Topic-Sensitive *Hidden-Web* Crawling

Panagiotis Liakos Alexandros Ntoulas

University of Athens

WISE – Paphos, November 2012

Motivation

- *Hidden-Web*
  - Considerably larger than the Surface Web [Ber00, CHL+04]
  - High quality information [Ber00]
  - Variety of topics
- We are often interested only in a small portion of a *Hidden-Web* site
  - Portal talking about politics
  - Mobile application focusing on the US Presidential Elections

Motivation

- *Hidden-Web*
  - Considerably larger than the Surface Web [Ber00, CHL+04]
  - High quality information [Ber00]
  - Variety of topics

- We are often interested only in a small portion of a *Hidden-Web* site
  - Portal talking about politics
  - Mobile application focusing on the US Presidential Elections

Problem:

How can we efficiently retrieve the interesting portion of a *Hidden-Web* site?

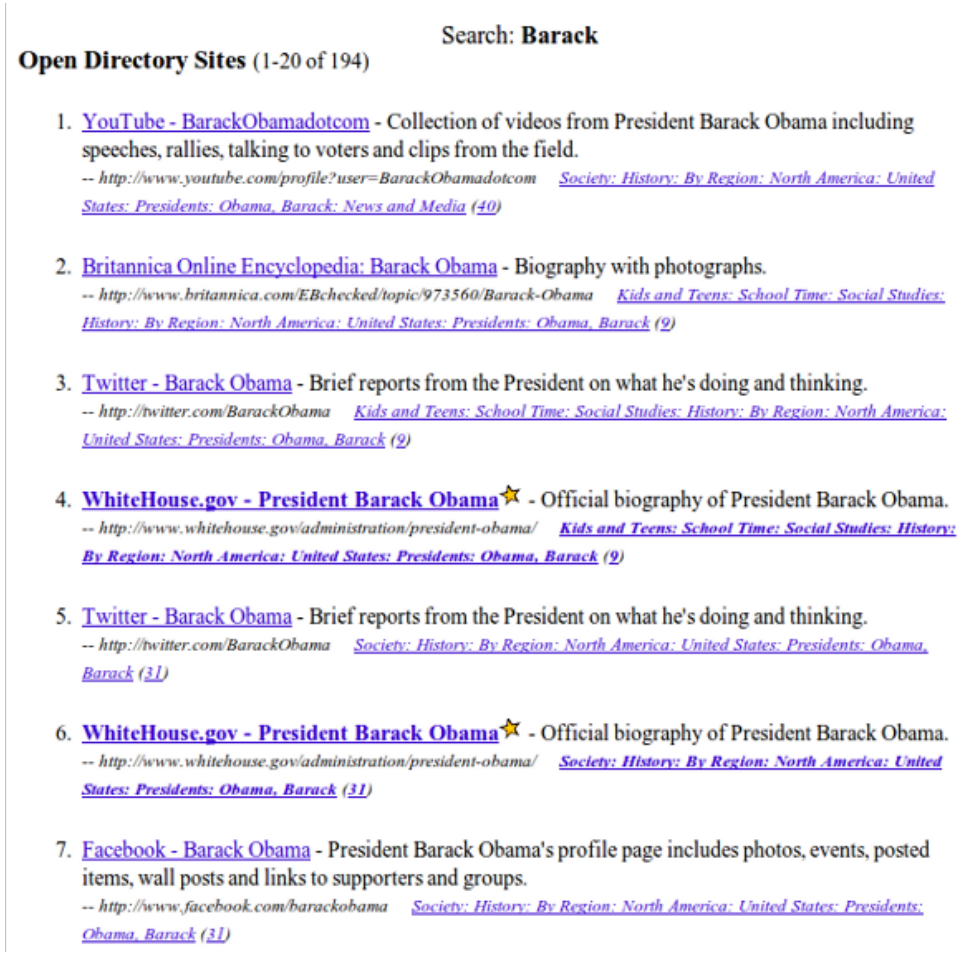
Interacting with a *Hidden-Web* site

1. User submits a query through a search interface
2. User receives a result index page
3. User navigates to a site of her choice

Open Directory Project search interface

Interacting with a *Hidden-Web* site

1. User submits a query through a search interface
2. User receives a result index page
3. User navigates to a site of her choice



Search: **Barack**

Open Directory Sites (1-20 of 194)

1. YouTube - BarackObamadotcom - Collection of videos from President Barack Obama including speeches, rallies, talking to voters and clips from the field.
   -- *http://www.youtube.com/profile?user=BarackObamadotcom* *Society: History: By Region: North America: United States: Presidents: Obama, Barack: News and Media (40)*

2. Britannica Online Encyclopedia: Barack Obama - Biography with photographs.
   -- *http://www.britannica.com/EBchecked/topic/973560/Barack-Obama* *Kids and Teens: School Time: Social Studies: History: By Region: North America: United States: Presidents: Obama, Barack (9)*

3. Twitter - Barack Obama - Brief reports from the President on what he's doing and thinking.
   -- *http://twitter.com/BarackObama* *Kids and Teens: School Time: Social Studies: History: By Region: North America: United States: Presidents: Obama, Barack (9)*

4. **WhiteHouse.gov - President Barack Obama** - Official biography of President Barack Obama.
   -- *http://www.whitehouse.gov/administration/president-obama/* ***Kids and Teens: School Time: Social Studies: History: By Region: North America: United States: Presidents: Obama, Barack (9)***

5. Twitter - Barack Obama - Brief reports from the President on what he's doing and thinking.
   -- *http://twitter.com/BarackObama* *Society: History: By Region: North America: United States: Presidents: Obama, Barack (31)*

6. **WhiteHouse.gov - President Barack Obama** - Official biography of President Barack Obama.
   -- *http://www.whitehouse.gov/administration/president-obama/* ***Society: History: By Region: North America: United States: Presidents: Obama, Barack (31)***

7. Facebook - Barack Obama - President Barack Obama's profile page includes photos, events, posted items, wall posts and links to supporters and groups.
   -- *http://www.facebook.com/barackobama* *Society: History: By Region: North America: United States: Presidents: Obama, Barack (31)*

Open Directory Project result page

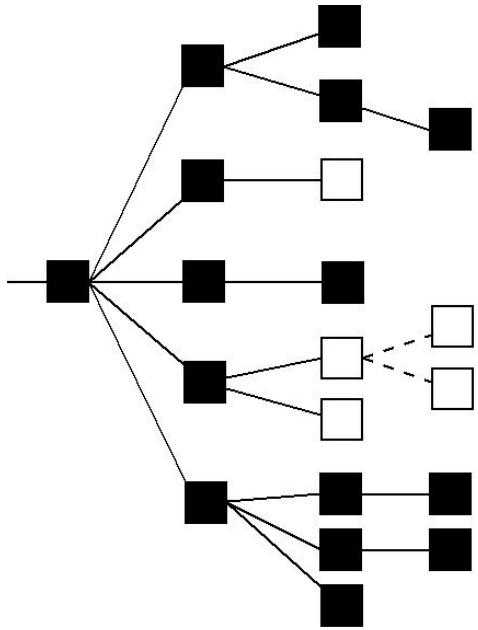
Interacting with a *Hidden-Web* site

1. User submits a query through a search interface
2. User receives a result index page
3. User navigates to a site of her choice

Related Website

Differences with the Surface Web

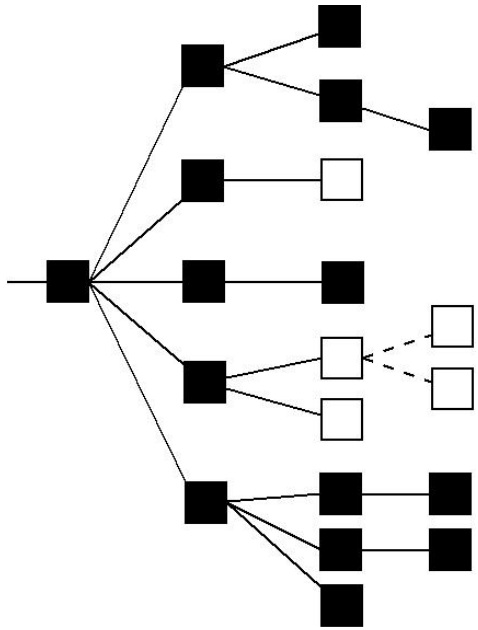
- Links - Queries
- Focused Crawlers
  - Follow links of **relevant** pages
  - Evaluate the content of a page to estimate the possibility that a link is **useful**



Focused Crawling

Differences with the Surface Web

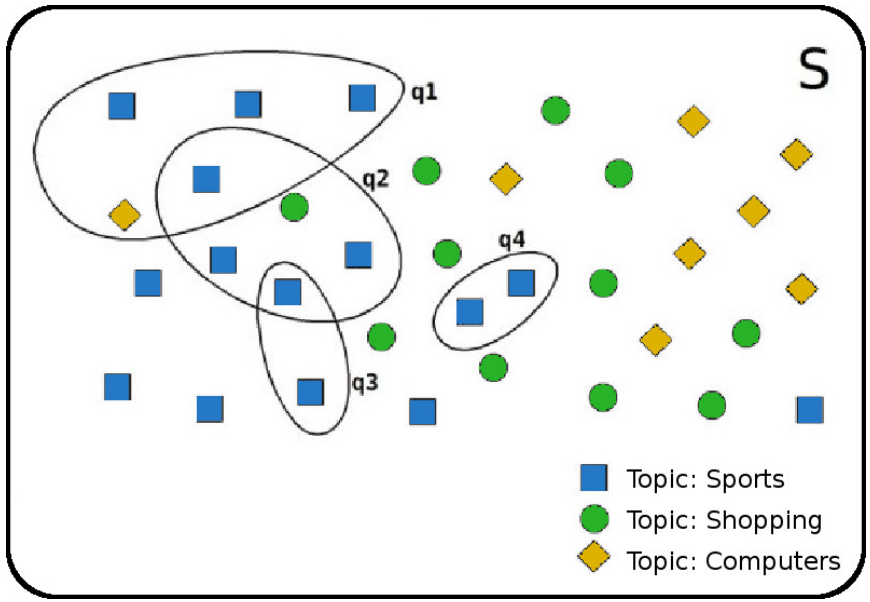
- Links - Queries
- Focused Crawlers
  - Follow links of **relevant** pages
  - Evaluate the content of a page to estimate the possibility that a link is **useful**
- Focused *Hidden-Web* Crawlers?



Focused Crawling

How can we select appropriate queries?

- S: set of pages in a *Hidden-Web* site
- q_i: set of pages returned after submitting q_i
- each q_i uses up resources (bandwidth, cpu etc.)



A *Hidden-Web* site as a set

Goal:

Retrieve all the pages for a given topic using the minimum resources

Algorithm

Algorithm 1 Pseudocode for a Topic-Sensitive *Hidden-Web* Crawler

while (available resources) **do**

q_i = selectTerm(WordCollection); (1)

R(q_i) = submitAndDownload(q_i); (2)

update(WordCollection); (3)

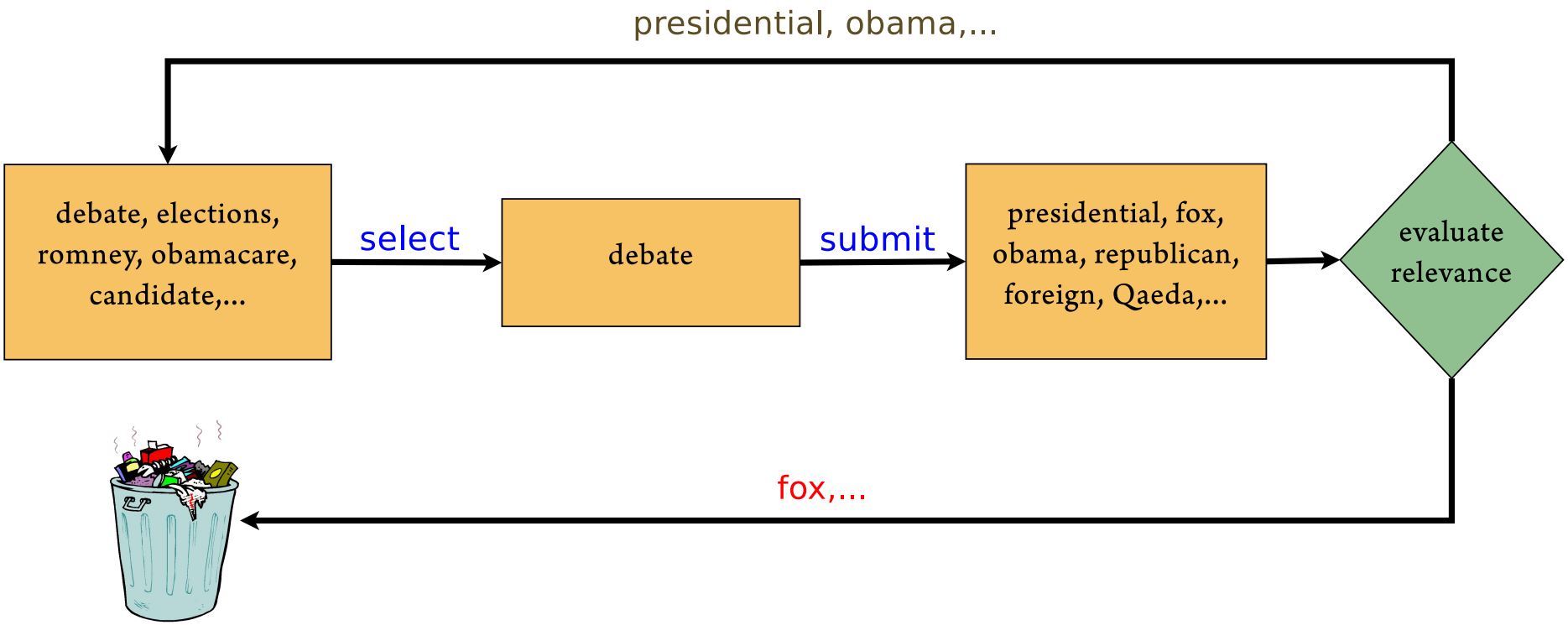
end while

Word Collection

- Pool of words for a specific topic
  - e.g. US Presidential Elections
- Initialization with an exemplary document
  - A few articles on the Barack Obama - Mitt Romney debates

- Must not be static!
  - Too small
  - Too specific
  - Adaptability is crucial

Maintaining a good Word Collection



Relevance Evaluation Policies

- **perfect:**
  - Uses categorization information from the *Hidden-Web* site itself
- **do-nothing:**
  - Accepts all results
- **NaiveBayes:**
  - Uses a Naive Bayes classifier for text categorization
- **CosineSimilarity:**
  - Examines the cosine similarity of every document with the initial one

Relevance Evaluation Policies

- **perfect:**
  - Uses categorization information from the *Hidden-Web* site itself
- **do-nothing:**
  - Accepts all results
- **NaiveBayes:**
  - Uses a Naive Bayes classifier for text categorization
  - Needs a training dataset
- **CosineSimilarity:**
  - Examines the cosine similarity of every document with the initial one

Relevance Evaluation Policies

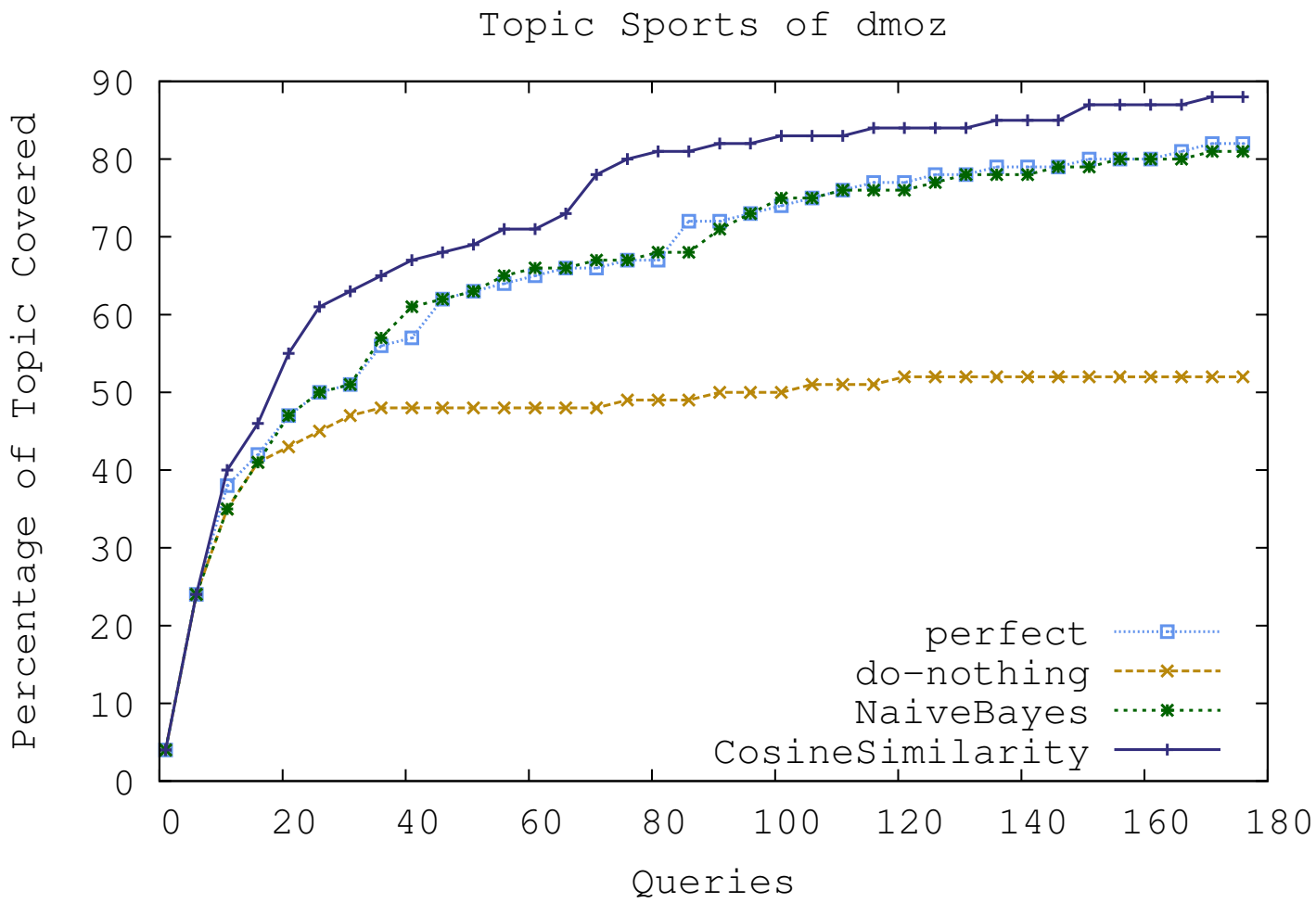
- **perfect:**
  - Uses categorization information from the *Hidden-Web* site itself
- **do-nothing:**
  - Accepts all results
- **NaiveBayes:**
  - Uses a Naive Bayes classifier for text categorization
  - Needs a training dataset
- **CosineSimilarity:**
  - Examines the cosine similarity of every document with the initial one
  - Quality of the initial document may affect performance

Experimental Setup

- *Open Directory Project* [1]
  - ≈ 5 million pages
- Public non-beta *Stack Exchange* [2] sites
  - $\approx 400,000$ questions
- Exemplary Documents
  - Relevant snippets from the site in search
- Performance Metric
  - Percentage of relevant documents retrieved

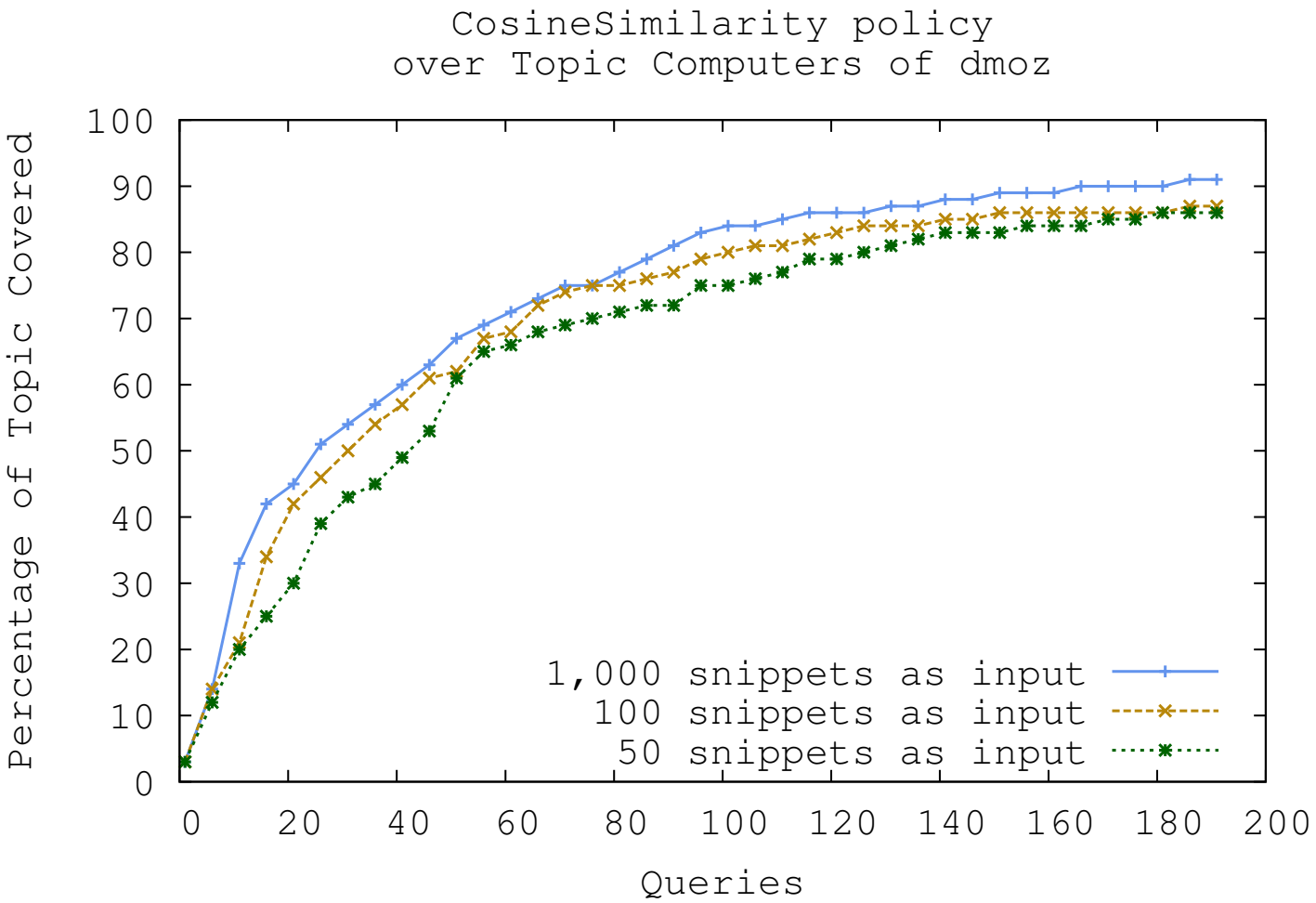
[1] `http://www.dmoz.org`
[2] `http://stackexchange.com`

Evaluation of different policies over the same topic



- *CosineSimilarity* outperforms the rest policies

Impact of input document size



- Almost insignificant impact on the performance of the policy

Comparison to Generic *Hidden-Web* Crawling

- Previous experiments [NZC05]: ≈**700 queries** needed to retrieve 70% of the *Open Directory Project* contents
- Topic Sports: **52** queries for 70%
- Topic Computers: **60** queries for 70%

Related Work

- Discovery of *Hidden-Web* forms [RGM01, BF07]
- Producing Meaningful Queries [NZC05, BF04]
  - Attempt to download all of a *Hidden-Web* site
- Focused Crawlers [CvdBD99]
  - Incombatible with the *Hidden-Web*

Future - Ongoing work

- Diverse query formulations
- Additional *Hidden-Web* sites
- Retrieve recently updated content incrementally

References


Michael K. Bergman, *The deep web: Surfacing hidden value*, 2000.


Luciano Barbosa and Juliana Freire, *Siphoning hidden-web data through keyword-based interfaces*, In SBBD, 2004, pp. 309–321.


Luciano Barbosa and Juliana Freire, *An adaptive crawler for locating hidden-web entry points*, Proceedings of the 16th international conference on World Wide Web (New York, NY, USA), WWW '07, ACM, 2007, pp. 441–450.


Kevin Chen-Chuan Chang, Bin He, Chengkai Li, Mitesh Patel, and Zhen Zhang, *Structured databases on the web: observations and implications*, SIGMOD Rec. **33** (2004), 61–70.


Soumen Chakrabarti, Martin van den Berg, and Byron Dom, *Focused crawling: a new approach to topic-specific web resource discovery*, Proceedings of the eighth international conference on World Wide Web (New York, NY, USA), WWW '99, Elsevier North-Holland, Inc., 1999, pp. 1623–1640.


Alexandros Ntoulas, Petros Zerfos, and Junghoo Cho, *Downloading textual hidden web content through keyword queries*, Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries (New York, NY, USA), JCDL '05, ACM, 2005, pp. 100–109.


Sriram Raghavan and Hector Garcia-Molina, *Crawling the hidden web*, Proceedings of the 27th International Conference on Very Large Data Bases (San Francisco, CA, USA), VLDB '01, Morgan Kaufmann Publishers Inc., 2001, pp. 129–138.

thank you

Queries Issued and Topic Accuracy

No	Term	Precision
1	results	42.83%
2	statistics	43.61%
3	roster	71.36%
10	men	27.26%
15	scores	27.89%
20	players	30.62%
25	hockey	41.85%
30	tennis	7.43%
40	rugby	14.43%
60	sport	3.82%
100	competition	12.28%

(a) *Perfect*

Term	Precision
results	42.83%
statistics	43.61%
roster	71.36%
schedules	10.31%
church	0.00%
coaching	17.89%
methodist	0.00%
beliefs	10.11%
stellt	0.00%
bietet	0.00%
nach	0.00%

(b) *Do-nothing*

Term	Precision
results	42.83%
statistics	43.61%
roster	71.36%
schedules	10.31%
standings	67.96%
baseball	38.29%
records	8.38%
membership	1.52%
county	0.39%
fc	5.45%
standing	26.66%

(c) *NaiveBayes*

Term	Precision
results	42.83%
statistics	43.61%
roster	71.36%
tables	5.61%
player	24.36%
players	33.70%
hockey	44.08%
baseball	32.20%
race	14.56%
conference	1.73%
competitive	11.61%

(d) *CosineSimilarity*

Queries issued by the different policies