Introduction to Data Mining I

Instructor: Dmitri A. Gusev

Fall 2007

CS 502: Computers and Communications Technology

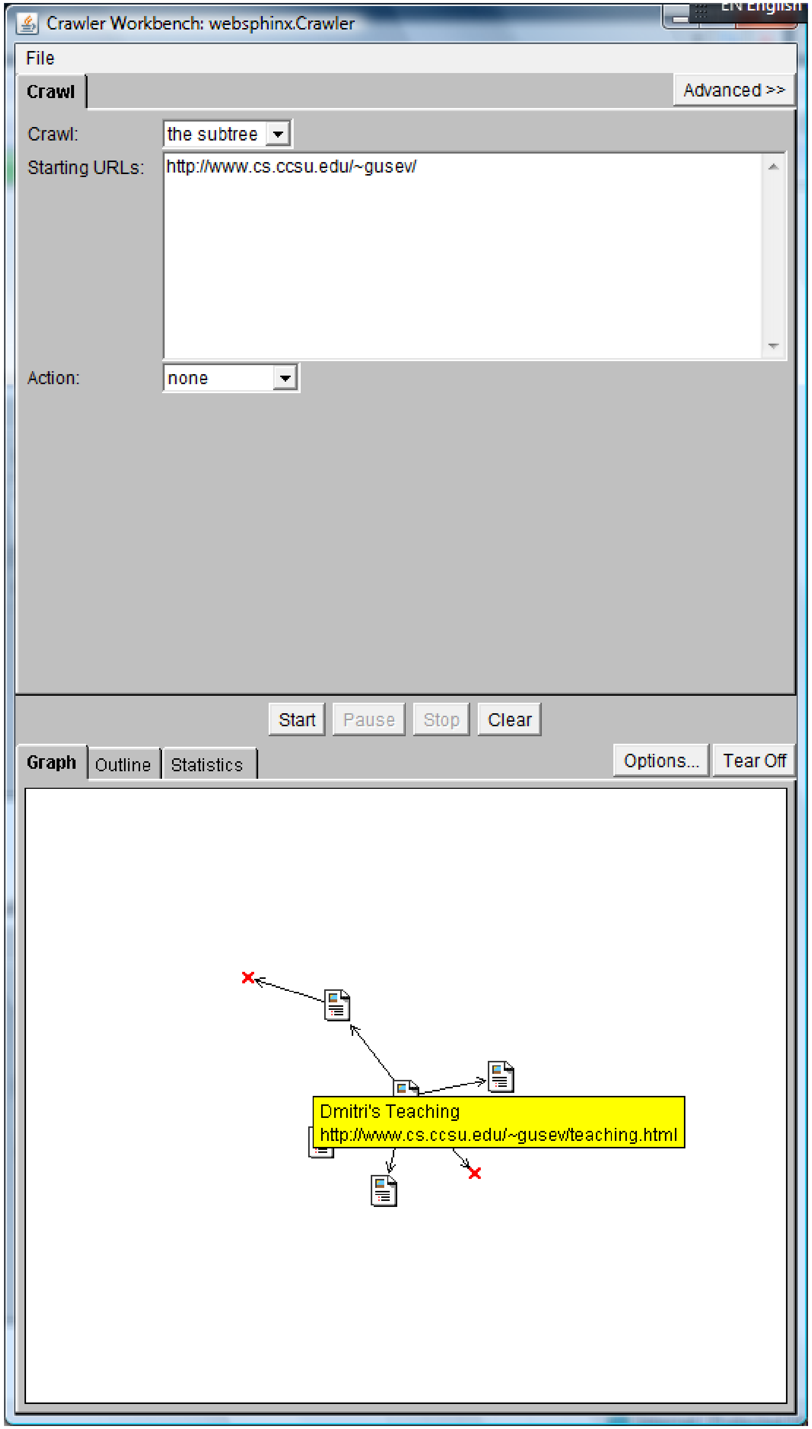
Lecture 26, December 10, 2007

Crawling the Web

- A *web crawler* (also called a *robot* or *spider*) is a program that browses and processes Web pages automatically.
- Web crawler example: http://www.cs.cmu.edu/~rcm/websphinx/

  WebSPHINX: A Personal, Customizable Web Crawler; WebSPHINX consists of two parts: the Crawler Workbench and the WebSPHINX class library. WebSPHINX is designed for advanced web users and Java programmers who want to crawl over a small part of the web (such as a single web site) automatically. Using the Crawler Workbench, you can:

  - Visualize a collection of web pages as a graph
  - Save pages to your local disk for offline browsing
  - Concatenate pages together for viewing or printing them as a single document
  - Extract all text matching a certain pattern from a collection of pages; extract images from a set of pages
  - Develop a custom crawler in Java or Javascript that processes pages however you want.



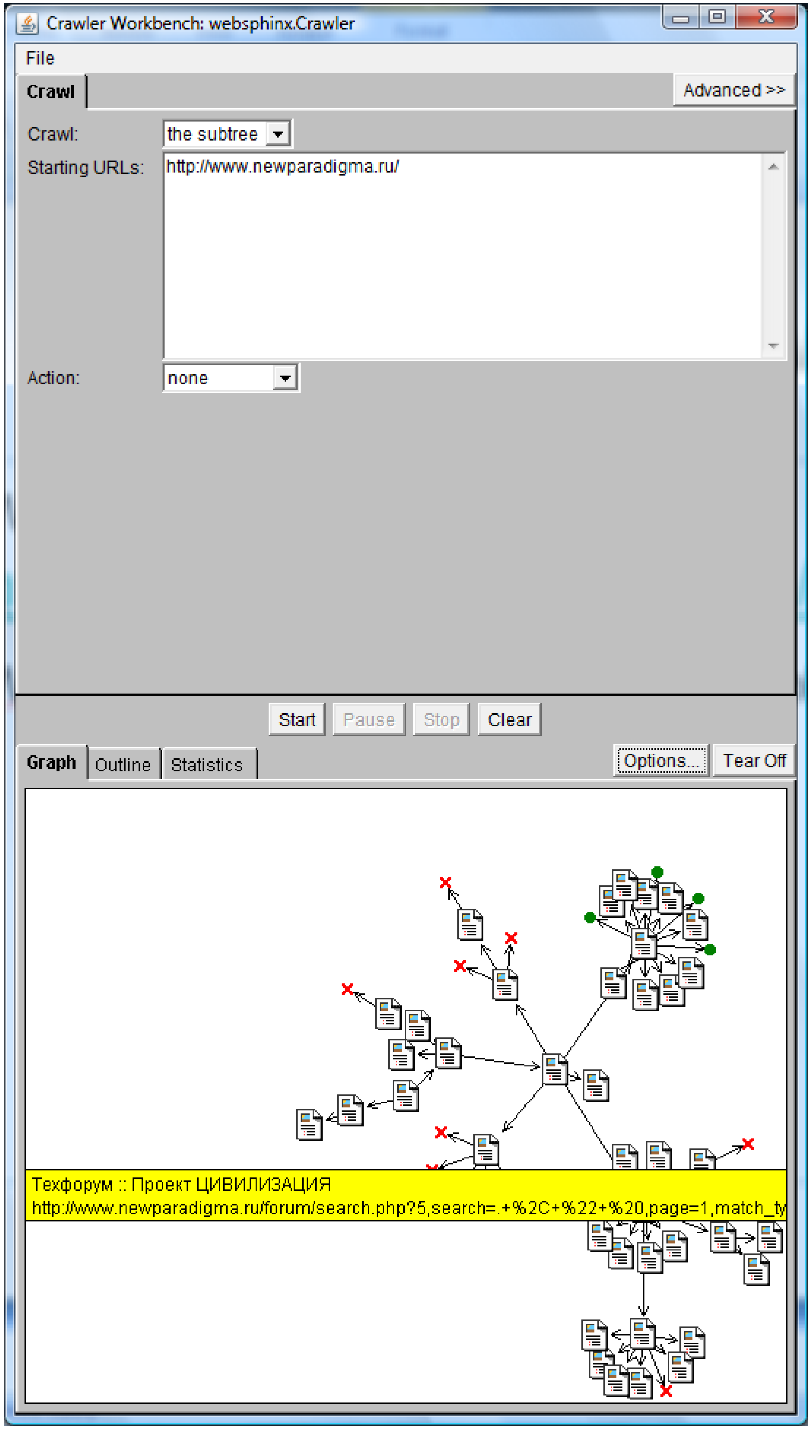
Crawler Workbench: websphinx.Crawler

Crawl: the subtree

Starting URLs: http://www.cs.ccsu.edu/~gusev/

Action: none

Dmitri's Teaching
http://www.cs.ccsu.edu/~gusev/teaching.html



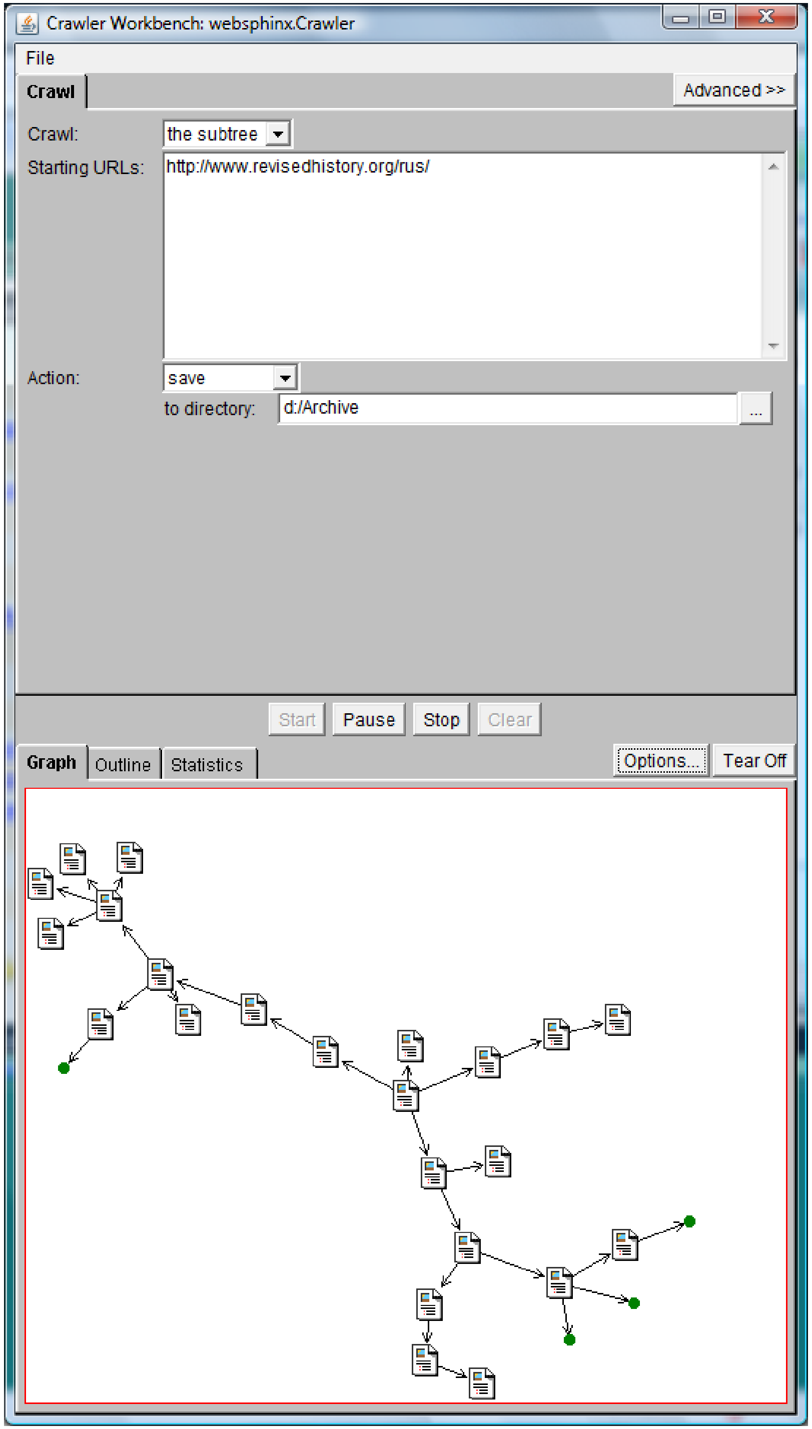
Crawler Workbench: websphinx.Crawler

Crawl: the subtree

Starting URLs: http://www.newparadigma.ru/

Action: none

Техфорум :: Проект ЦИВИЛИЗАЦИЯ
http://www.newparadigma.ru/forum/search.php?5,search=.+%2C+%22+%20,page=1,match_ty

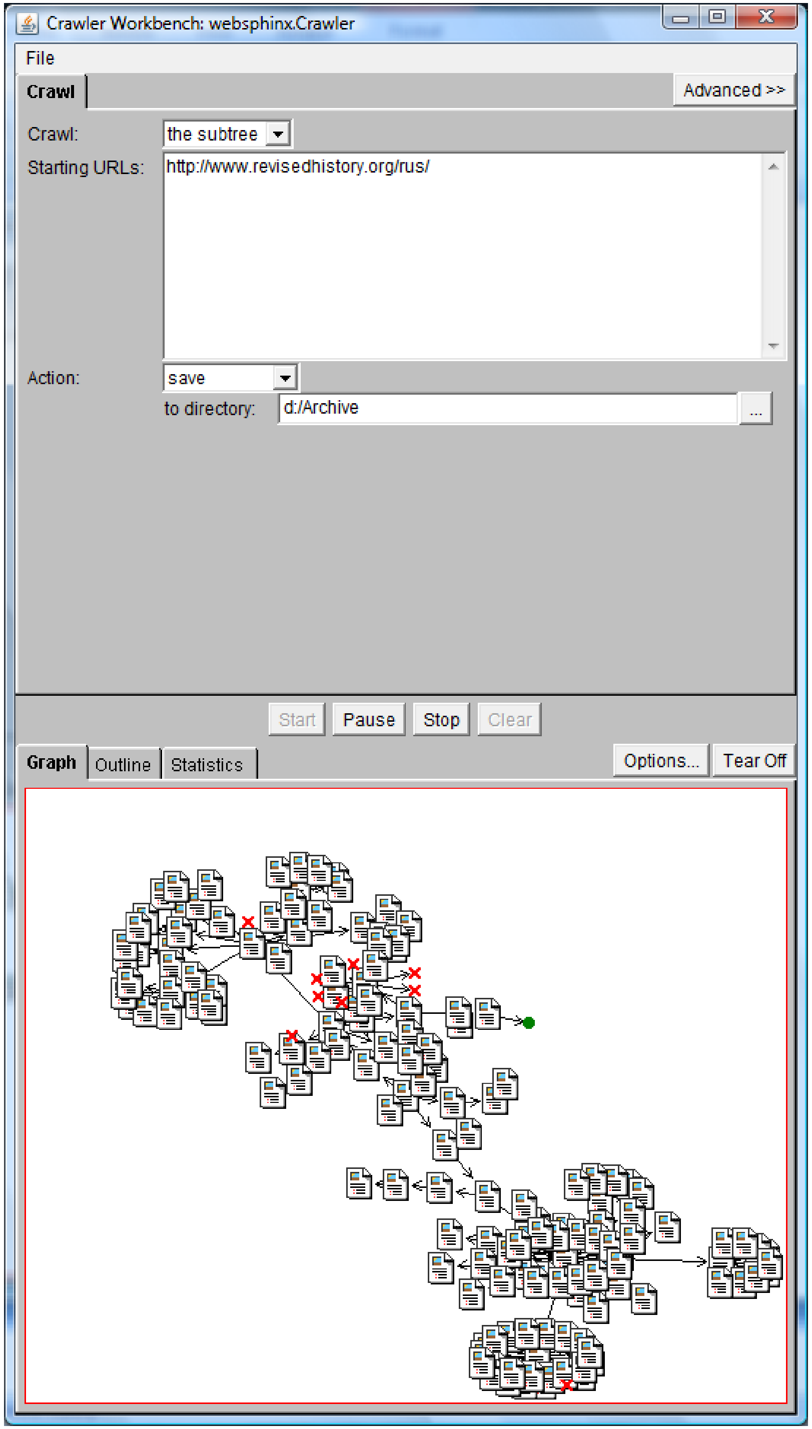


Crawl: the subtree

Starting URLs: http://www.revisedhistory.org/rus/

Action: save

to directory: d:/Archive



Crawl: the subtree

Starting URLs: http://www.revisedhistory.org/rus/

Action: save

to directory: d:/Archive

W3C Protocol Library

- http://www.w3.org/Library/
  - Libwww - the W3C Protocol Library - is a highly modular, general-purpose client side Web API (Application Programming Interface) written in C for Unix and Windows (Win32). It's intended for both small and large applications, like browser/editors, robots, batch tools, etc. The purpose of libwww is to serve as a testbed for protocol experiments.

Web Document Retrieval Links

- http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471666556.html
  “Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage” by Zdravko Markov, Daniel T. Larose: Chapter 1 is available
- http://dbpubs.stanford.edu:8090/pub/1999-66

  “The PageRank Citation Ranking: Bringing Order to the Web” by Larry Page, Sergey Brin, Rajeew Motwani, and Terry Winograd
- http://www.cs.cornell.edu/home/kleinber/auth.ps

  “Authoritative Sources in a Hyperlinked Environment” by Jon M. Kleinberg

Links on Document Classification and Clustering

- http://mitpress.mit.edu/catalog/item/default.asp?sid=B30F92E7-F5BC-4009-916E-D28B2CA762F2&ttype=2&tid=3525&mode=toc “Truth from Trash: How Learning Makes Sense” by Chris Thornton, Preface and Chapter 1 are available
- http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf

  “An Introduction to the WEKA Data Mining System” by Zdravko Markov and Ingrid Russel
- http://www.cs.ccsu.edu/~markov/ccsu_courses/playtennis.pdf

  “Play Tennis Example”, from lecture slides for “Machine learning” by Tom Mitchell, http://www.cs.cmu.edu/afs/cs.cmu.edu/user/mitchell/ftp/mlbook.html
- http://www.mathworks.com/access/helpdesk/help/toolbox/stats/dendrogram.html

  Dendrogram - a hierarchical clustering function

Web Mining

- *Web content mining* - discovery of Web document content patterns (text mining).
- *Web structure mining* - discovery of hypertext/linking structure patterns
  - – use hyperlinks to enhance text classification
  - – page ranking
  - – modeling and measuring the Web
- *Web usage mining* - discovery of web users activity patterns
  - – mining web server logs
  - – mining client machine access logs