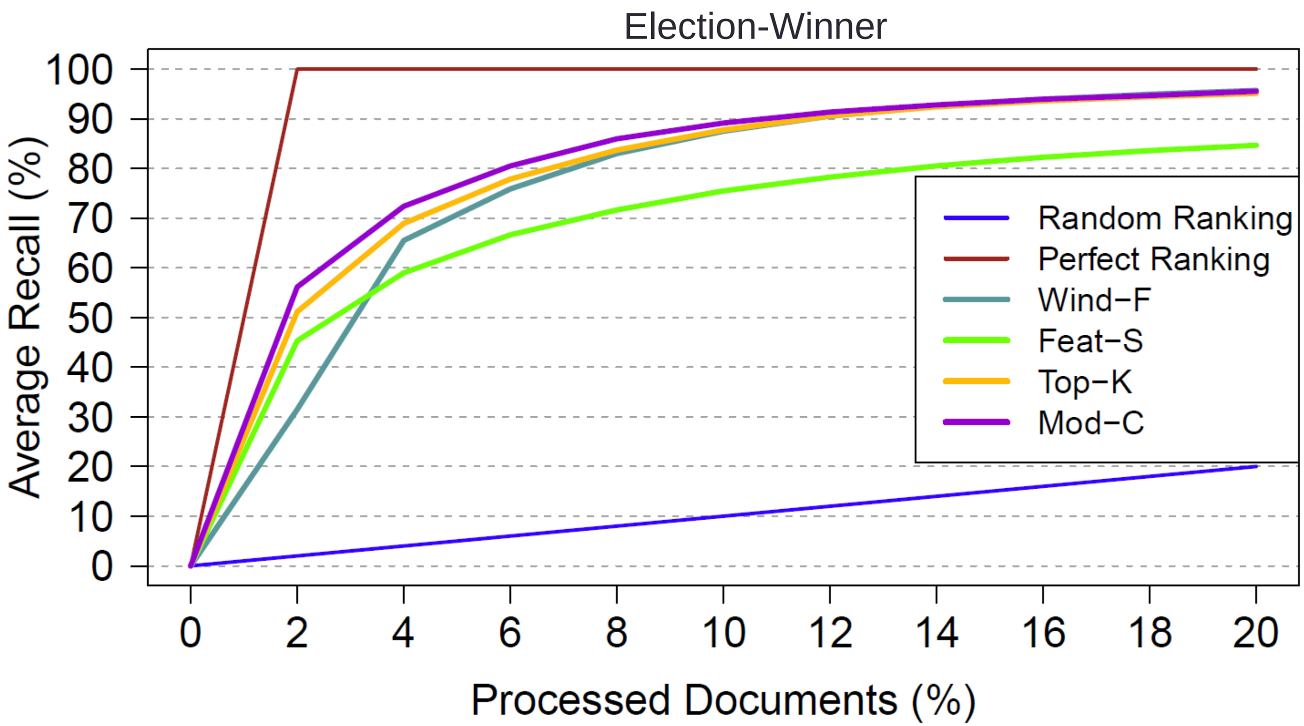
Does Learning Ranking Models Help?



- **Learning ranking models** leads to better document ranking
- **RSVM-IE** performs best at **early stages**
- **BAgg-IE** obtains high gains **later on**
- Objective function of **learning model shapes document ranking**

Additional experiments in paper: analogous conclusions over all relations

Does Update Detection Help?



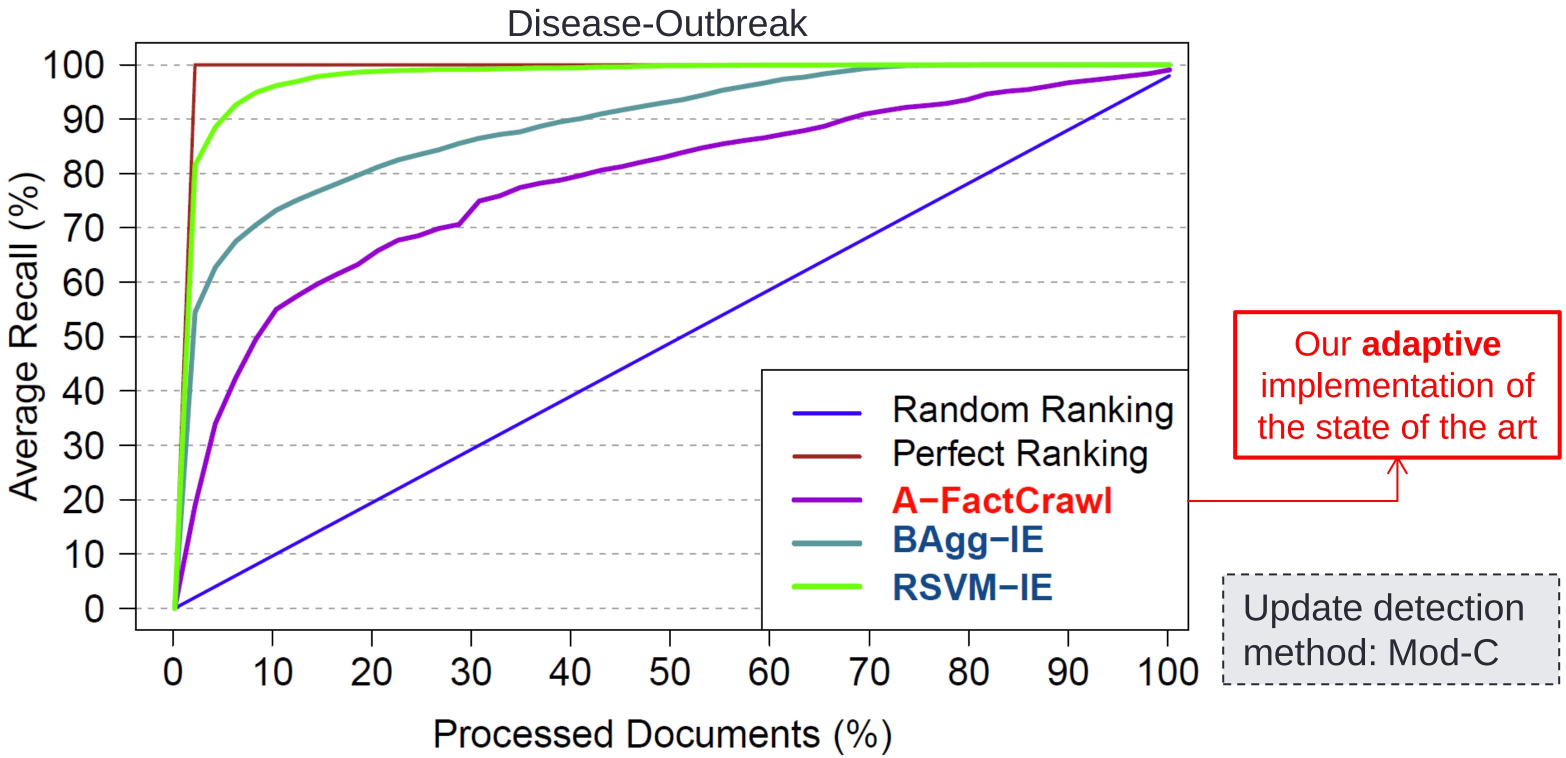
Update detection baselines:
Wind-F=Updates after processing 20,000 documents (2% of collection)
Feat-S=Update method based on Gaussian kernel *[Glazer, ICPR '12]*

Ranking Strategy: RSVM-IE

- **Feat-S** unable to evaluate over new features, crucial during adaptation
- **Top-*K*** and **Mod-C** improve the efficiency of the extraction process
- **Mod-C** leads to best execution using more efficient approach, with fewer models

Additional experiments in paper: analogous conclusions over all relations

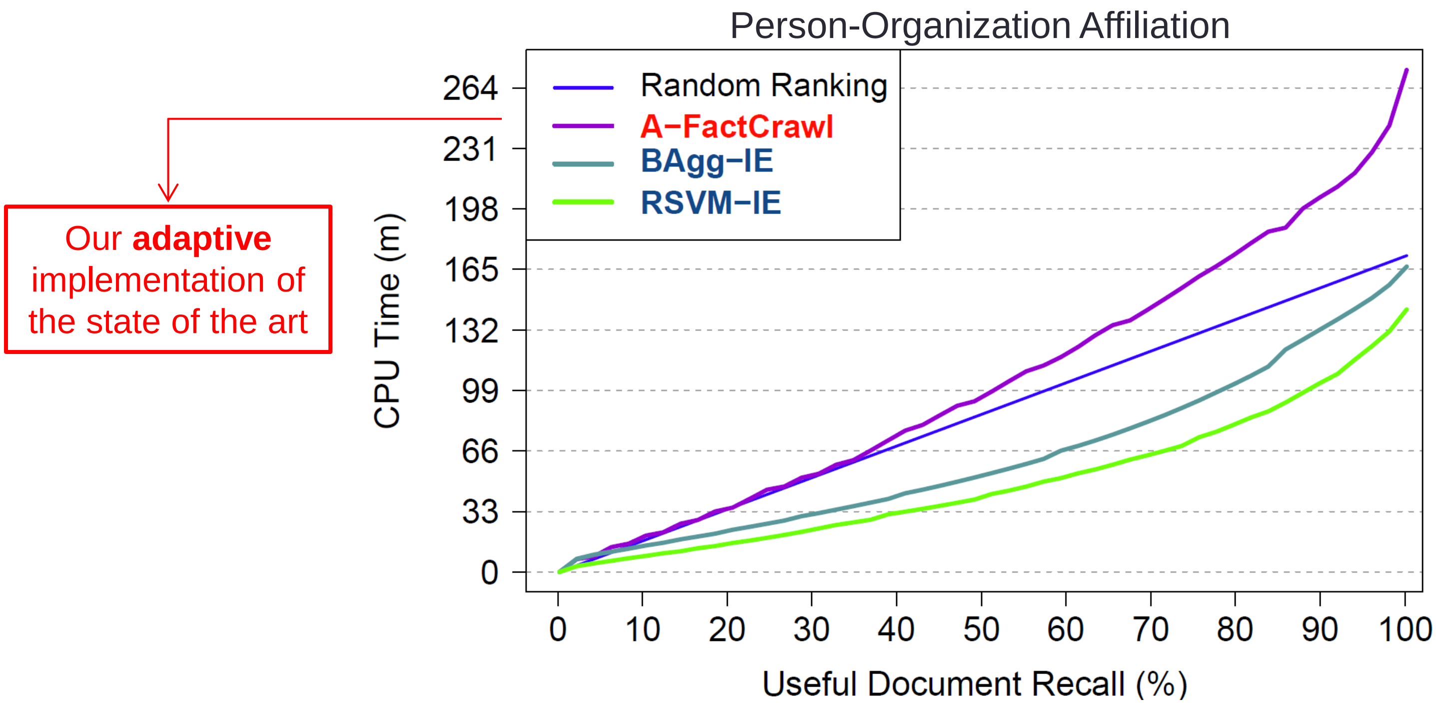
Putting Learning to Rank and Update Detection Together: Recall Analysis



- **Our techniques** bring **significant improvement** for sparse relations
- **RSVM-IE performs best**, as it prioritizes useful documents better, favoring adaptation

Additional experiments in paper: analogous conclusions over all relations

Putting Learning to Rank and Update Detection Together: Extraction Time



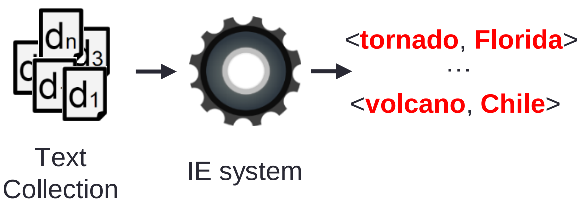
- Cost of adapting in **A-FactCrawl** hurts efficiency of extraction process
- **Our techniques improve** efficiency of process even for inexpensive IE systems

Additional experiments in paper for **our techniques**:

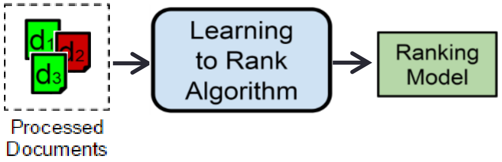
- Analogous conclusions also for expensive IE systems and sparse relations
- **Scale linearly** in the size of the collection

Document Ranking for Scalable Information Extraction: Summing Up

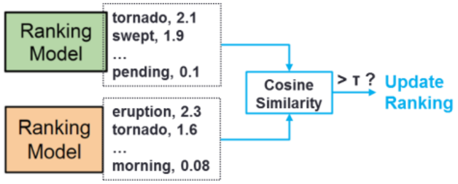
- Running IE system over **large** text collections is computationally **expensive**



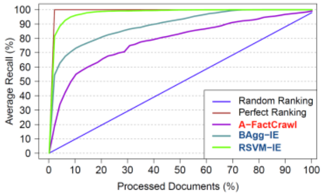
- Proposed lightweight, adaptive approach and learning-based alternatives
    - **Online learning** algorithms with **in-training feature selection**: RSVM-IE, BAgg-IE



    - **Update detection** based on feature changes: Mod-C, Top-*K*



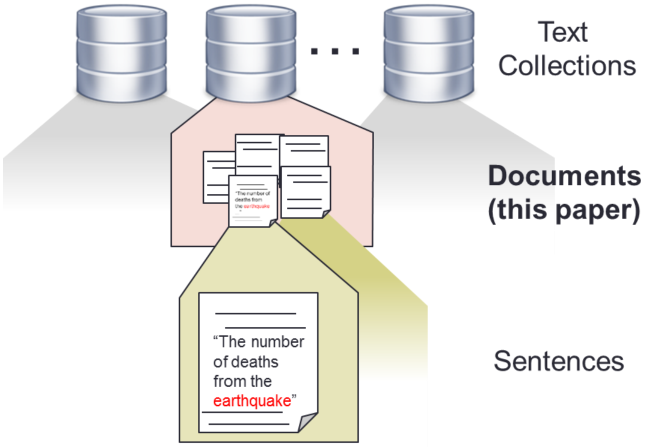
- **RSVM-IE** + **Mod-C performs best**: Useful documents are better prioritized, enabling richer, more efficient ranking adaptation



Future Work: Ranking at Different Granularities

- **Few collections** on the Web are relevant to an IE task

  Prioritize them based on number of useful documents

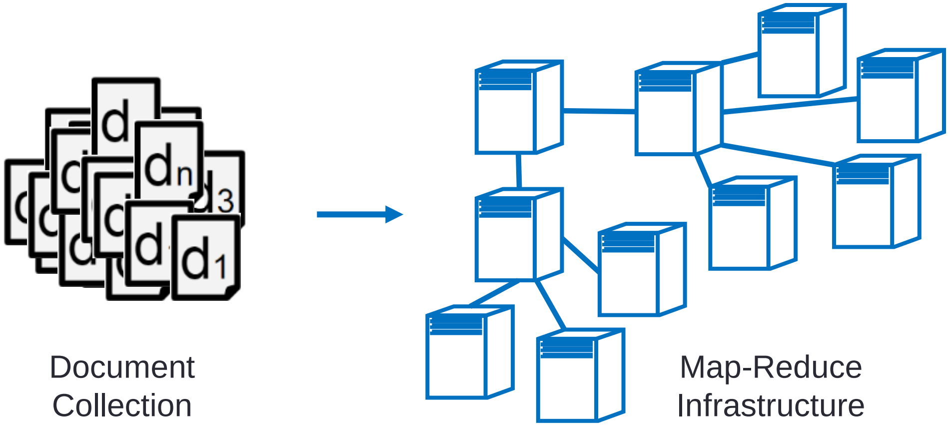


- **Few sentences** in a text document output tuples for an IE task

  Prioritize them based on usefulness and diversity

Future Work: Distributing the Execution of IE Systems

- Identify **optimal distributed execution strategy**
  - E.g., by determining document placement in distributed file system



But Before We Leave…

Try **REEL**, our toolkit to **easily develop and evaluate** IE systems

Open source and freely available at http://reel.cs.columbia.edu

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

NSF

IARPA

Thanks!

Information Extraction: Time Analysis

Task	Time per sentence (ms)		Toolkit or Algorithm	
Sentence splitting	0.1		PTB	
Tokenization	0.1		PTB	
Part-of-speech tagging	7.4		ClearNLP	
Shallow parsing	42		Search	
Dependency parsing	25.6		ClearNLP	
Semantic role labeling	8.4		ClearNLP	
Named Entity recognition (per entity)	1.1		SENNA	
Relation extraction	766	67	Tree Kernel	OLLIE
Total	**850.7**	**151.7**		

Experimental Settings: Data and Relations

- Dataset: The New York Times 1.8 million articles from 1987-2007
- Information Extraction Systems

Simple extraction systems:
HMMs, Text patterns

Person-Organization

Person	Organization
Larry Page	Google
Sergey Brin	Google

Google co-founders **Larry Page** and **Sergey Brin** recently sat down with billionaire venture capitalist Vinod Khosla for a lengthy interview.

Disease-Outbreaks

Disease	Time Period
Cholera	between 2010 and 2013

The Haiti cholera outbreak **between 2010 and 2013** was the worst epidemic of **cholera** in recent history.

Person-Career

Person	Career
Jim Kendall	President

"This is not a victimless crime," said **Jim Kendall**, **president** of the Washington Association of Internet Service Providers.

Dense relations

Man Made Disaster-Location

Disaster	Location
fire	Booneville

A **fire** destroyed a Cargill Meat Solutions beef processing plant in **Booneville**.

Person-Charge

Person	Charge
Ibrahim Muktar Said	Connection with bombing

Ibrahim Muktar Said was charged Sunday night in **connection with** the failed Hackney bus **bombing**.

Election-Winner

Person	Election
Boris Johnson	London mayoral election

Boris Johnson defeated Ken Livingstone in the **London mayoral election**.

Natural Disaster-Location

Disaster	Location
tornado	Florida

A **tornado** swept the coast of **Florida** on Wednesday.

Complex extraction systems:
CRFs, SVM Kernels

Sparse relations

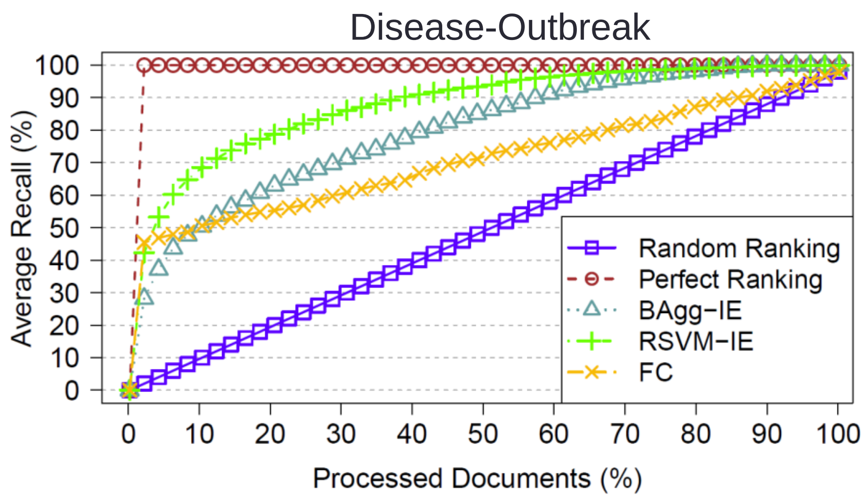
Experimental Settings: Extractors

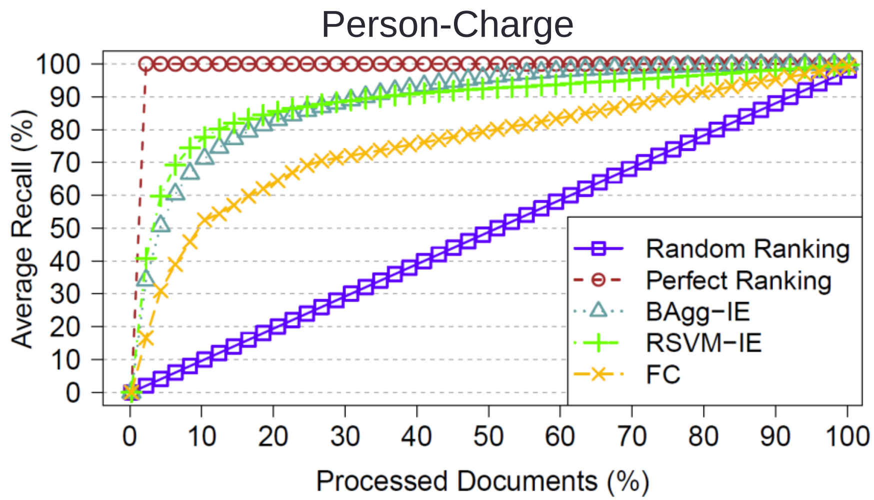
- Person-Organization Affiliation:
  - Entities: HMM and text patterns
  - Relation: SVM classifier
- Disease-Outbreak:
  - Entities: Dictionaries and manually crafted regular expressions
  - Relation: Distance between entities
- Others:
  - Entities: Stanford NLP (Person and Location), MEMM (Natural Disasters), and CRF (others)
  - Relation: Subsequences Kernel *[Bunescu and Mooney, NIPS '05]*

Experimental Settings: Details

- Document Sampling Strategies:
  - Simple Random Sampling (**SRS**): Documents are collected randomly from fully-accessible collection
  - Cyclic Querying Sampling (**CQS**): Queries learned from external collection and issued in a round-robin fashion
- Update Detection:
  - Feature Shifting (**Feat-S**): Gaussian kernel for one-class classification
    - Triggers an update for high geometrical difference
      [A. Glazer, "Feature Shift Detection." *ICPR* '12]
  - Fixed Window (**Wind-F**): Triggers after processing N documents

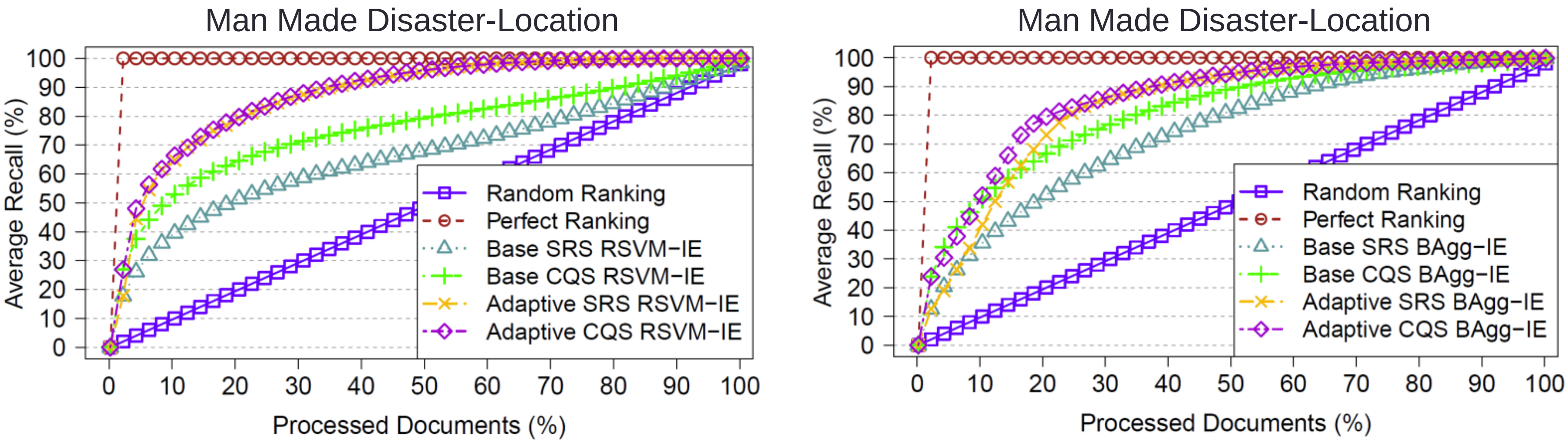
Ranking Models vs. FactCrawl





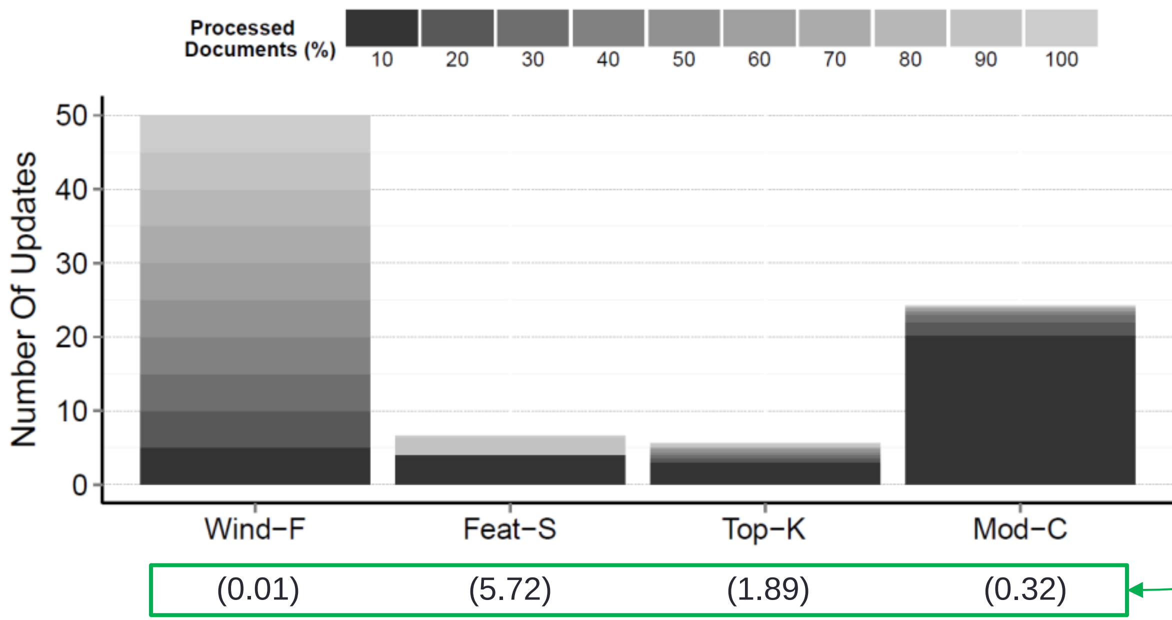
- Using **full document contents** leads to better document ranking
- RSVM-IE performs best at **early stages**
- BAgg-IE obtains high gains **later on**
- **Objective function shapes the document ranking**

Impact of Document Sampling



- CQS improves recall at early stages
- CQS obtains higher average precision and AUC
- Targeted sampling **improves the efficiency of** the extraction process

Update Detection: Time and Distribution of Updates



Update detection baselines:
Wind-F=Updates after processing 20,000 documents (2% of collection)
Feat-S=Update method based on Gaussian kernel *[Glazer, ICPR '12]*

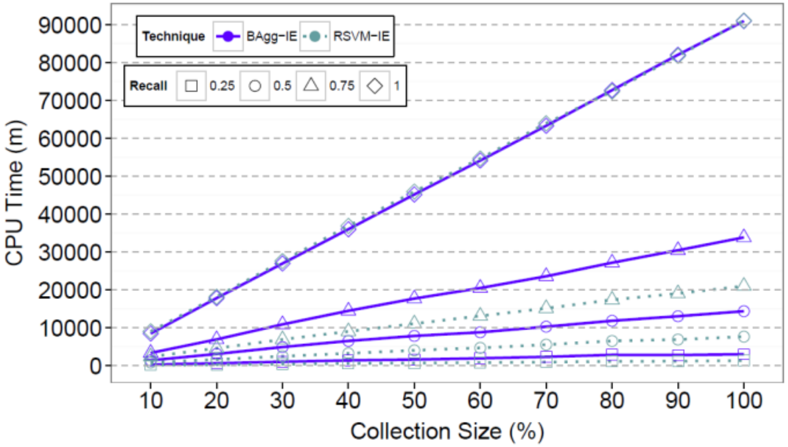
Average CPU time per document (ms)

- Wind-F is the **most efficient** but ignores document contents
- Feat-S performs fewer updates but is affected by kernel cost
- Top-*K* performs the **fewest** updates, relatively efficiently
- Mod-C exhibits best **number of updates-time balance**

Scalability Analysis: Running Time

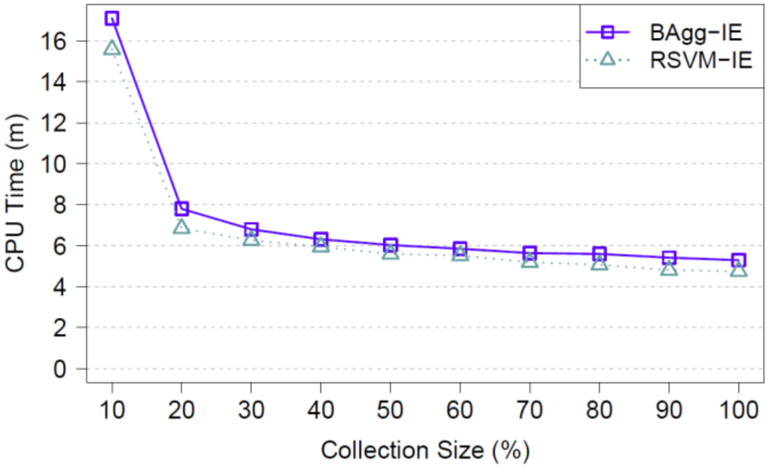
Target recall

Natural Disaster-Location



Fixed set of useful documents

Person-Organization Affiliation



- Our approach:
  - **Scales linearly** to collection size
  - **Improves** with the more information we find in larger collections
  - **Is a substantial step towards scalable information extraction**