

Focused Crawling with Scalable Ordinal Regression Solvers

Rashmin Babaria, J Saketha Nath, Krishnan S, KR Sivaramakrishnan,
Chiranjib Bhattacharyya, M N Murty

Department of Computer Science and Automation
Indian Institute of Science, INDIA

ICML-2007

Focused Crawling & Large scale OR

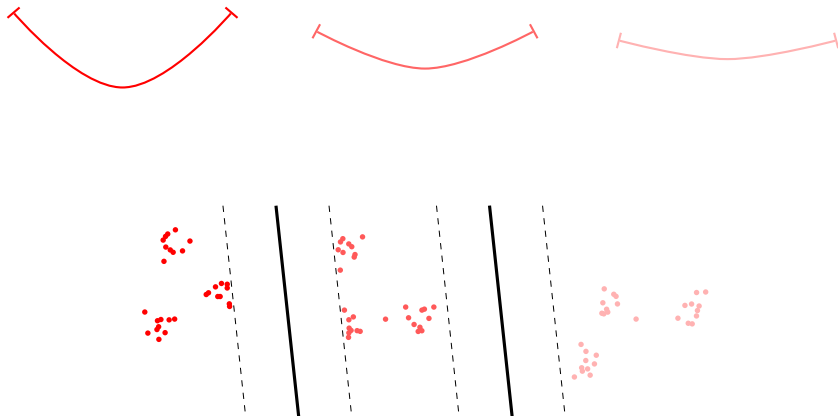
Focused Crawling

- Given a topic (seed pages) find out relevant pages from the web
- Pose Focused Crawling as a large scale OR problem

Ordinal Regression

- Fast OR training algorithm — scales to millions of datapoints
 - Fast algorithm to solve an SOCP with one SOC constraint
- Low prediction time

Baseline OR Formulation [Chu & Keerthi, 2005]



Clustering based scalable OR Formulation

- Describe data using **clusters** instead of data points

Clustering based scalable OR Formulation

- Describe data using **clusters** instead of data points
 - Class conditional distributions — mixture models with spherical covariance

Clustering based scalable OR Formulation

- Describe data using **clusters** instead of data points
 - Class conditional distributions — mixture models with spherical covariance
- Using second order moments $(\mu, \sigma^2 \mathbf{I})$, classify **clusters**

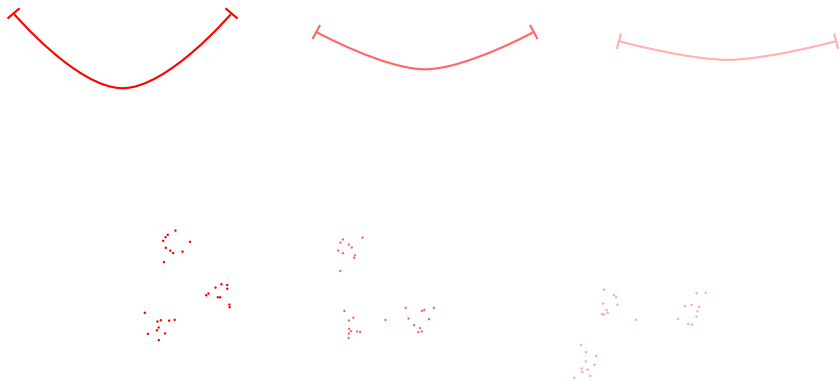
Clustering based scalable OR Formulation

- Describe data using **clusters** instead of data points
 - Class conditional distributions — mixture models with spherical covariance
- Using second order moments $(\mu, \sigma^2 \mathbf{I})$, classify **clusters**
- Proposed formulation will have constraints **per cluster**

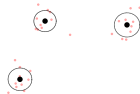
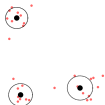
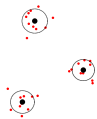
Clustering based scalable OR Formulation

- Describe data using **clusters** instead of data points
 - Class conditional distributions — mixture models with spherical covariance
- Using second order moments $(\mu, \sigma^2 \mathbf{I})$, classify **clusters**
- Proposed formulation will have constraints **per cluster**
- Size of optimization problem $O(\text{clusters})$ rather than $O(\text{datapoints})$

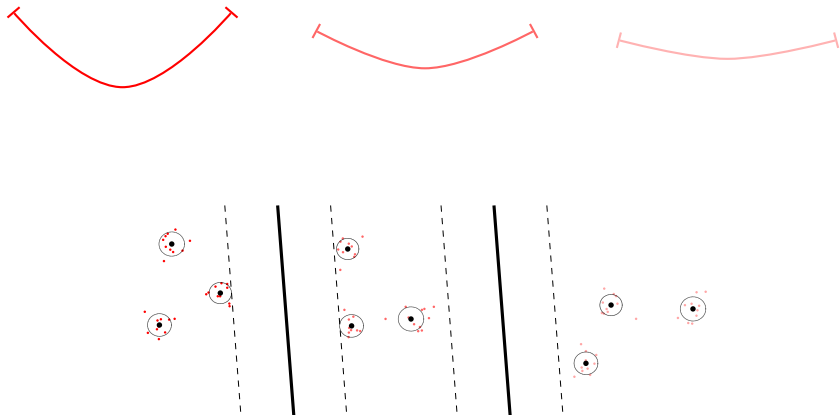
Proposed OR formulation's solution



Proposed OR formulation's solution



Proposed OR formulation's solution



Proposed OR formulation

Features:

- SOCP Problem with **one SOC constraint**
- $T_{train} = T_{clust} + T_{SOCP} = O(n)$
 - Cluster moments estimated using BIRCH [Zhang et.al., 1996]
 $T_{clust} = O(n)$
 - SOCP solved using SeDuMi^a. T_{SOCP} is **independent** of n
- Can be Kernelized — using input space cluster moments
 - No. of Support Vectors at max. k — low prediction time

^a<http://sedumi.mcmaster.ca/>

Clustering + SOCP gives speedup

Table: Training times (sec) with **SeDuMi** and **SMO-OR** [Chu & Keerthi, 2005] on synthetic dataset.

S-Rate	S-Size	SMO-OR	SeDuMi
0.002	10,000	182	1
0.0025	12,500	260	1
0.003	15,000	340	1
0.3	1,500,000	×	9
1	5,000,000	×	36

Table: Training times (sec), test error rate with **SeDuMi** and **SMO-OR** [Chu & Keerthi, 2005] on CS-Census dataset.

	S-Size	SMO-OR	SeDuMi
		sec (err)	sec
	5,690	893 (.128)	20.4 (.109)
	11,393	5281.6 (.107)	108.8 (.112)
	15,191	9997.5 (.107)	271.1 (.108)
CS	22,331	×	435.7 (.119)

Large number of clusters is still challenging

Table: Training times (sec), test error rate with **SeDuMi** and **SMO-OR** [Chu & Keerthi, 2005] on CH-California Housing dataset.

	S-Size	SMO-OR	SeDuMi
		sec (err)	sec
CH	10,320	551.9 (.619)	112 (.623)
	13,762	1033.2 (.616)	768.8 (.634)
	15,482	1142 (.617)	×
	17,202	1410 (.617)	×
	20,230	1838.5 (.62)	×

CB-OR Solver

Key Idea:

- Exploit special SOCP form — SOCP problem with **one SOC** constraint
 - Erdougan et.al., 2006 — specialized solvers scale better
- Fast algorithm similar in spirit to Platt's SMO for QP

Features:

- More scalable than generic solvers
- Easy to implement, uses no optimization tools

CB-OR Solver

Rewrite Dual as follows:

$$\begin{aligned} \min_{\alpha, \alpha^*} \quad & W \sqrt{(\alpha^* - \alpha)^\top \mathbf{K} (\alpha^* - \alpha)} - \mathbf{d}^\top (\alpha + \alpha^*) \\ \text{s.t.} \quad & 0 \leq \alpha \leq 1, 0 \leq \alpha^* \leq 1 \\ & s_i^* \leq s_i, \forall i = 1, \dots, r-2, s_{r-1}^* = s_{r-1} \end{aligned}$$

\mathbf{K} is Gram matrix for cluster centers

$$s_i = \sum_{k=1}^i \sum_{j=1}^{n_k} \alpha_k^j \quad \text{and} \quad s_i^* = \sum_{k=2}^{i+1} \sum_{j=1}^{n_k} \alpha_k^{*j}$$

CB-OR Solver

Minimization wrt. two multipliers

$$\begin{aligned} \min_{\Delta\alpha} \quad & \sqrt{a(\Delta\alpha)^2 + 2b(\Delta\alpha) + c} - e\Delta\alpha \\ \text{s.t.} \quad & lb \leq \Delta\alpha \leq ub \end{aligned}$$

Has closed form solution:

$$\Delta\alpha = \begin{cases} \left[\frac{e\sqrt{\frac{ac-b^2}{a-e^2}} - b}{a} \right]_{lb}^{ub} & \text{if } ac - b^2 > 0, a - e^2 > 0 \\ \left[\frac{-b}{a} \right]_{lb}^{ub} & \text{if } ac - b^2 = 0, a - e^2 > 0 \\ ub & \text{if } e - \sqrt{a} \geq 0 \\ lb & \text{if } e + \sqrt{a} \leq 0 \end{cases}$$

CB-OR Solver

CB-OR Algorithm

- Step 1 Pick two most KKT violators
- Step 2 Solve the 1-d minimization problem
- Step 3 Update unknowns
- Step 4 Check for KKT violators. If none terminate. Else Step 1

CB-OR — Evaluation

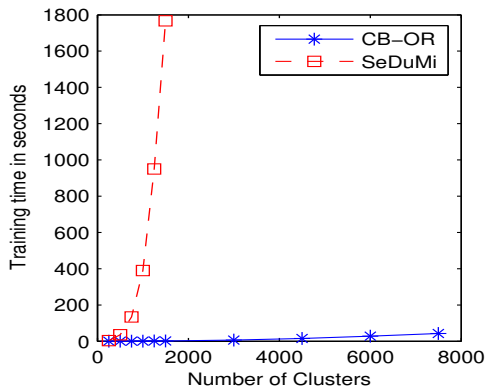


Figure: Dashed line represents training time with **SeDuMi** and continuous line that with **CB-OR** on a synthetic dataset.

CB-OR — Evaluation

Table: Comparison of training times (in sec) with **CB-OR**, **SMO-OR** and **SeDuMi** on benchmark datasets. The test set error rate is given in brackets. (CH-California Housing, CS-Census datasets).

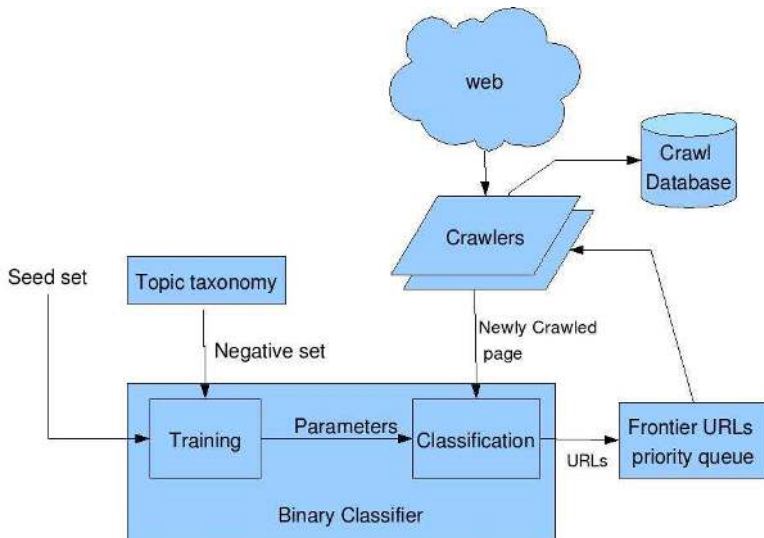
	S-Size	CB-OR	SMO-OR	SeDuMi
		sec (err)	sec (err)	sec
CH	10,320	.5 (.623)	551.9 (.619)	112
	13,762	1.5 (.634)	1033.2 (.616)	768.8
	15,482	8.4 (.618)	1142 (.617)	×
	17,202	14.3 (.621)	1410 (.617)	×
	20,230	10.4 (.62)	1838.5 (.62)	×
CS	5,690	.3 (.109)	893 (.128)	20.4
	11,393	.7 (.112)	5281.6 (.107)	108.8
	15,191	1 (.108)	9997.5 (.107)	271.1
	22,331	1.5 (.119)	×	435.7

Focused Crawling

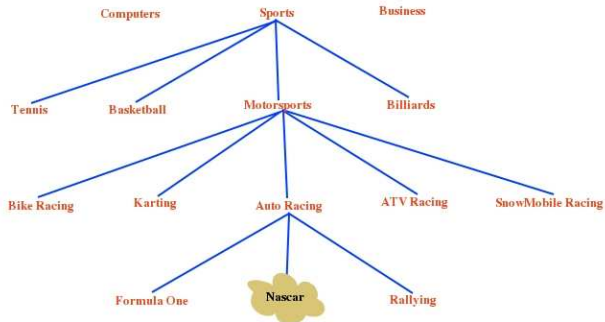
Focused Crawling

- Given a topic (seed pages) find out relevant pages from the web.
- S. Chakrabarti et.al (1999,2002), C. Aggarwal et.al (2001), M. Diligenti et.al (2000)
- Requires low bandwidth and low disk space.
- Small updation cycle.

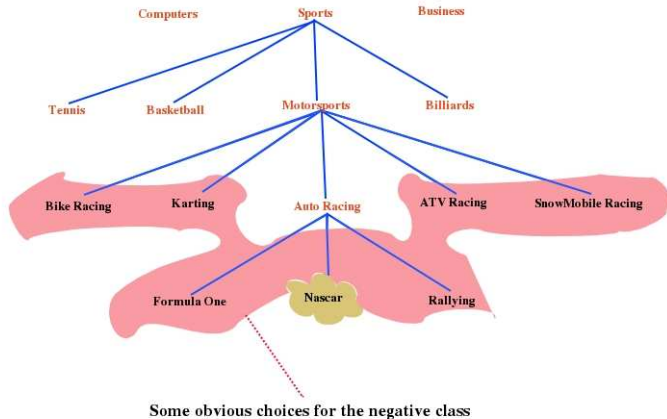
Baseline Focused Crawler [Chakrabarti et.al., 1999]



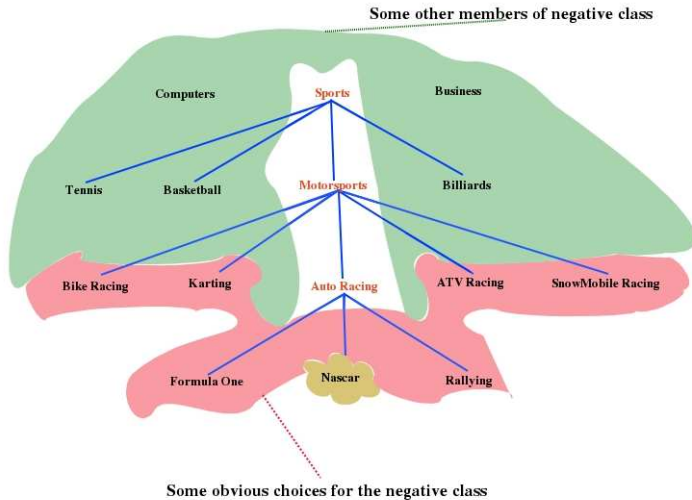
Topic Taxonomy



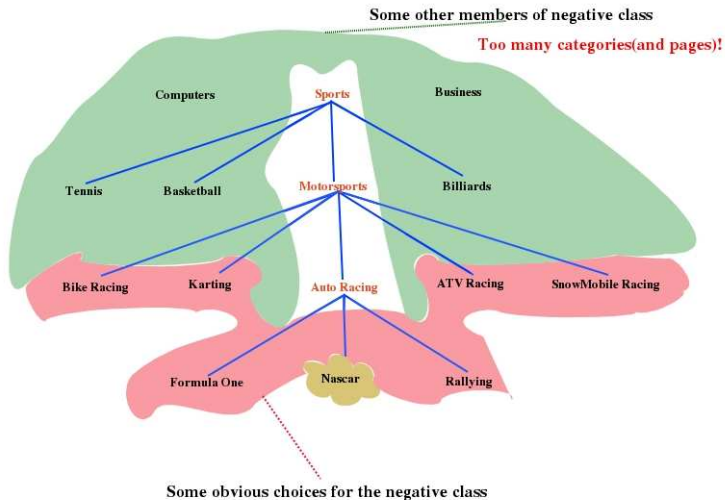
Topic Taxonomy



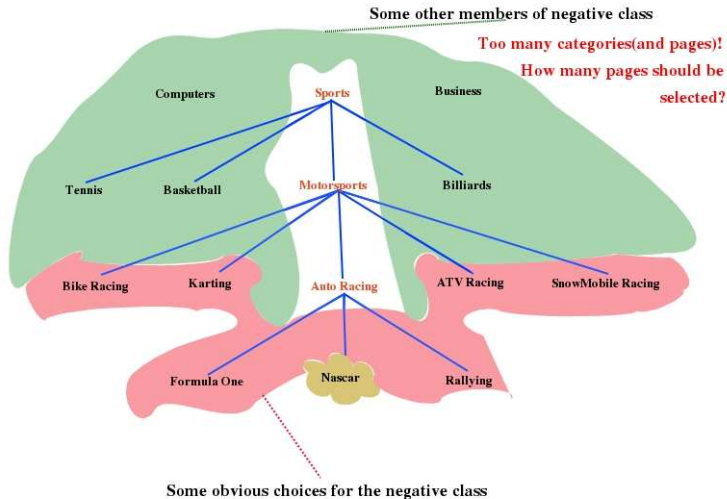
Topic Taxonomy



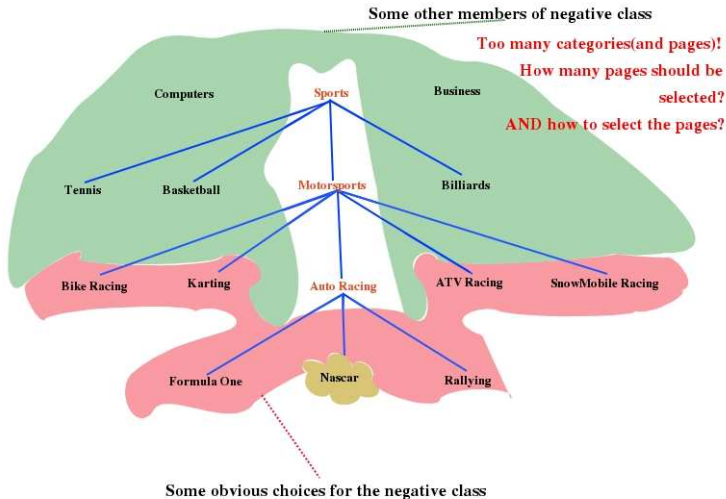
Topic Taxonomy



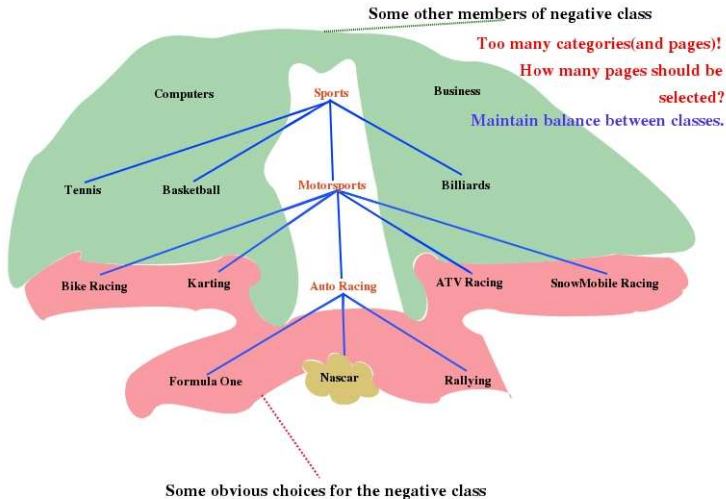
Topic Taxonomy



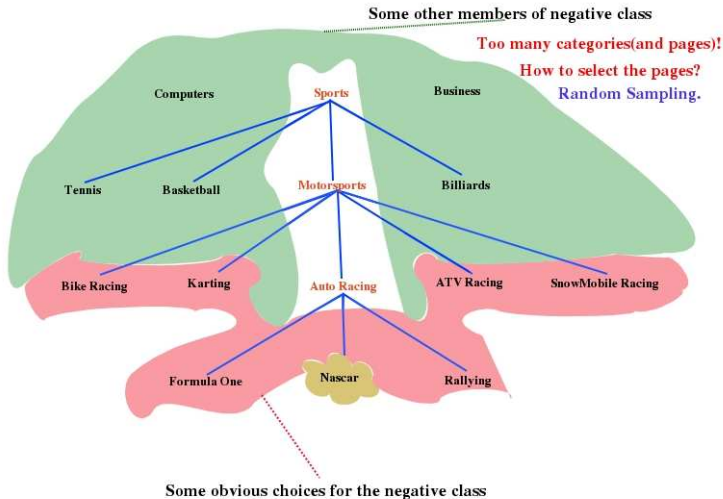
Topic Taxonomy



Topic Taxonomy



Topic Taxonomy



Exploit link structure

- Grangier and Bengio observe that hyperlinked documents are semantically closer.
- One link away pages are more similar to seed pages compare to two link away pages.

Link structure in web

The image shows a screenshot of the Yahoo! homepage. At the top, there is a navigation bar with links for "Web", "Images", "Video", "Local", "Shopping", and "more". Below this is a search bar with the text "Search:" and a "Y! Answers" link. The main content area is divided into sections: "Featured", "Entertainment", "Sports", and "Life". The "Sports" section is active, displaying a featured article titled "Roy the role model" about NBA Rookie of the Year Brandon Roy. Below the article, there are several smaller links, including "Role model Roy more than just best rookie", "NASCAR: Waltrip survives scary accident", and "Which team will win the NHL Stanley Cup?". The "NASCAR" and "Which team will win the NHL Stanley Cup?" links are circled in red. On the right side, there is a "Check your" section with links for "Mail", "Weather 34°F", "Yahoo! Shopping", and "Electronics". The left sidebar contains various utility links like "Autos", "Finance", "Games", "GeoCities", "Groups", "HotJobs", "Maps", "Movies", "Music", "Personals", "Photos", "Real Estate", and "Shopping".

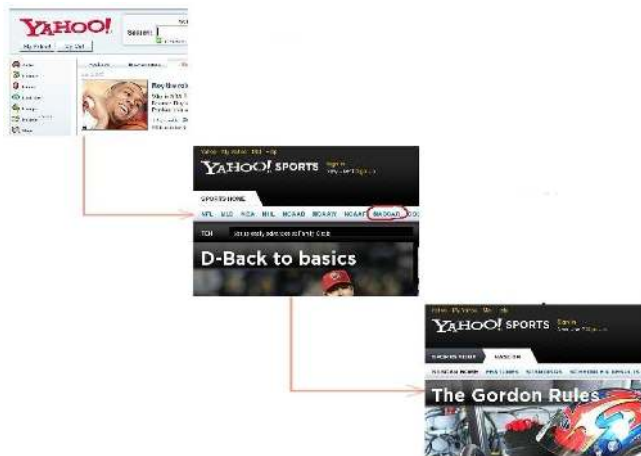
Link structure in web

The image shows a screenshot of the Yahoo! Sports website. A blue arrow points from a smaller, partially visible Yahoo! page in the top left to the main page. The main page features a navigation bar with the following links: NFL, MLB, NBA, NHL, NCAAB, NCAAW, NCAAF, **NASCAR** (circled in red), GOLF, TENNIS, SOCCER, and ALLSPORTS. Below the navigation bar, there is a search bar and a main content area. The main content area includes a large article titled "D-Back to basics" featuring a photo of a baseball pitcher. To the right of the main article, there are two smaller sections: "NHL playoffs" with a photo of a hockey player and a list of links, and "Decision day" with three small portraits of men.

Link structure in web



Focused Crawling as OR problem — exploit link structure



Focused Crawling as OR problem — exploit link structure



Focused Crawling as OR problem — exploit link structure



Level 1 - Page has many links to level 0 pages(Hub)



Level 0 - Pages belong to topic

Focused Crawling as OR problem — exploit link structure



Level 2 - Some of the links on this page will lead to topic pages.

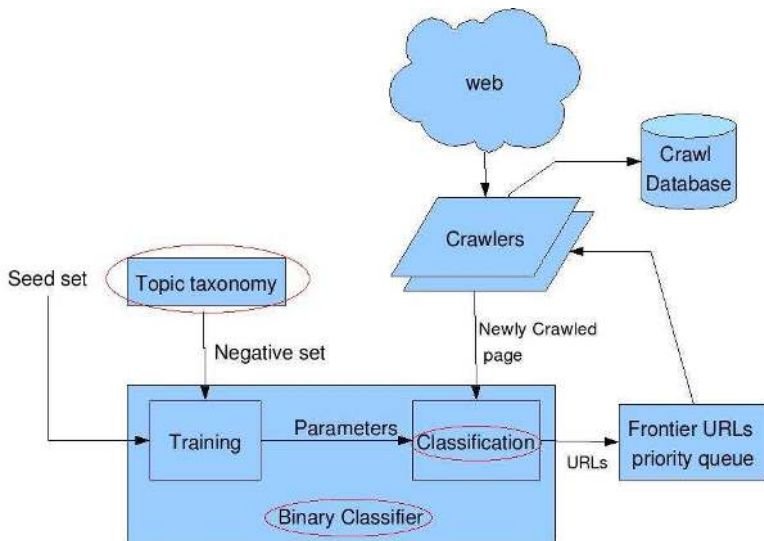


Level 1 - Page has many links to level 0 pages(Hub)

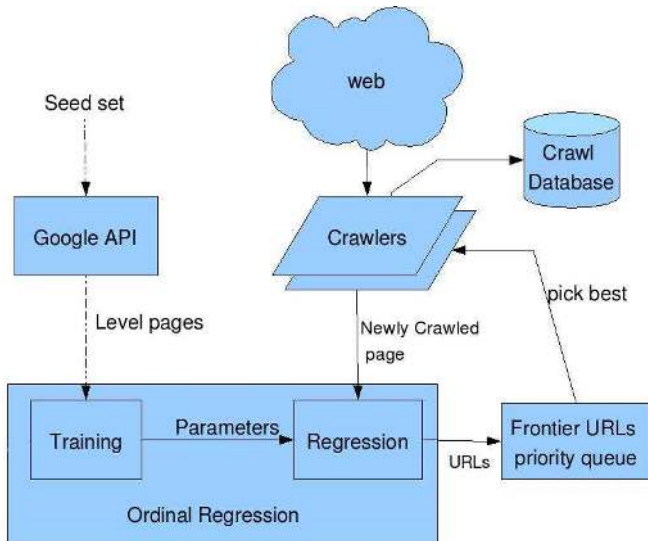


Level 0 - Pages belong to topic

Baseline Focused Crawling architecture



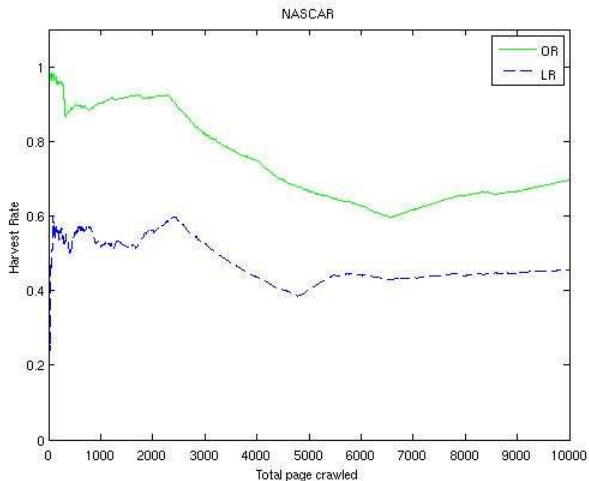
Proposed Focused Crawling architecture



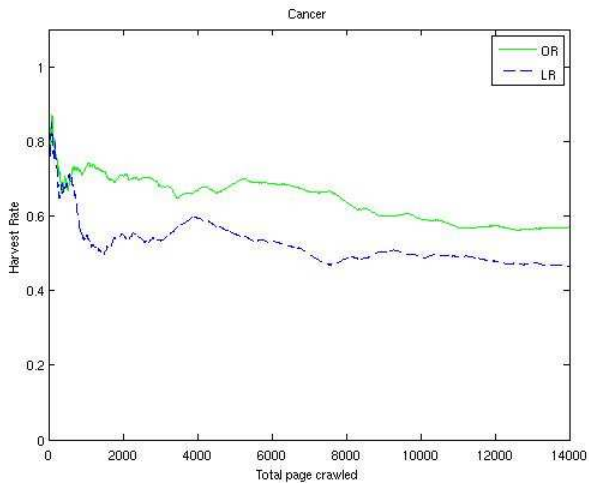
Focused Crawling is a large scale OR problem

Category	Seed	1	2	3	4
NASCAR	1705	1944	1747	1464	1177
Soccer	119	750	1109	1542	3149
Cancer	138	760	895	858	660
Mutual Funds	371	395	540	813	1059

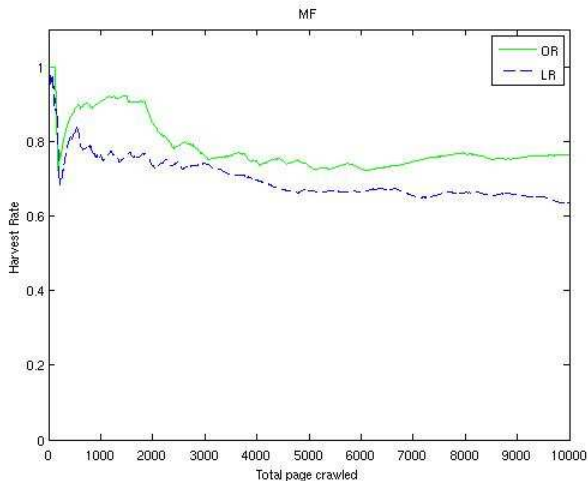
NASCAR harvest rate



Cancer harvest rate



Mutual Funds harvest rate



Harvest rate comparison

Dataset	Baseline	OR
NASCAR	.3698	.6977
Cancer	.4714	.58
Mutual Fund	.526	.5969
Soccer	.34	.4952

Conclusions

- Proposed a scalable clustering based OR formulation
 - Training time $O(\text{datapoints})$
 - Support Vectors $O(\text{clusters})$
- Exploited special structure of the formulation to develop a fast solver, CB-OR
 - Scalable to tens of thousands of **clusters**
- We formulated focused crawling as large scale ordinal regression
 - No need for negative class definition
 - **Independent** of topic taxonomy
 - OR captures **link structure** of web graph.

Focused crawler code available at

<http://mllab.csa.iisc.ernet.in/downloads/focusedcrawler.html>

Acknowledgments

This project is partially supported by AOL India Pvt Ltd and DST, Government Of India (DST/ECA/CB/660)

Questions?