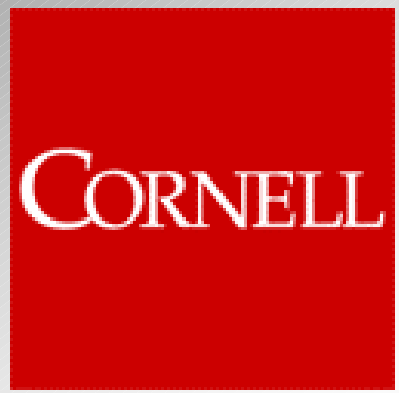


Automatically Constructing Descriptive Site Maps

Pavel Dmitriev

Cornell University

Joint work with Carl Lagoze

8th Asia Pacific Web Conference, January 17, 2006

- The Prime Pages (prime number research, records and resources)
  - Lists of primes
    - Single primes
      Keywords: **prime**, **certificate**, **primes**, **bit**, **random**.
      A 36,007 **bit** "Nearly **Random**" **Prime** A 36,007 **bit** "nearly **random**" **prime** This 10,839 digit **prime** does not have a nice short description!... While new records with cyclotomy are being boiled, I have another kind of large **prime** for you: a 36007 **bits** almost **random**, proved **prime**.... Mersenne Glossary **Prime** Curios!...
      http://www.utm.edu/research/primes/lists/single_primes/ - 7pages
    - Lists of small primes (less than 1000 digits)
      Keywords: **primes**, **first**, **digits**, **prime**, **twin**.
      Move up one level] From other sites All the **primes** below 100,711,433 (5.8 million **primes**) All the **primes** below 2,000,000,009 (98 million **primes**)... **Prime** Lists FAQ e-mail list Titans **Prime** Links Submit **primes** What is small? Depends on your context, but since this site focuses on titanic **prime**... Lists of small **primes** (less than 1000 **digits**) Lists of small **primes** (Another of the **Prime** Pages' resources) Home Search Site Largest The 5000...
      http://www.utm.edu/research/primes/lists/small/ - 4pages
    - Primes just less than a power of two
      Keywords: **prime**, **bits**, **primes**, **69**, **45**.
      Pages: 8-100 **bits**, 101-200 **bits**, 201-300 **bits**, 301-400 **bits**. n ten least k's for which 2n-k is **prime**.... **Prime** Lists FAQ e-mail list Titans **Prime** Links Submit **primes** Here is a frequently asked question at the **Prime** Pages: I am working on an algorithm and... **Prime** Lists FAQ e-mail list Titans **Prime** Links Submit **primes** When designing algortihms, sometimes we need a list of the **primes** just less than a power...
      http://www.utm.edu/research/primes/lists/2small/ - 5pages
  - Modular restrictions on Mersenne divisors
    Keywords: **prime**, **mod**, **p-1**, **theorem**, **proof**.
    Let p be a **prime** and a any integer, then ap = a (**mod** p).... 1 (**mod** p)). Finally, multiply this equality by **p-1** to complete the **proof**.... Let p be a **prime** which does not divide the integer a, then a**p-1** = 1 (**mod** p)....
    http://www.utm.edu/research/primes/notes/proofs/MerDiv.html - 5pages
  - Proofs that there are infinitely many primes
    Keywords: **prime**, **primes**, **primality**, **theorem**, **test**.
    Prime Lists FAQ e-mail list Titans **Prime** Links Submit **primes** Euclid may have been the first to give a proof that there are infintely many **primes**.... **Prime** Page References **Prime** Page References (references for the **Prime** Pages) Home Search Site Largest Finding How Many?... Notice that different a's can be used for each **prime** q.) **Theorem** 2 can be improved even more: if F...
    http://www.utm.edu/research/primes/notes/proofs/infinite/index.html - 17pages

Outline

- Motivation
- The Approach
- Evaluation & Discussion
- Conclusion & Future Work

Motivation

- Web sites are getting more and more complex
  - Increased number of pages
  - Structured for SEO
- Difficult for users to navigate
- Site maps contain little information, often outdated

Applications

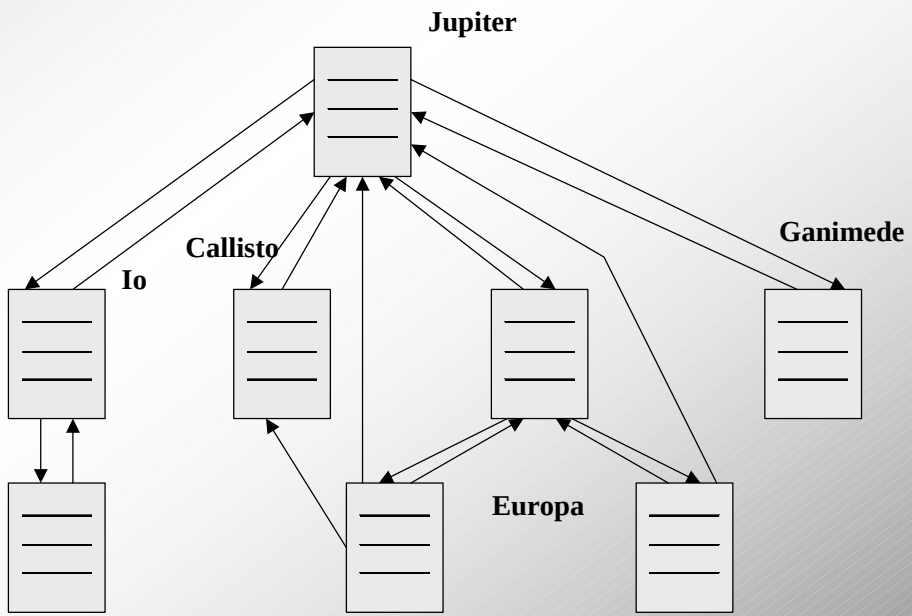
- Site maps can be constructed by
  - Web site administrator
  - Search engine
- Can be used to
  - Aid navigation
  - Organize search results
  - Improve search quality

The Approach

- Identify important sections on a web site
- Build a site map
- Generate descriptions

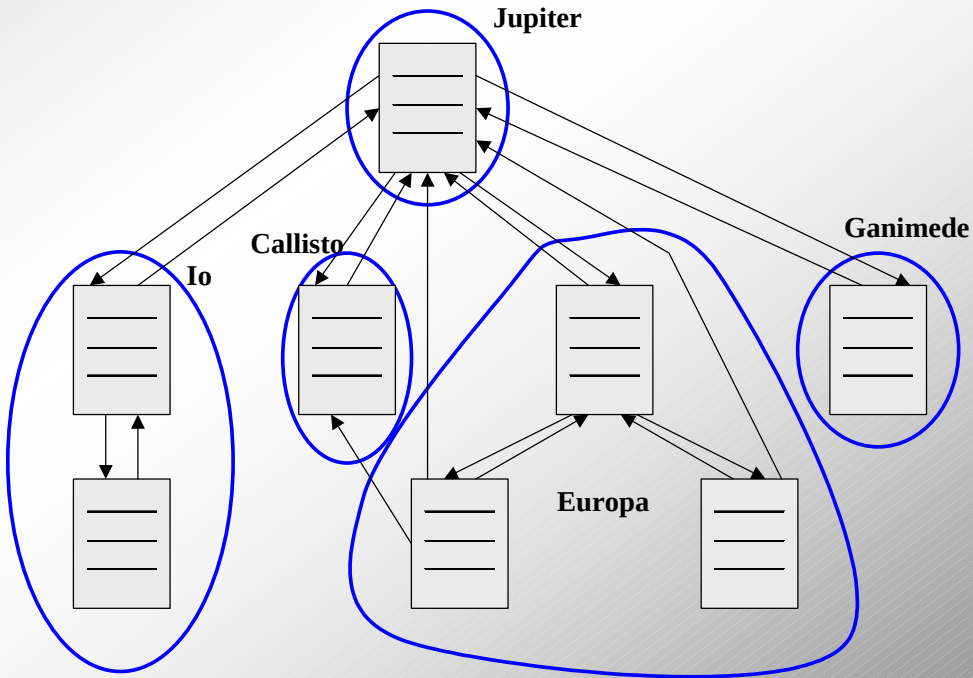
Identifying Site Sections

- Use the system for identifying compound documents (cDocs) we developed earlier



Identifying Site Sections

- Use the system for identifying compound documents (cDocs) we developed earlier



Identifying Site Sections

- cDoc = important section of a web site

- The process consists of two stages:
  - Learn a "profile" of a cDoc from examples
  - Use the "profile" to identify cDocs on new web sites

More details:

Dmitriev et al. "*As We May Perceive: Inferring Logical Documents from Hypertext*" Hypertext-2005, Salzburg, Austria.

Building a Site Map

- The Approach:
  - Identify site sections
  - **Build a site map**
  - Generate descriptions
- Build site map in two stages:
  - Identify section leader pages
  - Arrange sections into a hierarchy, according to user preferences

Identifying Leader Pages

- Use a heuristic approach
  - If have “index.html” etc., make it a leader page
  - Otherwise, take leader page be the page having largest number of inlinks from other sections

Building a hierarchy

- Build a section graph
  - Nodes are sections
  - Edge from s_1 to s_2 iff there is a down- or same-level link from some page in s_1 to the leader page of s_2.
- Hierarchy is a BFS traveral of the graph starting from the node containing the root page of the site

Satisfying User Constraints

- Allow 3 types of constraints
  - Min number of pages in a section
  - Max depth of the site map
  - Max number of leaf items in the site map
- Enforce using a DFS traversal of the site map, merging the nodes which violate the constraints

Generating Descriptions

- The Approach:
  - Identify site sections
  - Build a site map
  - **Generate descriptions**
- For every item, generate
  - Title (anchortext)
  - Summary

Multi-doc Summarization

- Feature extraction
  - Extract features relevant to estimating significance of a sentence
- Sentence ranking
  - Order sentences according to their significance
- Summary generation
  - Generate summary by picking most significant sentences, but avoiding redundancy

Summary Generation

- Use title, content, and anchortext as features
- Identify sentence boundaries using *http://www.answerbus.com/sentence/*
- Rank sentences according to similarity to the centroid, or to *k* most frequent keywords
- Pick *m* most significant sentences, but have a similarity threshold to make sure that the sentence is not too similar to the existing summary

Experimental Results

- Dataset:
  - 20 web sites on educational topics
  - Important sections identified manually
- Evaluate every component separately
  - Here focus on anchortext and summary generation

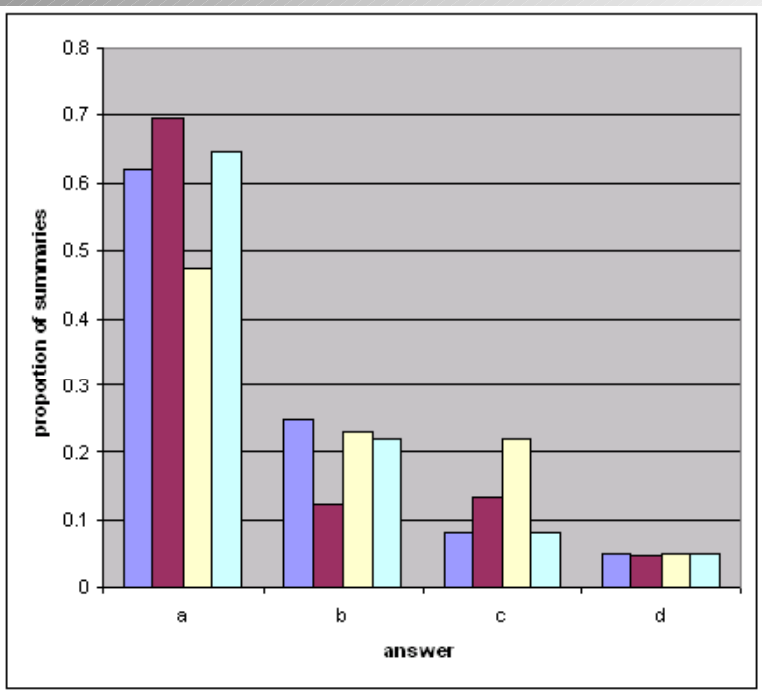
Anchortext Generation

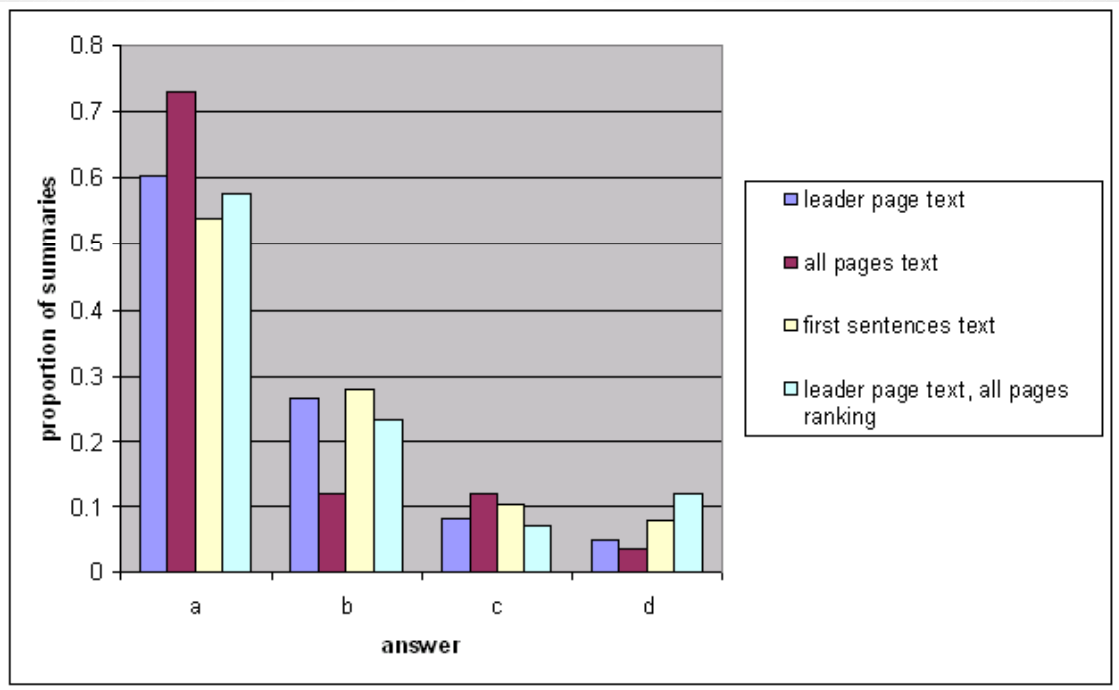
- Titles are much better than anchortexts
- Anchortexts within the same web site seem to contain little descriptive information
  - Search engines may ignore anchortexts from the same web site

Summary Generation

- Does the summary give the user a good idea of the topic/function/goal of the site map item it describes?
  (a) Gives a very good idea
  (b) Gives a pretty good idea
  (c) Gives a rough idea of it
  (d) There are some clues for it in the summary, but they are not very easy to see
  (e) The summary contains no useful information

Summary Generation





Results for summary generation using centroid (left) and 5 most frequent keywords (right)

Conclusion

- Described an approach for automatic generation of descriptive site maps
- Evaluation on 20 web sites showed that our approach can generate high quality site maps
- Future work:
  - Large scale evaluation across multiple domains
  - Explore new browsing paradigms resulting from integrating our approach with a search engine

The End

Thank you!

- The Prime Pages (prime number research, records and resources)
  - Lists of primes
    - Single primes
      Keywords: **prime**, **certificate**, **primes**, **bit**, **random**.
      A 36,007 **bit** "Nearly **Random**" **Prime** A 36,007 **bit** "nearly **random**" **prime** This 10,839 digit **prime** does not have a nice short description!... While new records with cyclotomy are being boiled, I have another kind of large **prime** for you: a 36007 **bits** almost **random**, proved **prime**.... Mersenne Glossary **Prime** Curios!...
      http://www.utm.edu/research/primes/lists/single_primes/ - 7pages
    - Lists of small primes (less than 1000 digits)
      Keywords: **primes**, **first**, **digits**, **prime**, **twin**.
      Move up one level] From other sites All the **primes** below 100,711,433 (5.8 million **primes**) All the **primes** below 2,000,000,009 (98 million **primes**)... **Prime** Lists FAQ e-mail list Titans **Prime** Links Submit **primes** What is small? Depends on your context, but since this site focuses on titanic **prime**... Lists of small **primes** (less than 1000 **digits**) Lists of small **primes** (Another of the **Prime** Pages' resources) Home Search Site Largest The 5000...
      http://www.utm.edu/research/primes/lists/small/ - 4pages
    - Primes just less than a power of two
      Keywords: **prime**, **bits**, **primes**, **69**, **45**.
      Pages: 8-100 **bits**, 101-200 **bits**, 201-300 **bits**, 301-400 **bits**. n ten least k's for which 2n-k is **prime**.... **Prime** Lists FAQ e-mail list Titans **Prime** Links Submit **primes** Here is a frequently asked question at the **Prime** Pages: I am working on an algorithm and... **Prime** Lists FAQ e-mail list Titans **Prime** Links Submit **primes** When designing algortihms, sometimes we need a list of the **primes** just less than a power...
      http://www.utm.edu/research/primes/lists/2small/ - 5pages
  - Modular restrictions on Mersenne divisors
    Keywords: **prime**, **mod**, **p-1**, **theorem**, **proof**.
    Let p be a **prime** and a any integer, then ap = a (**mod** p).... 1 (**mod** p)). Finally, multiply this equality by **p-1** to complete the **proof**.... Let p be a **prime** which does not divide the integer a, then a**p-1** = 1 (**mod** p)....
    http://www.utm.edu/research/primes/notes/proofs/MerDiv.html - 5pages
  - Proofs that there are infinitely many primes
    Keywords: **prime**, **primes**, **primality**, **theorem**, **test**.
    Prime Lists FAQ e-mail list Titans **Prime** Links Submit **primes** Euclid may have been the first to give a proof that there are infintely many **primes**.... **Prime** Page References **Prime** Page References (references for the **Prime** Pages) Home Search Site Largest Finding How Many?... Notice that different a's can be used for each **prime** q.) **Theorem** 2 can be improved even more: if F...
    http://www.utm.edu/research/primes/notes/proofs/infinite/index.html - 17pages