Focused crawling: a new approach to topic-specific Web resource discovery

Authors
Soumen Chakrabarti
Martin van den Berg
Byron Dom

Presented By: Mohamed Ali Soliman
m2ali@cs.uwaterloo.ca
Outline

- Why Focused Crawling?
- Contributions
- Applications
- System Architecture
- Evaluation
- Related Work
- Comments
Why Focused Crawling?

- Current general crawlers operate with high cost.
- They have a limited coverage of the web.
- Huge web growth should not affect users with specific interests.
- Huge index size is undesired when the task is to find focused resources.
Outline

• Why Focused Crawling?
• Contributions
• Applications
• System Architecture
• Evaluation
• Related Work
• Comments
Contributions

- Reduce network and hardware crawling costs.
- Provide the ability to manage web content using a distributed team of focused crawlers.
- Control the crawler behavior using other integrated hypertext mining processes:
  - Classifier
  - Distiller
Outline

• Why Focused Crawling?
• Contributions
• Applications
• System Architecture
• Evaluation
• Related Work
• Comments
Potential Applications

• Discovering linkage sociology.
• Locating highly relevant sites.
• Enriching training base for human-supervised topic learning.
• Detecting community behavior.
• Estimating topic change rate.
Outline

• Why Focused Crawling?
• Contributions
• Applications
• System Architecture
• Evaluation
• Related Work
• Comments
Focused Crawler (User’s View)

- Backbone Taxonomy
- User Selected Files
- Classifier
  - Classes proposed by system
- User Refinements
  - User tailored class taxonomy

User Feedback

Distiller

Crawler

Classifier

Classifier Models Updates

Classifier Knowledge base

Training

Operation

Focused Crawling
Focused Crawler (User’s View)

- **Classify**
- **Page currently viewed**
- **Examples from selected topic**
- **Neighboring pages can be added to examples by drag and drop**
- **Interesting topics are marked**

2/8/05  Focused Crawling
System Architecture

- **Classifier**: makes relevance judgments on pages crawled to decide on expanding links found in these pages.

- **Distiller**: determines a measure of centrality of crawled pages to determine visit priorities.

- **Crawler**: allows dynamically reconfigurable priority controls by the classifier and distiller.
System Architecture

Focused Crawling

- Predefined class hierarchy
- Calculate words probabilities within different classes contexts
- Words statistics
- Calculate relevance scores for new pages
- Get new URLs to be crawled
- Crawl new URLs received from Watchdog
- Distiller operates periodically to select top hubs to visit

Browser-based Administration Interface

Mark Ratings

Pick URLs

Watchdog Priority Controls

Memory Buffers

Worker Threads

Classifier (Training)

Topic Models

Classifier (Filtering)
Classifier

• Given a document, what is the probability that it belongs to some class?
  – Given a document $d$ and a set of predefined classes \( \{c_i; i=1..n\} \), calculate $Pr(c_i|d)$; $i=1..n$

• **Hard Classification**: Select the class with the maximum probability.

• **Soft Classification**: Produce a ranked list of classes according to probabilities.
Bayes Classifier

- $Pr(class|doc) = Pr(doc|class) \times Pr(class)/Pr(doc)$
  - $Pr(class)$: frequency of class documents inside collection.
  - $Pr(doc) = \sum_{i=1}^{n} Pr(doc | c_i) \times Pr(c_i)$
  - $Pr(doc|class)$ ??
Classifier

**Bayes Classifier** [McCallum, 1998]

- Multinomial Model
  - Document is generated by independently selecting words from a *bag of words* representing combined vocabulary for all classes.
  - A document occurrence probability, given some class, is the product of occurrence probabilities of its words within the context of that class.
Classifier

Bayes Classifier [McCallum et al., 1998]

• Multinomial Model

\[
\Pr(d \mid c_i) = \Pr(\mid d \mid)^* \mid d \mid \prod_{t \in d} \frac{\theta(c, t)^{n(d, t)}}{n(d, t)!}
\]

- \(n(d, t)\): Number of occurrences of word \(t\) inside document \(d\).
- \(\theta(c, t)\): Occurrence probability of word \(t\) inside class \(c\).

- For each class, the classifier stores \(\theta(c, t)\) for each vocabulary word \(t\), and uses that to calculate the tested document occurrence probability.
Distiller

- For each visited document $d$, the classifier produces a relevance score $R(d)$ that is used to give future crawl priorities.
- In addition, hub pages that point to authoritative sources need to be located.
- Due to web authorship diversity, relevant pages could point to irrelevant ones, e.g. pointing to famous search engine or html editors.
Distiller

- Assign non unit weights to edges.
- Edges are grouped into forward and backward:
  - $E_F[u,v] = R(v)$
  - $E_B[u,v] = R(u)$
- Iterate over graph nodes updating edges weights.
- A threshold $\rho$ is used to include potential authorities with high enough relevance scores.
Evaluation

Unfocused  Hard Focused  Soft Focused

Moving Average of Relevance

2/8/05  Focused Crawling
Evaluation

URL overlap between 2 crawlers using disjoint startup URL sets
Evaluation

Distance between top servers and seed URL sets
Outline

• Why Focused Crawling?
• Contributions
• Applications
• System Architecture
• Evaluation
• Related Work
• Comments
Related Work

Accelerated Focused Crawling [Chakrabarti et al., 2002]

• Only a fraction of out-links from a page are worth following.
• Documents were modeled as tag trees using DOM (Document Object Model).
• The text surrounding hyperlinks is used to decide on the relevance of target pages to be crawled before actually crawling them.
Related Work

Accelerated Focused Crawling [Chakrabarti et al., 2002]
Related Work

Classifying Web Pages using Links only

[Furnkranz, 2001]

- Hyperlinks that point to test documents are used as indicators for the classes of these documents.
- Diversity of web authorship is used to make good predictions.
- Different combinations of anchor, headings, paragraph and phrases feature sets derived from hyperlinks were used to make class predictions.
- Accuracy ranged from 57% to 87% for different methods used for combining links predictions.
Outline

• Why Focused Crawling?
• Contributions
• Applications
• System Architecture
• Evaluation
• Related Work
• Comments
Comments

• No consideration was made to using different kinds of text classifiers.
• Using the embedded classifier to judge crawl relevance is unconvincing.
• The scheme used by the crawler to refresh the contents of crawled pages is not described.
• Results were illustrated using mainly two classes although calculating overall estimates using all classes was possible.
References


